# A Word-and-Paradigm Workflow for Fieldwork Annotation

**Maria Copot**[1], **Sara Court**[2], **Noah Diewald**[2], **Stephanie Antetomaso**[2], and **Micha Elsner**[2]

[1]Laboratoire de Linguistique Formelle, Université Paris Cité
[2]Department of Linguistics, Ohio State University

mcopot@etu.u-paris.fr, {court.22,diewald.21,antetomaso.2}@osu.edu, melsner0@gmail.com

## Abstract

There are many challenges in morphological fieldwork annotation: it heavily relies on segmentation and feature labeling (which have both practical and theoretical drawbacks), it's time-intensive, and the annotator needs to be linguistically trained and may still annotate things inconsistently. We propose a workflow that relies on unsupervised and active learning grounded in Word-and-Paradigm morphology (WP). Machine learning has the potential to greatly accelerate the annotation process and allow a human annotator to focus on problematic cases, while the WP approach makes for an annotation system that is word-based and relational, removing the need to make decisions about feature labeling and segmentation early in the process and allowing speakers of the language of interest to participate more actively, since linguistic training is not necessary. We present a proof-of-concept for the first step of the workflow: in a realistic fieldwork setting, annotators can process hundreds of forms per hour.[1]

## 1 Introduction

A major component of current workflows for linguistic fieldwork is the creation and curation of Interlinear Glossed Texts (IGT), in which morphological forms are segmented into meaning-bearing units. These are expensive and time-consuming to produce, but constitute important training data for computational fieldwork methods. While IGT are a valuable resource for the study of endangered and under-described languages (Zamaraeva, 2016), annotations that directly segment and label morphemes may have both practical and theoretical shortcomings. Segmentation-based analyses may not always straightforwardly account for the diversity of phenomena attested in the world's languages, making them especially problematic in the early stages of understanding a morphological system. An alternative is provided by analyses that characterize morphological relations at the word level, such as those associated with Word-and-Paradigm approaches (WP; named by Hockett (1954); see Blevins (2016) for a general overview), which do not require segmentation and may allow for more efficient and informative morphological annotation in a low-resource fieldwork setting.

WP theories classify word forms in terms of the shared relationships they exhibit within a connected lexicon. These relationships may be conceptualised as tabular paradigms in which one axis groups items sharing a lexeme and the other groups items sharing a morphosyntactic cell.

| PRS | PRS.3S | PST | PTCP.PRS |
|---|---|---|---|
| run | runs | ran | running |
| live | lives | lived | living |

Table 1: A partial WP paradigm table in English

Note that paradigmatic tables would still be informative about the identities of morphosyntactic cells even without cell labels. Such an unlabeled table can be assembled first and then serve as an aid for post-hoc decisions about how to label the contrasts.

WP-style analyses inform recent work on unsupervised paradigm discovery (Kann et al., 2020b; Wiemerslage et al., 2021; Erdmann et al., 2020) as well as neural inflection and reinflection models without internal segmentation (Kann and Schütze, 2016; Anastasopoulos and Neubig, 2019; Silfverberg and Hulden, 2018). Since WP theories have proven to be such a good fit for "big data" morphology (Elsner et al., 2019), this paper asks whether they can also benefit the "small data" fieldwork setting. We see several potential advantages: modern computational tools can be used to provide initial analyses or suggestions for the annotator; grouping forms together as belonging to the same cells or

[1] All code for the paper can be found at https://github.com/CopotM/WP-workflow-ComputEL2022

*Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 159 - 169
May 26-27, 2022 ©2022 Association for Computational Linguistics

lexemes may be faster and easier than segmenting on a first pass; finally, segment-free annotations of some morphological phenomena may be preferable on theoretical grounds. We conduct pilot experiments in three languages, including a true under-resourced language, to show that trained human annotators can rapidly improve on the results of an unsupervised morphological analyzer (Jin et al., 2020). The workflow we propose takes these corrections as input to bootstrap iterated active learning. We collaborate with non-linguist native speakers of Wao Terero, a language isolate spoken in Ecuador, to evaluate the potential of the proposed methodology to increase community engagement in the annotation process. Finally, we discuss possible next steps in the design of an interactive annotation environment for Word-and-Paradigm morphology.

## 2 Background

### 2.1 Word-and-Paradigm Morphology

Linguistic theory necessarily informs documentary and descriptive methodology (Himmelmann, 1998). Standard workflows for linguistic fieldwork can be seen as theoretically aligned with Item-and-Arrangement (IA) analyses of morphological structure: field linguists often annotate collected texts or transcriptions by slicing words into morphemic subunits, each with a consistent form-meaning association. The resulting IGT can be useful for illustrating morphological structure in certain well-described languages, but IA-based approaches to morphological annotation have two main theoretical drawbacks. First, segmentation may not be able to capture important morphological generalizations, as many linguistic patterns are not strictly segmental. For example, an IA-style gloss b. below is unable to directly convey information about what exactly makes *caught* a past tense in the same way that is possible for *seemed*.

| | | | |
|---|---|---|---|
| a. | *seem-ed* | b. | *caught* |
| | seem-PST | | PST\catch |

Second, it is not always the case that morphological systems exhibit reliable form-meaning correspondences. The meaning of a segmented unit may instead only be interpretable by contrasting it with other forms of the same word, or other words with the same grammatical function. Table 2 shows how the same segmental unit can have different (in this case, opposite) meanings which may only be interpreted in the context of other forms of the same lexeme: in Spanish, there is no unambiguous verb ending for IND.PRS.3SG, nor is there an unambiguous meaning for -a/-e. Instead, both segments are interpretable as IND.PRS.3SG markers only when contrasted with other related forms, like the SUBJ.PRS.3SG. In this case, if one is marked by -a, the other will be marked by -e.

| | IND.PRS.3SG | SUBJ.PRS.3SG |
|---|---|---|
| TO EAT | com-e | com-a |
| TO BUY | compr-a | compr-e |

Table 2: Morphological exchange pattern in Spanish

In addition to the theoretical shortcomings of segmentation-based approaches, IA-style annotation workflows may pose more concrete problems during descriptive or documentary fieldwork. Morphemic segmentation and labeling requires important decisions about the structure of a language's morphology to be made from the very start of the annotation process, even when the researcher lacks sufficient information to do so. WP theories instead take words themselves as the smallest meaning-bearing unit of analysis. By doing so, it is possible to characterize a morphological system as a set of parallel relationships among words. To derive the paradigmatic structure of a system (Bonami and Strnadová, 2019), one must start by establishing pairwise formal relationships that mark a functional contrast. For example, *run* ∼ *runs* and *eat* ∼ *eats* both mark PRS.NON3SG∼PRS.3SG by means of the formal *X*∼*Xs* relationship. Chains of words linked together by morphological relationships make up morphological families (e.g., {*run, running, ran, runner, runners*}; {*eat, eating, ate, eater, eaters*}), which may in turn be aligned according to word forms exhibiting parallel contrasts in meaning. In this way, paradigmatic structure gives rise to both morphological families (sets of inflectionally or derivationally related word forms) and paradigm cells (sets of forms that occupy the same place in the system of contrasts) as structured objects of morphological analysis.

Since decisions about the boundaries and labeling of subword units are unnecessary in such a framework, WP is well-suited to bootstrap morphological annotation. An annotation workflow that labels related structures of words and paradigms as opposed to segmented morphemes can help avoid the pitfalls of making incorrect assumptions about the system at an early stage, which can lead to problematic conclusions about the grammar of the language and be hard to recover from once adopted. Nevertheless, morpheme-style segmentations can still be extracted from WP alternations and paradigms, and morphosyntactic labels can easily be added to paradigmatic cells. In practice, a WP-based annotation system a) captures non-segmental morphological patterns just as naturally as segmental ones, since it's based on alternations at the word level and b) relies on judgements about which relationships are the same and which are different, a more straightforward task for untrained linguistic consultants. The latter aspect may in turn serve to boost community engagement and facilitate crowdsourcing data on a larger scale.

### 2.2 Machine-aided morphological annotation

Morphological annotation is part of a larger language documentation and description workflow which may be systematized as a sequence of data collection, transcription, analysis, annotation, and archival (Thieberger and Berez, 2012). Fieldwork projects are inherently collaborative in nature and their outcomes are ultimately shaped by the unique needs of the multiple stakeholders involved, including the community, the researcher, and the funding organization, among others. A project's goals may include the development of materials for language maintenance and revitalization, community access to digital language technologies, or the collection of language data for linguistic analysis. For this reason, field linguists often use software tools such as SIL's FLEx/FieldWorks (Rogers, 2010) to create digital lexica and collections of IGT that may serve as input for downstream applications or analyses and facilitate the creation of community-facing resources (Schreiner et al., 2020). To gloss a text using FieldWorks, the analyst separates each word into canonical morphemes, associating each one with a lexical or grammatical meaning.

While FieldWorks is perhaps the most widely used software for morphological annotation, other low-resource systems have successfully implemented rule-based finite state transducers (FST) for machine-aided development of IGT, digital lexica, and searchable corpora (Alnajjar et al., 2020; Kazeminejad et al., 2017; Arppe et al., 2016). While standard FSTs are limited in their ability to represent non-concatenative alternations and allomorphy, alignment-based transduction and two-level methods may be paired with probabilistic rule- or feature-based models to achieve higher performance using inflection tables or parallel texts (Hulden et al., 2014; Ahlberg et al., 2015; Palmer et al., 2010). Still other studies use IGT for morphological paradigm or grammar induction (Zamaraeva et al., 2019; Moeller et al., 2020). Since each of these methods assume pre-existing linguistic analyses of the data being processed, they may be suitable for later stages of annotation and resource development, but they run the risk of obscuring morphological patterns, hindering the discovery of important generalizations across word forms in the data early on in description and analysis.

For machine-assisted morphological annotation at the level of the unsegmented word, we draw on recent studies investigating low resource applications of neural sequence taggers for morphological analysis, POS tagging, and NER. While such models are known to require large amounts of consistently annotated data typically unavailable for under-described languages (Kann et al., 2020a), Garrette and Baldridge (2013) show that a POS tagger can be successfully trained with data annotated in as little as two hours when appropriate noise reduction techniques are applied. Experiments comparing model architectures suggest BiLSTMs with attention may be used for sequence tagging in low resource settings when combined with strategies for noise reduction and data augmentation, including character-level cross-domain and cross-lingual transfer methods (Adelani et al., 2021, 2020; Cotterell and Heigold, 2017; Hedderich et al., 2020), and data augmentation strategies involving external resources and collaborative curation (Adelani et al., 2021; Hedderich et al., 2021).

For linguistic fieldwork on languages that are not only under-resourced but also under-described, these methods may be complemented by unsupervised models to aid in the discovery of morphological phenomena and patterns that have yet to be documented or analyzed (Erdmann et al., 2020). For this reason, our proposed workflow utilizes unsupervised paradigm discovery methods to cluster and tag related word forms according to both lexeme and paradigm cell without the need for prior analysis or segmentation. Eventual implementation of an active learning component would allow for automated semi-supervised annotation to further increase efficiency and accuracy. Previous work on active learning for NER suggests the ideal model architecture may depend on the amount of data available for input (Erdmann et al., 2019). The modular nature of our proposed workflow would therefore allow for the option of interchanging models at different points within data collection and annotation. In summary, we believe WP-based annotation brings field linguistic representations closer to those used in the NLP community at large. This can speed up the early stages of the analytical process by enabling the use of unsupervised methods. Moreover, it enables relatively off-the-shelf adoption of new tagging models, rather than development of specific solutions for IGT.

### 2.3 Benefits for community engagement

In addition to allowing the fieldworker to begin annotation without committing to a segmentation-based analysis of the data, our proposed workflow aims to increase collaborative research with the language community by facilitating native speaker involvement in the annotation task. In conjunction with a growing focus on the ethical collection of linguistic data and collaborative fieldwork (Rice, 2006), we must remember that the diversity of language communities means there is no "one-size-fits-all" approach to ethical research or community engagement (van Driem, 2016). It is therefore important to position speakers as self-sufficient researchers of their own language by involving them at every step of the process, from data collection to analysis (Czaykowska-Higgins, 2009). Failure to engage with speakers of the language being studied can have far-reaching consequences.[2] We assume that the fieldworker is engaging with a community that desires resources such as dictionaries, grammars and educational materials. These are all the product of linguistic analysis. If the optimal outcome of community engagement is that community members have maximum agency in achieving their goals, lowering barriers to entry for non-specialists by providing more accessible tools and methods for analysis is one strategy for decreasing the community's reliance on outside specialists. Our proposal is only the first of many steps that would need to be taken to allow technology to facilitate such an outcome.

Existing fieldwork tools and methodologies may hinder community-based research by making data analysis difficult or inaccessible for native speakers of the language. One of the authors is involved in ongoing fieldwork on Wao Terero, a linguistic isolate spoken in the Ecuadorian Amazon. Education levels are low in rural Ecuador, and some of the native speakers involved in this project have less than a United States high school equivalent in formal schooling. The use of research tools that require extensive formal training or prior education in linguistic theory bars this subset of the community from fully participating in data analysis. Since these speakers are often older and monolingual, this can also skew scientific results, producing linguistic materials or analyses which may not represent the general Wao population.

Our proposed workflow directly contributes to the development of community-based research and linguistic tools. The relational WP approach, coupled with the concordance-based interface we propose, asks native speakers to identify patterns in the data by matching like with like, without requiring them to first learn technical vocabulary or a theory of morphemes. Section 4.1 presents the results of our collaboration with Wao Terero speakers to evaluate the potential of our workflow to increase community engagement in the field.

## 3 Workflow and experiments

We propose an annotation workflow (Figure 1) that begins with the collaborative collection of primary data within a language community. For morphological analysis, the model makes use of both naturalistic transcriptions or texts as well as a list of target lemmas for analysis. These files are used as input to an unsupervised model that identifies potential instances of lexemes and cells and outputs a sample of occurrences for each category to be annotated. The annotation process involves both excluding occurrences that don't belong in their assigned group and seeking out new occurrences to add to the group. The annotated output could then be used along with additional primary data as input to a supervised classifier within an active learning framework. Since the workflow is modular, each individual component can be updated over time in line with future advances in state of the art machine learning.

As a proof of concept that WP annotation is viable, we implement the first stage of our proposed workflow in which fieldworkers begin with the results of

[2] As one example, the ISO 639-3 codes used for identifying some languages, such as Wao Terero, Shuar (Chicham), and Ho-Chunk, are references to slurs for these communities. Since these codes are referenced by both the HTML and XML standards of the W3C, these communities cannot currently use the web in their language without reference to hate speech. A minimum of effort to engage with these communities could have avoided this.

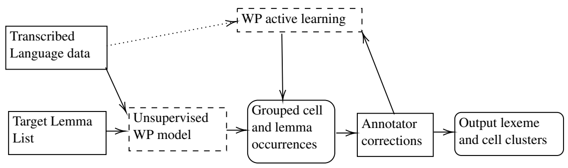
Transcribed Language data
WP active learning
Target Lemma List
Unsupervised WP model
Grouped cell and lemma occurrences
Annotator corrections
Output lexeme and cell clusters

Figure 1: The proposed annotation workflow

an unsupervised analysis. We conduct a series of pilot experiments demonstrating that annotators can rapidly improve the results of such a system. While our full proposed workflow would integrate active learning within an interactive annotation environment, in the present work, we do not currently use these annotations as additional input to the model.

### 3.1 Model

Our initial annotation system is the baseline paradigm discovery system from SIGMORPHON 2020 Task 2 (Jin et al., 2020). Given raw text and a set of target lemmas, the model uses edit trees and an unsupervised HMM to identify and return potential inflected forms of those lemmas. The original model has both a retrieval and a generation component - the retrieval component uses edit distance to find potential forms of the given lemmas and cluster them into paradigm cells, while the generation component produces potential forms of the lemmas which are not present in the text. Because fieldworkers and language consultants must see forms in context in order to make decisions about their morphological status, we apply only the first (retrieval) component to cluster forms which are attested in the raw text corpus. For instance, the model produces the following paradigm for the English word HEAR: *hear, heard, hearing, heart* (each occupying a different numerically labeled cell). This model's requirements determine the amount of raw text necessary for our proposed workflow. While Jin et al. (2020) requires a moderate amount of text, a less resource-hungry method could be substituted for very early fieldwork where little transcribed text is available.

We generate concordance-style datasets for the annotation of each proposed lexeme and cell using the examples of their proposed forms identified within the corpus by the unsupervised model (Figure 2). These datasets are generated separately for each individual lexeme and paradigm cell proposed by the model. For lexemes, we aim for up to 7 examples per form; for cells, we sample 20 instances in total. These numbers were selected based on initial estimates to account for known trade-offs between annotation speed and quality.

Our annotation guidelines set the following rules: The annotator should ensure all accepted examples in a set correspond to the same lexeme or morphophonological paradigm cell. Incorrect instances include forms that are derivationally related, homophonous, or belong to other lexical categories or (non-syncretic) paradigm cells.[3] All of our experiments targeted verbs, and the annotators were instructed to reject any form they did not believe was a verb. By presenting examples within a concordance format, the annotator is able to use word context to filter out forms which do not occur in the correct paradigm as well as tokens which are homophonous with a member of the paradigm. For instance, when annotating files for the verbal lexeme *hear*, the annotator would exclude both *heart* (incorrect paradigm) and the noun *hearing* (homophonous with the V.PTCP.PRS form).

Inspection of these datasets allows annotators to correct precision errors in the model's output (incorrect forms added to paradigms/cells) but not recall errors (missing forms). We therefore give annotators the opportunity to include missing items for each paradigm cell in the dataset by showing them a few examples of forms from each of the other model-proposed paradigm cells. Because the number of lexemes is much larger than the number of cells, we cannot augment the lexeme datasets in the same way. Instead, we give the annotator a tool which conducts a regular expression search in the corpus and adds up to 20 detected examples to the annotation dataset. For instance, the annotator could type *hear.?* to find additional forms similar to *hears* that were not originally captured by the model. The annotator may then apply the same process of comparison to provide additional positive examples for a downstream classifier.

### 3.2 Data

In order to evaluate the effect of a human annotator on model performance, we experiment on two languages, English and Croatian, with gold standard annotations from the Universal Dependencies data set. For English, we use the GUM treebank (Zeldes, 2017) and for Croatian the SETimes treebank (Agić

[3] Especially from a WP perspective, there is no strong theoretical reason for positing that paradigms may not span derivationally related items and different lexical categories (Bonami and Strnadová, 2019). These choices were made for the current study so that the annotators' decisions could be compared to a gold standard.

LEXEME

| annotator | | form | model output | |
|---|---|---|---|---|
| | . . . you're still going to | hear | True | them. |
| | She thought she could | hear | True | Gomez laughing. |
| X | . . . signalling of problems of | hearing | True | and understanding. |

CELL

| annotator | | form | model output | |
|---|---|---|---|---|
| | . . . mechanisms underlying the | learning | True | and processing of L2 grammar . . . |
| | . . . periods of limited . . . exposure | following | True | L2 training are not uncommon . . . |
| | . . . may be found in different situations | including | True | when one studies a language . . . |
| X | . . . such as listening and | reading | True | comprehension . . . |
| | The training | lasted | False | varying lengths of time. . . |

Figure 2: A selection of instances for annotation of the lexeme HEAR (top) and for a system-proposed morphological cell (bottom). Ellipses are for presentational purposes; the annotators saw full sentential contexts. The baseline's decision about the token is displayed as True/False in column 4, and the annotator marks an X in column 0 to indicate that they disagree.

and Ljubešić, 2015). In each case, we extract the entire treebank training file as raw text for model input. For the list of target lemmas, we select verbal lemmas with frequency ranks 10-111;[4] we skip the top 10 lemmas because they are the most likely to have atypical paradigms (Bybee, 1988), and the early stages of fieldwork should focus on identifying typical paradigm structure.

In each case, two annotators from a Linguistics Ph.D. program spent 30 minutes annotating lexeme data and 30 minutes on cell data. Experimenting on English, a language well known by the annotators, provides an upper estimate of model-plus-human annotator performance. The Croatian experiment provides a potentially more realistic example of model-plus-human performance on a language still relatively unknown to the fieldworker. While our annotators speak several Indo-European languages, neither of them is fluent in a Slavic language nor has ever studied Croatian.

It is relatively common to develop tools for endangered or under-resourced languages by applying them to small or unannotated datasets from well-resourced languages, since this allows for evaluation against a curated gold standard. However, well-resourced languages may differ typologically from real endangered languages, leading to poor generalization (Kann et al., 2019; Mager et al., 2018). Therefore, we conduct a third experiment on Wao Terero. The annotators for this experiment were native speakers of Wao Terero who have never taken a course in linguistics. Both were Spanish-Wao Terero bilinguals. One recently completed high school and the other has attended university courses. A Linguistics Ph.D. student who is currently conducting field research on Wao Terero but not a fluent speaker also performed annotations on the same data. To run the model, we use Wao Terero text of the New Testament as the raw text corpus. As the model did not perform well with single character verbs, the fieldworker specifically selected multi-syllable seed verbs for the model by searching for common inflectional endings with a regular expression search and compiled these into a list of 108 target lemmas. Six resulting lexeme files were removed since they had only two items and were potentially ambiguous. The annotators were provided approximately 10 minutes of instruction using Spanish verbal paradigms as examples before completing the task for Wao Terero. Instead of using the technical term "paradigm," the fieldworker described the concept as a collection of all the forms a verb might take while remaining the same word. Annotators were told that the goal was to assess the effectiveness of the annotation method, rather than tackling the issue of paradigm discovery directly. A guided practice preceded an hour of annotation. The guided practice featured the Spanish verb *cazar*, 'to hunt', with some errors that involved the homophone *casarse*, 'to marry', and a lexical-category-altering derivation *cazador*, 'hunter'. Speakers were allowed to ask any questions they wished during annotation and additionally provide clarification to one another. A request was made that they limit their communication among themselves so that it would be possible

[4] This selection procedure is not entirely unsupervised since we are guaranteed to select verbs as targets and our assessment of their frequency counts all their actual forms. However, it is similar to the task setup used in SIGMORPHON 2020, which also used only verbal targets; we believed that keeping this simplification from the shared task would help the baseline model to perform as reported.

to compare their annotation choices. They were given a directory with 102 annotation files named in a numerical order corresponding to the alphabetical order of stems, and instructed to annotate files in order until the hour was complete. Since we do not have an independently developed gold standard for Wao Terero, we focus our evaluation of the results of this experiment on measurements of annotation speed and additional qualitative observations.

### 3.3 Evaluation

For English and Croatian, we measure annotation accuracy at the token level. As in many unsupervised applications, this requires a preliminary mapping, since some of the model's proposed lexemes or cells may mix together multiple actual cells, from which the annotator must try to select one. We find the most likely interpretation of a set of forms by taking the most common gold label among its accepted instances. If a lexeme file contains *hear.V, hear.V, hear.V, hearing.N, heart.N*, the best interpretation is *hear.V*; If the annotator accepts examples 1, 2, and 4, they have 2/3 correct acceptances and 1/2 correct rejections, for an accuracy of 3/5. Annotator accuracies are micro-averaged across the dataset. Given the imposed time limits of the experiment, annotators did not inspect every annotation file output by the model. Final scores reflect the entire dataset; cases the annotator did not reach are left at their baseline values.[5]

## 4 Results

Table 3 shows lexeme evaluations before considering regular expression search results. Annotation in English is much faster than in Croatian. English annotators inspected an average of 444 lexeme tokens, including regex search results, in their 30 minutes, while Croatian annotators inspected 306. This is expected, since Croatian was selected to simulate the early stages of fieldwork in which the linguist is still relatively unfamiliar with the language being analyzed. However, annotators for each language are capable of reliably rejecting incorrect forms proposed by the model. Annotator mistakes on the English data tend to involve confusion between adjectival and verbal interpretations of forms like *leading*.

| | Acc. | Marked | Corr. |
|---|---|---|---|
| English | | | |
| Base | 81 | - | - |
| A1 | 84 | 58 | 50 |
| A2 | 83 | 43 | 33 |
| Croatian | | | |
| Base | 66 | - | - |
| A3 | 67 | 19 | 19 |
| A4 | 66 | 12 | 12 |

Table 3: Evaluation of lexeme annotations. Marked shows a count of instances altered from the baseline by annotators; Corr. shows a count of correct alterations.

| | Acc. | Marked | Corr. |
|---|---|---|---|
| English | | | |
| Base | 67 | - | - |
| A1 | 97 | 129 | 120 |
| A2 | 94 | 119 | 108 |
| Croatian | | | |
| Base | 90 | - | - |
| A3 | 90 | 8 | -1 |
| A4 | 90 | 28 | 16 |

Table 4: Evaluation of cell annotations.

Table 4 shows cell evaluations. Annotation of cells in English is rapid and highly accurate; the English annotators were able to review all proposed cells in 30 minutes. The English baseline produces several candidate paradigm cells containing mostly function words or other non-verbal material. These are easily rejected, as are spurious members of real cells. Annotation in Croatian is slower and comparatively more error-prone. Annotators reviewed an average of 1384 items in 30 minutes, but without marking many forms. However, even with only 30 minutes, one annotator did contribute useful information on cell membership in Croatian.

Regular expression search did not contribute usefully to the results. While all four annotators labeled the results of their searches with high accuracy ($\geq$ 80), it seems to have been too difficult to write good regular expressions that would elicit valid but undetected paradigm members. In English, annotators found 6 and 4 correct novel forms; in Croatian, 0 and 1. We believe an interactive environment for search and annotation could be more effective, a point we return to below.

[5] We follow the SIGMORPHON 2020 task in grouping syncretic cells for evaluation. Thus, English has 5 valid paradigm cells (e.g., for the lexeme *show*: nonpast *show*, nonpast 3rd person *shows*, gerund/present participle *showing*, past participle *shown* and past active *showed*).

### 4.1 Wao Terero results

The Wao Terero speaking annotators each made an assessment for all items in 4 files, constituting 67 tokens, in one hour. The linguist assessed 776 tokens. The speakers rejected 11 and 9 items respectively, agreeing on 3 items. The fieldworker rejected 15 items, agreeing with the speakers 6 and 5 times respectively. All annotators agreed on rejections for a total of 2 items. Notably, in one file consisting of four items, the linguist's annotations were the complement of one of the speaker's. This indicates that the file was ambiguous between two lexeme options and that finding ways to address ambiguity would increase annotator agreement. For instance, a specification of heuristics might be used. In this case simply choosing the option that resulted in the fewest rejections would have been adequate.

The difference in the number of items annotated by the speakers as compared to the fieldworker reflects the different approaches taken when completing the annotation task. Specifically, speakers attempted to understand the sentences that provided context for the words in question. Because the Bible constitutes atypical Wao Terero, one speaker complained that the data contained non-words.[6] This complicated the task for both speakers and may explain the low number of tokens assessed. The linguist instead assessed tokens based on orthographic regularities. Strictly speaking, this was exactly what the Wao Terero speakers were asked not to do, since they were provided homophones as examples of items that should be rejected.

In response to exit questions following the annotation task, the speakers stated that they had understood the task and its goals. One speaker stated that this type of investigation is valuable. Neither speaker claimed to find the task dull. One indicated that they didn't find it difficult except for when they had initially started. The other stated that it was very difficult because of unusual words. That is, one answered the question based on the conceptual difficulty of the task while the other answered according to practical issues with the data.

Despite issues with the data and what might be considered an initially slow annotation pace, our claim that speakers would find the task intuitive was borne out. Considering that existing IGT workflows require a great deal of specialist knowledge, the fact that speakers with no linguistic training can begin annotating using a WP-based workflow with only 10 minutes of training is notable.

## 5 Discussion

A more relational, word-based approach to morphological annotation for language documentation is desirable for both theoretical and practical reasons. Theoretically, Word-and-Paradigm annotation allows fieldworkers to avoid, or at least defer, difficult decisions about both morphological form and function. Our proposal separates the identification of a morphological cell from the application of a morphosyntactic label or set of features, a difficult analytical task often involving comparison of many related examples. For instance, distinguishing past tense from perfect aspect is a sensitive task (Bybee and Dahl, 1989) which might best be done once the forms in question can be reliably separated from the rest of the verbal paradigm. Thus, even if the desired end goal of annotation is IGT, we believe that our proposed annotation methodology can speed up the early stages of annotation and prevent the researcher from having to commit to an analysis too early.

From a practical standpoint, by starting with an automatically generated concordance set for each proposed cell and lexeme, the annotator can focus on making direct comparisons – is the proposed grouping coherent in terms of its surface form and distributional context? Annotating at the level of unsegmented words in context makes it possible for a native speaker who has no knowledge of technical or grammatical terms to easily identify the patterns represented by the concordances and contribute their expert knowledge. The proposed workflow is therefore designed to facilitate greater community involvement in the development of language resources and technologies in line with community needs. Our experiment results show that even in an unfamiliar language for which the

[6] The New Testament is not the optimal corpus for Wao Terero but there was little other choice. The language of New Testiment translations can be atypical and stylized, even in English. Given the distance between Mediterranean and Amazonian cultures the translation of the New Testiment into Wao Terero is filled with neologisms, unfamiliar concepts and atypical phrasal constructions. Although there is a sizable Wao Terero deposit at the Endangered Language Archive (ELAR), the orthography of that collection is inconsistent and the restrictions placed on its use do not allow for practical use by researchers or Wao community members. One of the authors is currently involved in an effort to create an open corpus. This is in line with the wishes of Wao speakers, who have no access to ELAR materials. Unfortunately, this alternative corpus is still under development and the only suitably large corpus with a consistent orthography in existences is the New Testament.

annotator does not yet understand the functions of different morphological markers, there is still some capacity to weed out forms that do not belong. The tight integration of this workflow with unsupervised learning technology also means that any future improvements in unsupervised paradigm discovery can immediately benefit fieldworkers.

While our pilot experiment focuses on the initial steps of a computationally-aided field documentation project, our full proposed workflow envisions both an interactive environment and active learning. An interactive environment would allow annotators to view proposed paradigm tables alongside the text, helping to see where forms might be missing. Annotators could also search more effectively for forms that could fill in these gaps, for instance by viewing a word cloud of similar forms. This could make it easier to fix recall errors in system-proposed paradigms. Even a few labeled instances can vastly improve the performance of part of speech taggers (Stratos and Collins, 2015; Søgaard, 2010), and the same is true in a relational setting (Li et al., 2020). In our environment, we envision active learning directing the annotator's attention to the least certain distinctions, eliminating repetitive annotation of "easy" instances.

Our proposal shows promise as a faster and more theoretically grounded alternative to existing tools. Its modular structure makes it easy to integrate advances in the field of computational linguistics, and it allows the fieldworker to quickly and easily involve the intuitions of native speakers with little linguistic training, boosting community engagement.

## 6 Acknowledgements

We would like to thank Flora and Alberto Boyotai, the Wao Terero speakers who tested the annotation system and provided us with helpful feedback. A portion of this research was funded by the Lewis and Clark Fund for Exploration and Field Research from the American Philosophical Society, and by the Labex EFL.

## References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. 2021. Masakhaner: Named entity recognition for african languages. *arXiv preprint arXiv:2103.11811.*

David Ifeoluwa Adelani, Michael A Hedderich, Dawei Zhu, Esther van den Berg, and Dietrich Klakow. 2020. Distant supervision and noisy label learning for low resource named entity recognition: A study on Hausa and Yorùbá. *arXiv preprint arXiv:2003.08370.*

Željko Agić and Nikola Ljubešić. 2015. Universal Dependencies for Croatian (that work for Serbian, too). In *The 5th Workshop on Balto-Slavic Natural Language Processing*, pages 1–8, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

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

Khalid Alnajjar, Mika Hämäläinen, Jack Rueter, and Niko Partanen. 2020. Ve'rdd. narrowing the gap between paper dictionaries, low-resource nlp and community involvement. *arXiv preprint arXiv:2012.02578.*

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 984–996, Hong Kong, China. Association for Computational Linguistics.

Antti Arppe, Jordan Lachler, Trond Trosterud, Lene Antonsen, and Sjur N Moshagen. 2016. Basic language resource kits for endangered languages: A case study of Plains Cree. In *Proceedings of the 2nd Workshop on Collaboration and Computing for Under-Resourced Languages Workshop (CCURL 2016), Portorož, Slovenia*, pages 1–8.

James P. Blevins. 2016. *Word and paradigm morphology*. Oxford University Press.

Olivier Bonami and Jana Strnadová. 2019. Paradigm structure and predictability in derivational morphology. *Morphology*, 29.

Joan Bybee. 1988. Morphology as lexical organization. In Michael Hammond and Michael Noonan, editors, *Theoretical morphology: Approaches in modern linguistics*, pages 119–142. Academic Press.

Joan L Bybee and Östen Dahl. 1989. *The creation of tense and aspect systems in the languages of the world*. John Benjamins Amsterdam.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual character-level neural morphological tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759, Copenhagen, Denmark. Association for Computational Linguistics.

Ewa Czaykowska-Higgins. 2009. Research models, community engagement, and linguistic fieldwork. *Language Documentation & Conservation*, 3(1):15–50.

Micha Elsner, Andrea D. Sims, Alexander Erdmann, Antonio Hernandez, Evan Jaffe, Lifeng Jin, Martha Booker Johnson, Shuan Karim, David L. King, Luana Lamberti Nunes, Byung-Doh Oh, Nathan Rasmussen, Cory Shain, Stephanie Antetomaso, Kendra V. Dickinson, Noah Diewald, Michelle McKenzie, and Symon Stevens-Guille. 2019. Modeling morphological learning, typology, and change: What can the neural sequence-to-sequence framework contribute? *Journal of Language Modelling*, 7(1):125–170.

Alexander Erdmann, Micha Elsner, Shijie Wu, Ryan Cotterell, and Nizar Habash. 2020. The paradigm discovery problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7778–7790, Online. Association for Computational Linguistics.

Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen-Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice Joyeux-Prunel, and Marie-Catherine de Marneffe. 2019. Practical, efficient, and customizable active learning for named entity recognition in the digital humanities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2223–2234, Minneapolis, Minnesota. Association for Computational Linguistics.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 138–147, Atlanta, Georgia. Association for Computational Linguistics.

Michael A. Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. 2020. Transfer learning and distant supervision for multilingual transformer models: A study on African languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2580–2591, Online. Association for Computational Linguistics.

Michael A. Hedderich, Lukas Lange, and Dietrich Klakow. 2021. Anea: Distant supervision for low-resource named entity recognition.

Nikolaus P Himmelmann. 1998. Documentary and descriptive linguistics.

Charles F. Hockett. 1954. Two models of grammatical description. *Word*, 10:210–234.

Mans Hulden, Markus Forsberg, and Malin Ahlberg. 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*, pages 569–578, Gothenburg, Sweden. Association for Computational Linguistics.

Huiming Jin, Liwei Cai, Yihui Peng, Chen Xia, Arya McCarthy, and Katharina Kann. 2020. Unsupervised morphological paradigm completion. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6696–6707, Online. Association for Computational Linguistics.

Katharina Kann, Kyunghyun Cho, and Samuel R. Bowman. 2019. Towards realistic practices in low-resource natural language processing: The development set. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3342–3349, Hong Kong, China. Association for Computational Linguistics.

Katharina Kann, Ophélie Lacroix, and Anders Søgaard. 2020a. Weakly supervised pos taggers perform poorly on truly low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8066–8073.

Katharina Kann, Arya D. McCarthy, Garrett Nicolai, and Mans Hulden. 2020b. The SIGMORPHON 2020 shared task on unsupervised morphological paradigm completion. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 51–62, Online. Association for Computational Linguistics.

Katharina Kann and Hinrich Schütze. 2016. Single-model encoder-decoder with explicit morphological representation for reinflection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 555–560, Berlin, Germany. Association for Computational Linguistics.

Ghazaleh Kazeminejad, Andrew Cowell, and Mans Hulden. 2017. Creating lexical resources for polysynthetic languages—the case of Arapaho. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 10–18, Honolulu. Association for Computational Linguistics.

Pengshuai Li, Xinsong Zhang, Weijia Jia, and Wei Zhao. 2020. Active testing: An unbiased evaluation method for distantly supervised relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 204–211, Online. Association for Computational Linguistics.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sarah Moeller, Ling Liu, Changbing Yang, Katharina Kann, and Mans Hulden. 2020. IGT2P: From interlinear glossed texts to paradigms. In *Proceedings of the*

*2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5251–5262, Online. Association for Computational Linguistics.

Alexis Palmer, Taesun Moon, Jason Baldridge, Katrin Erk, Eric Campbell, and Telma Can. 2010. Computational strategies for reducing annotation effort in language documentation. *Linguistic Issues in Language Technology*, 3(4):1–42.

Keren Rice. 2006. Ethical issues in linguistic fieldwork. *Journal of Academic Ethics*, 4:123–155.

Chris Rogers. 2010. Review of fieldworks language explorer (flex) 3.0. *Language Documentation & Conservation*, 4:78–84.

Sylvia LR Schreiner, Lane Schwartz, Benjamin Hunt, and Emily Chen. 2020. Multidirectional leveraging for computational morphology and language documentation and revitalization. *Language documentation and conservation*, 14.

Miikka Silfverberg and Mans Hulden. 2018. An encoder-decoder approach to the paradigm cell filling problem. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2883–2889, Brussels, Belgium. Association for Computational Linguistics.

Anders Søgaard. 2010. Simple semi-supervised training of part-of-speech taggers. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 205–208, Uppsala, Sweden. Association for Computational Linguistics.

Karl Stratos and Michael Collins. 2015. Simple semi-supervised POS tagging. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 79–87, Denver, Colorado. Association for Computational Linguistics.

Nicholas Thieberger and Andrea L Berez. 2012. Linguistic data management. Oxford University Press.

George van Driem. 2016. Endangered language research and the moral depravity of ethics protocols. 10:243–252.

Adam Wiemerslage, Arya D. McCarthy, Alexander Erdmann, Garrett Nicolai, Manex Agirrezabal, Miikka Silfverberg, Mans Hulden, and Katharina Kann. 2021. Findings of the SIGMORPHON 2021 shared task on unsupervised morphological paradigm clustering. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 72–81, Online. Association for Computational Linguistics.

Olga Zamaraeva. 2016. Inferring morphotactics from interlinear glossed text: Combining clustering and precision grammars. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 141–150, Berlin, Germany. Association for Computational Linguistics.

Olga Zamaraeva, Kristen Howell, and Emily M. Bender. 2019. Handling cross-cutting properties in automatic inference of lexical classes: A case study of chintang. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 28–38, Honolulu. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.