Remote Elicitation of Inflectional Paradigms to Seed Morphological Analysis in Low-Resource Languages

John Sylak-Glassman*, Christo Kirov*, David Yarowsky**
*Center for Language and Speech Processing
**Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218, USA
jcsg@jhu.edu, ckirov@gmail.com, yarowsky@jhu.edu

Abstract

Structured, complete inflectional paradigm data exists for very few of the world's languages, but is crucial to training morphological analysis tools. We present methods inspired by linguistic fieldwork for gathering inflectional paradigm data in a machine-readable, interoperable format from remotely-located speakers of any language. Informants are tasked with completing language-specific paradigm elicitation templates. Templates are constructed by linguists using grammatical reference materials to ensure completeness. Each cell in a template is associated with contextual prompts designed to help informants with varying levels of linguistic expertise (from professional translators to untrained native speakers) provide the desired inflected form. To facilitate downstream use in interoperable NLP/HLT applications, each cell is also associated with a language-independent machine-readable set of morphological tags from the UniMorph Schema. This data is useful for seeding morphological analysis and generation software, particularly when the data is representative of the range of surface morphological variation in the language. At present, we have obtained 792 lemmas and 25,056 inflected forms from 15 languages.

Keywords: Low-Resource Languages, Morphology, UniMorph Schema, Seed Corpus, Crowdsourcing, Linguistic Fieldwork

1. Introduction

For most of the world's languages, no structured, complete inflectional paradigms in a machine-readable format are available for human language technology (HLT) applications. These paradigms are necessary as seed data for downstream tasks, especially morphological analysis and generation. Increased global internet access has made it possible to gather inflectional paradigm data directly from remotely-located speakers of many of the world's languages. We present methods for data collection grounded in the practice of linguistic fieldwork. The methods developed thus far have been used to gather a corpus of inflectional paradigm data for 15 languages, with materials ready for eliciting data from speakers of 33 additional languages.[1] The methods we have developed entail first constructing detailed elicitation templates in which speakers supply desired inflectional forms (§2.). These templates include: 1) A prompt in a high-resource world language (e.g. English) that speakers can translate into the target low-resource language (e.g. Uzbek), 2) a short description of the desired inflectional form in pedagogical or linguistic terminology (e.g. 'aorist'), and 3) a detailed specification of the inflectional form in terms of the UniMorph Schema, introduced by Sylak-Glassman et al. (2015a, 2015b).[2] These elicitation templates are distributed to remotely-located speakers, who may be either professional translators or untrained native speakers recruited via crowdsourcing platforms (§3.). Even a very small amount of inflectional paradigm seed data, particularly when it is representative of the surface morphological variation of the language, can serve as effective training data for machine learning, particularly for generating possible inflected forms (§4.). We demonstrate this by assessing the ability of recent morphological paradigm completion software by Durrett and DeNero (2013) to predict inflected word forms using the small amount of training data supplied by elicitation, the full amount of Wiktionary data, and a small random subset of that Wiktionary data.

2. Elicitation Templates

Language informants, including professional translators and untrained native speakers, are asked to supply inflectional forms by translating prompts in English or another major world language with which they may be more familiar (e.g. French, Spanish). These prompts are carefully constructed and phrased to make the use of a particular inflectional form obligatory in the given context. Contextual material is necessary to make distinctions that are not overtly made in the prompt language, but are made in the target language. Parentheses are used to separate contextual material from the material that is to be translated. For example, a bare noun like "house" may be realized differently in different grammatical contexts in languages with nominal case, and to cause speakers to supply an accusative (direct object) case form, they are prompted with "(the fire destroyed the) house." The practice of eliciting translations of carefully constructed prompts is a standard technique used in

[1] These 15 languages are listed in Table 1, and the additional 33 languages include: Acholi, Armenian, Balochi, Burmese, Cambodian, Coptic, Dinka, Gujarati, Hausa, Hindi, Kikuyu, Kongo (Fioti), Kurdish, Lao, Lingala, Luganda, Maguindanao, Malay, Mongolian, Nahuatl, Nubian, Nuer, Pashto, Samoan, Somali, Spanish, Surubu (Fiti), Swahili, Tausug, Thai, Vietnamese, Wolof, and Zande.

[2] The name used in those works for this schema, the Universal Morphological Feature Schema, is deprecated in favor of the name UniMorph Schema.

linguistic fieldwork to reveal grammatical distinctions.[3]

For each requested form, additional descriptive information is given using either pedagogical terminology commonly used when teaching the elicited language or (as a secondary option) standard technical terminology used by linguists. This information is an additional resource that helps translators pinpoint which inflectional form is desired. Fig. 1 (on the following page) shows a template given to translators and includes both the additional descriptive information and the English prompt.

Each inflectional form is associated with a set (i.e. a vector) of features from the UniMorph Schema, as shown on the left in Fig. 2. These features, which are not shown to language informants, precisely describe the meaning of each form using a language-independent annotation scheme that allows for straightforward comparison with inflectional forms in other languages that have a similar meaning. The UniMorph Schema contains over 277 features from 25 dimensions of meaning (i.e. morphological categories such as tense, number, case, evidentiality, etc.). These features are motivated by overt, usually affixal morphological contrasts in at least one natural language. Each feature represents a semantic "atom" within a dimension of meaning that is not decomposed further in any natural language.

For example, within the dimension of number, the Austronesian language Sursurunga distinguishes singular (1), dual (2), paucal (>3), greater paucal (>4), and plural (many).[4] Larike (also Austronesian; Corbett 2000:21) marks singular, dual (exactly 2), trial (exactly 3), and plural (>4). The UniMorph Schema includes the features necessary to distinguish all these categories, which are marked by surface contrasts in each language and are not decomposed further in any natural language.[5] However, were a language to be discovered that distinguished a blended dual-trial (2/3) from singular, paucal, greater paucal, and plural, the UniMorph Schema would combine the minimal dual (exactly 2) and trial (exactly 3) features together additively to annotate the blended category (as DU+TRI).

The features of the UniMorph Schema are therefore motivated by findings from the linguistic typology literature that demonstrate the finest-grained, overt morphological distinctions that are made in natural languages. These features can be combined additively or disjunctively to specify non-minimal morphological categories. For full details on the UniMorph Schema, see Sylak-Glassman et al. (2015a, 2015b).[6]

Because inflected wordforms in every language in the corpus are labeled using a single annotation scheme with consistent, meaningful, non-overlapping features, any application that is designed to take in data from one language can also process data from any other language in the corpus as well as data from any other source annotated using the UniMorph Schema. Similarly, because the UniMorph Schema can be used with any language to capture the meaning encoded by its inflectional morphology, its use in annotating inflected wordform data allows for the design of systems which can be language agnostic or independent without disregarding linguistic diversity (Bender, 2009).

3. Gathering Inflectional Paradigm Data

The templates described above were used to gather inflectional paradigm data from professional translators and from untrained native speakers recruited through Amazon's Mechanical Turk crowdsourcing platform.

One of the primary advantages of gathering paradigm data from professional translators was their greater metalinguistic awareness, including their overall knowledge of the language's structure and their comfort with pedagogical and linguistic terminology. Professional translators can therefore use descriptive information to pinpoint which form is desired, and can more reasonably be asked to choose lexemes/lemmas that illustrate the range of surface morphological variation in the language (e.g. different noun declensions or verb conjugations). Professional translators can also supply lemmas that illustrate variation in fixed lexical properties, such as animacy with nouns or transitivity with verbs.

By contrast, untrained native speakers cannot be assumed to have a high degree of metalinguistic awareness. For this reason, only the contextual natural language prompts in a major world language can be used to convey the differences between inflectional forms, and lemmas must be chosen in advance for the speakers. Lemmas can be chosen during the construction of elicitation paradigms based on knowledge of the language gleaned from a descriptive resource, but this entails additional effort for the linguist.

An alternate strategy is to choose lemmas automatically and gather data in sufficient quantities to increase the likelihood that examples of most surface variation will be collected. To do this, two problems must be solved: 1. Choosing lemmas in a way that is likely to maximize variation, and 2. constructing glosses that are intelligible to untrained speakers. Lemmas which maximize lexical variation can be chosen using information on lexical properties available in the entries of lexical resources like PanLex (Kamholz et al., 2014).

Given a set of lemmas whose paradigms we would like to elicit and an initial template designed for professional translators, we can automatically generate new prompts for each new lemma to display to untrained speakers. First, we excise the inflected lemmas from the existing prompts, replacing them with a placeholder tag indicating inflection (e.g. 'I

[3] Although elicitation has significant limitations and is typically combined with textual analysis, it is easily adaptable to crowdsourcing, useful for gathering paradigmatic data (especially when all or most distinctions within the paradigm are already known), and allows for the collection of highly specific forms, particularly those whose use requires an atypical context. For a nuanced discussion of elicitation and its limitations, see Chelliah (2001).

[4] Corbett (2000:27-29) writes that the paucal is used for 3-4 people or nuclear families of any size while the greater paucal is used for strictly 4 or more people, including larger-sized groups, with no strict dividing line between the greater paucal and plural.

[5] Corbett (2000) devotes considerable discussion to showing that no language uses a quadral (4) category, especially in distinction from the paucal and greater paucal in Sursurunga, even though such a distinction is in principle possible.

[6] Additional detailed documentation on the UniMorph Schema will be made available at the temporary site for the UniMorph Project: http://ckirov.github.io/UniMorph/

Please do NOT include pronouns in your translations	English Lemma:	...
Please use SINHALESE script, not Roman.	Sinhalese Lemma:	
Additional Detail (consult if needed)	*English Source*	
infinitive	to come	
verbal noun	(the man's) coming (will be of interest to others)	
present participle	(the man) coming (to the building saw his friend)	
past participle	(the man who had) come (from the north visited us)	
future participle (of obligation)	(the man) must come (OR the man is to come)	
conditional participle	when coming (to town, the man saw his friend)	
"while" converb / adverbial participle	while coming (to town, the man saw many birds)	
past, passive converb	having come (to town, the man met with his friend)	
agentive participle	one who comes	
prospective participle	about to come (home, the man saw that the gate was open)	
general (present)	he comes	
present continuous	he is coming	
past continuous	he was coming	
future (general)	he will come	
definite future	he will (certainly) come	
present habitual	he comes (here every day)	
past habitual	he used to come (here every day)	
presumptive habitual	he probably comes (here every day)	
counterfactual habitual	had he come (here every day, then he would have received his pay)	
simple perfect	he came	
present perfect	he has come	
past perfect	he had come	
future perfect	he will have come	

Figure 1: Excerpt of a blank elicitation template for Sinhalese verbs as presented to a translator.

Universal Morphological Features				English Lemma:	house	car
				Turkish Lemma:	ev	araba
Case	Number	Part of Speech	*Additional Detail (consult if needed)*	*English Source*		
NOM	SG	N	nominative singular	(the) house (stood on that land)	ev	araba
GEN	SG	N	genitive singular	(the fence) of (the) house / (the) house's (fence)	evin	arabanın /arabanın
DAT	SG	N	dative singular	(the new roof was beneficial) to (the) house	eve	arabaya
ACC	SG	N	accusative singular	(the fire destroyed the) house	evi	arabayı
ESS	SG	N	locative singular	(he is) at (the) house	evde	arabada
ABL	SG	N	ablative singular	(he went) from (the) house (to town)	evden	arabadan
EQTV	SG	N	equative singular	(a big apartment is) like (a) house	ev gibi	araba gibi
COMPV	SG	N	comparative singular	(a cottage is smaller) than (a) house	evden	arabadan
ALL	SG	N	directive singular	(he went) toward (the) house	eve doğru	arabaya doğru
INS	SG	N	instrumental singular	(the view was blocked) by (the) house	ev tarafından	araba tarafından
COM	SG	N	comitative singular	(the land will be sold) with (the) house	ev ile	araba ile
NOM	PL	N	nominative plural	(the) houses (stood on that land)	evler	arabalar
GEN	PL	N	genitive plural	(the fences) of (the) houses / (the) houses' (fences)	evlerin	arabaların
DAT	PL	N	dative plural	(the new roofs were beneficial) to (the) houses	evlere	arabalara
ACC	PL	N	accusative plural	(the fire destroyed the) houses	evleri	arabaları
ESS	PL	N	locative plural	(they are) in (the) houses	evlerin içinde	arabaların içinde
ABL	PL	N	ablative plural	(they went) from (the) houses (to town)	evlerden	arabalardan
EQTV	PL	N	equative plural	(big apartments are) like houses	evler gibidir	arabalar gibidir
COMPV	PL	N	comparative plural	(cottages are smaller) than houses	evlerden	arabalardan
ALL	PL	N	directive plural	(they went) toward (the) houses	evlere doğru	arabalara doğru
INS	PL	N	instrumental plural	(the view was blocked) by (the) houses	evler tarafından	arabalar tarafından
COM	PL	N	comitative plural	(the land will be sold) with (the) houses (on it)	evler ile	arabalar ile

Figure 2: Excerpt of an elicitation template for Turkish nouns with professional translator's input in the shaded area and the UniMorph Schema annotation on the left for each form.

am going' becomes 'I am VBG', where VBG is the standard Penn Treebank tag for the *-ing* verb form in English). This placeholder can be replaced with the inflected form of any lemma we wish to insert to make a new prompt (e.g. 'I am VBG' + 'try' = 'I am trying'). Since the prompts are always in a high-resource language like English, it is possible to look up the inflected forms of each replacement lemma. If a replacement lemma is transitive, we add a generic object pronoun to an intransitive gloss template (e.g. 'I am VBG it').

The elicited output of untrained speakers can be scored to determine its accuracy. Scoring consists of comparing the elicited output to either verified or predicted word-forms using string edit distance to determine the extent to which they diverge. Divergence can then be penalized, with thresholds established for acceptance, human inspection, and rejection. Speaker output can be compared to human-produced, gold standard data from reference grammars, professional translators, or a resource such as Wiktionary. It may also be compared to the output predicted by state-of-the-art paradigm completion software that is known to achieve high accuracy, such as Durrett and DeNero (2013), Ahlberg et al. (2014, 2015), and Nicolai et al. (2015).

The current corpus of inflectional paradigms gathered via elicitation from professional translators includes data from 15 languages, and is described in detail in Table 1.

Language	*L/F*	*Adj*	*N*	*Pro*	*V*
Akan (Twi)	L	20	20	–	12
	F	101	80	25	564
Amharic	L	4	6	–	10
	F	208	180	20	885
Azeri	L	15	15	–	15
	F	105	990	503	1070
Dari	L	20	20	–	20
	F	180	620	66	1398
Farsi	L	25	20	–	20
	F	225	620	65	1400
Igbo	L	14	14	–	14
	F	56	196	12	459
Indonesian	L	20	20	–	60
	F	140	200	16	449
Punjabi	L	6	9	–	12
(Indian)	F	237	397	389	763
Sinhala	L	6	9	–	12
	F	237	397	389	763
Tagalog	L	20	20	–	60
	F	320	120	24	1440
Tajik	L	20	20	–	20
	F	180	620	65	1423
Tigrinya	L	3	8	–	6
	F	138	160	12	378
Turkish	L	13	13	–	38
	F	91	862	–	940
Uzbek	L	15	15	–	18
	F	105	990	503	1265
Yoruba	L	21	22	–	22
	F	126	308	31	550
TOTAL: (L)	792	222	231	–	339
(F)	25056	2449	6740	2120	13747

Table 1: The content of the current elicited inflectional paradigm corpus, with the number of lemmas (L) and inflected forms (F) for each language indicated, with cross-language totals shown in the bottom row.

	Prof. Trans.	Wikt. Sm.	Wikt. Full
Azeri N	55.0%	70.7%	95.7%
Training	15	15	278
Test	30	30	30
Turkish N	23.1%	57.3%	94.7%
Training	13	13	1800
Test	200	200	200
Turkish V	56.0%	66.6%	92.4%
Training	13	13	421
Test	46	46	46

Table 2: Per form accuracy of Durrett and DeNero's (2013) morphological paradigm learner on datasets of varying sizes for three language–part-of-speech combinations.

4. Effectiveness of Seed Paradigms in Morphological Generation

To test the utility of elicited paradigms, data provided by professional translators was compared to the full quantity of data available from Wiktionary and a random subset of that full dataset that contained the same number of lemmas as in the data from translators.[7] Comparing to a small quantity of randomly chosen data from Wiktionary allows the effect of translators' manual choice of lemmas to be assessed. It also simulates accurate output from untrained speakers with unconstrained lemma choice since the lemmas in the randomly chosen subset likely do not capture the full range of surface morphological variation in the language.

Each dataset served as training input to Durrett and DeNero's (2013) morphological paradigm learner, which learns position-anchored string transformations from the lemma to the inflected form. The transformations that may apply to predicting a given inflectional form from a given lemma are ranked in a classification step using a log-linear model with character 5-grams in both directions as well as n-grams for preserved segments used as features. Table 2 shows the individual form accuracy for each dataset, along with the source and size of the training data. Within each language and part-of-speech, a single test set randomly drawn from the full Wiktionary dataset was used. Each test set represented approximately 10% of the full amount of data available on Wiktionary.

Although accuracy is lowest when the professional translators' data is used to train the model, the model cannot run on every possible random subset of the Wiktionary data, as Wiktionary paradigms are not guaranteed to be complete or contain the same set of inflections. While the first randomly chosen subset of Wiktionary's Turkish verb data happened to allow the model to generate predictions, Turkish nouns required one resampling of the random subset, and Azeri nouns required seven such resamplings before the model could generate predictions using the small Wiktionary subset.

This indicates that the range of inflectional variation among lemmas provided in the professional translators' data, which reflects the range of possible surface variation in the language as a whole, supplied sufficiently diverse data to allow the model to learn diverse string transformations and generate a sufficiently wide range of predictions. Hold-one-out validation on the lemmas in the translators' training data set showed, for example, that certain lemmas contributed strongly to the success of the Durrett and DeNero (2013) learner, as much as 25% of the individual form accuracy in Turkish verbs.

While the informants' chosen lemmas were highly diverse as a set and informative on an individual basis, the lack of more than one example for any given inflectional class variant appears to have a negative effect on the model's ability to accurately predict the forms of new lemmas in each class. Because the Wiktionary data as a whole is likely to better represent the type frequency of lemma variants (most lemmas come from one of a small number of regular inflection classes), even a randomly chosen small subset is more likely to contain multiple examples of some class. This allows the model to learn the appropriate string transformation rules for higher frequency inflectional classes and pro-

[7] Because our elicitation efforts focused on gathering new data for low-resource languages, only Turkish and Azeri had sufficient representation on Wiktionary to allow for the comparisons performed here.

vides more data for inflecting lemmas as belonging to one of those classes.

Overall, the results presented here indicate that even a very small amount of representative data is potentially useful for generating possible morphological forms, and that an increase in the number of lemmas per variant would likely lead to better classification results and improved accuracy. As data are collected from languages with additional gold standard morphological data (e.g. from Wiktionary), further testing will allow the disambiguation of the effects of the representation of variation in data, the number of examples of each variant, and the amount of data needed given different degrees of inflectional variation.

5. Conclusion

To address the need for structured, complete morphological paradigm data for HLT applications, particularly for low-resource languages, we have developed methods to gather inflectional paradigm data from remotely-located language professionals and untrained native speakers. These methods are grounded in traditional linguistic fieldwork practices, can be tailored to experts of varying skill levels, and are designed to be implemented at scale with remotely-located speakers of any language. The application of these methods has resulted in a corpus of inflectional paradigms for verbs, nouns, adjectives, and personal pronouns in 15 languages, with materials for 33 additional languages ready to be deployed and materials for other languages under development.

Ahlberg, M., Forsberg, M., and Hulden, M. (2014). Semi-supervised learning of morphological paradigms and lexicons. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 569–578, Gothenburg, Sweden. Association for Computational Linguistics.

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

Bender, E. M. (2009). Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics (ILCL): Virtuous, Vicious or Vacuous?*, pages 26–32, Stroudsburg, PA. Association for Computational Linguistics.

Chelliah, S. L. (2001). The role of text collection and elicitation in linguistic fieldwork. In Paul Newman et al., editors, *Linguistic Fieldwork*, pages 152–165. Cambridge University Press, Cambridge.

Corbett, G. G. (2000). *Number*. Cambridge University Press, Cambridge.

Durrett, G. and DeNero, J. (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 (NAACL-HLT)*, pages 1185–1195. Association for Computational Linguistics, Atlanta.

Kamholz, D., Pool, J., and Colowick, S. M. (2014). PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 3145–3150, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Nicolai, G., Cherry, C., and Kondrak, G. (2015). Inflection generation as discriminative string transduction. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL (NAACL)*, pages 922–931, Denver, CO. Association for Computational Linguistics.

Sylak-Glassman, J., Kirov, C., Post, M., Que, R., and Yarowsky, D. (2015a). A universal feature schema for rich morphological annotation and fine-grained cross-lingual part-of-speech tagging. In Cerstin Mahlow et al., editors, *Proceedings of the 4th Workshop on Systems and Frameworks for Computational Morphology (SFCM)*, Communications in Computer and Information Science, pages 72–93. Springer, Berlin, September.

Sylak-Glassman, J., Kirov, C., Yarowsky, D., and Que, R. (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, July. Association for Computational Linguistics.