COMPUTATIONAL MORPHOLOGY
FOR LANGUAGE DESCRIPTION AND DOCUMENTATION

SARAH MOELLER

University of Colorado Boulder

While the field of linguistics has slowly but surely widened the world's knowledge about human language, thanks in part to the recent emphasis on language documentation and description of under-documented languages, computational linguistics has barely expanded beyond a handful of economically or politically powerful languages. This paper is a synthesis of natural language processing (NLP) models and methods and a history about how the models and methods have been applied to the study of morphological structure, particularly in low-resource languages (LRL). The paper assumes that the study of morphology has an important role to play in both NLP and linguistics. It explores the potential for discovering newer and more efficient methods while training computational morphological models on data produced during language documentation and description (LDD) field projects.[1]

Keywords: language documentation, natural language processing, NLP, low-resource languages, machine learning

1. INTRODUCTION

Morphology comprises word-building properties in human languages and their accompanying (morpho-)syntactic phenomena. Historically, computational linguists and "paper-and-pencil linguists" have taken different and sometimes seemingly incompatible approaches to morphology (Karttunen & Beesley 2005). Yet, despite their out-of-sync approaches, both computational linguistics and "traditional" linguistics benefit from morphological analysis (Cotterell et al. 2015). For natural language processing (NLP), work with low-resource languages (LRL) is still largely uncharted territory. This paper explores the limited work in morphology by asking this question: "What [computational] methods...can detect [morphological] structure in small, noisy data sets, while being directly applicable to a wide variety of languages?" (Bird 2009).

The paper is organized as follows. Section 2 describes the workflow and activities of LDD. Section 3 sketches the history of NLP work with LRL. Section 4 defines morphological analysis and looks specifically at NLP applied to morpheme segmentation and glossing and Section 5 looks at the application to learning morphological inflectional paradigmatic patterns.

2. LANGUAGE DOCUMENTATION AND DESCRIPTION

In linguistics, morphological description of a broad range of languages is a foundational step towards any reasonable linguistic theory. A focus of documenting and describing under-documented languages, including their morphological structure, has been emphasized since the 1990's along with the development of language documentation as a distinct subfield. Himmelmann (1998) defines language documentation as "a comprehensive and representative sample of communicative events [that are] as natural as possible." Woodbury (2003) defines it similarly as "comprehensive and transparent records supporting wide ranging scientific investigations of the language." Language description can be defined as work that analyzes language documentation to create "systematic presentations of the phonology, morphology, syntax, and semantics of the language" (Bird & Chiang 2012). The emphasis on endangered languages over the past three decades has established best practices documenting a new language (Bowern 2008; Czaykowska-Higgins 2009; Lupke 2010; Vallejos 2014; Rice & Thunder 2017). However, the specific activities that divide the two subfields are not rigid. Therefore, the current work generally refers to them together as "language documentation and description (LDD)" or "documentary and descriptive linguistics".

The workflow of LDD is not standardized, although most projects seem to follow a similar sequence. A version of one common sequence (Bird & Chiang 2012) is given below (the numbers are used to refer to each task, e.g., "task 2a" refers to transcription). Each subsequent task progressively encompasses more description than documentation, except archiving, which comes strictly under language documentation but is logically a last step.

(1) Collect (audio/video recordings) naturally occurring speech

(2) a) Transcribe and b) translate

(3) Perform basic morphosyntactic analysis by segmenting the morphemes and creating morphological glosses and/or a lexicon

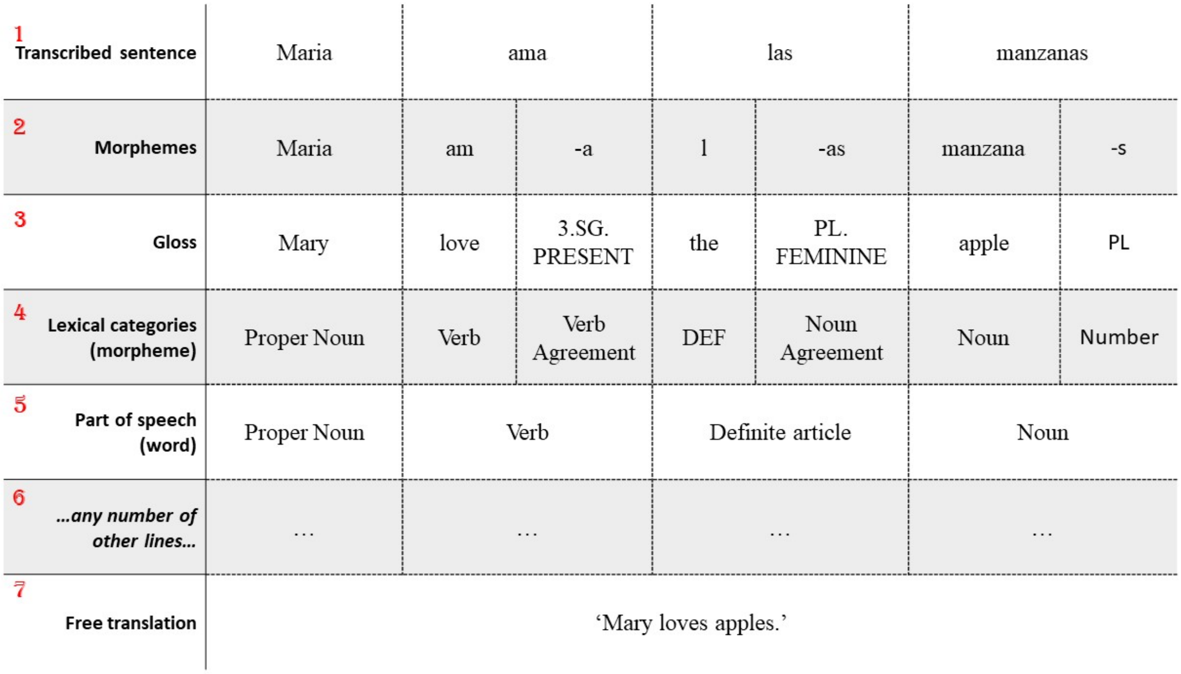
(4) Elicit morphological paradigms that reveal underlying patterns

(5) Prepare descriptive reports that outline the language's structure

(6) Archive data in a long-term repository

One primary output of this workflow is interlinear glossed texts (IGT), a data format distinctive to linguistics (Figure 1). Interlinearization is the primary task after transcription. It moves the workflow beyond simple documentation but still serves as a "preprocessing step" to language description (strictly defined) (Moon, Erk & Baldridge 2009). It comprises annotation tasks that enrich the data with analytic information added as lines under the transcribed text (task 1 in above workflow; line 1 in Figure 1). The most common lines are shown in Figure 1. Lines can be added in any order, but translations (task 2a; line 7) morpheme boundaries and morpheme glosses (task 3; lines 2 and 3, respectively) are usually added first. Doing more annotation (e.g. lines 4-6) often happens in field projects, but translation, morpheme segmentation, and morpheme glossing are usually given the highest priority.

FIGURE 1. INTERLINEARIZATION:
INTERLINEAR GLOSSED TEXTS ADD LINES OF ANNOTATION TO THE ORIGINAL TEXT.

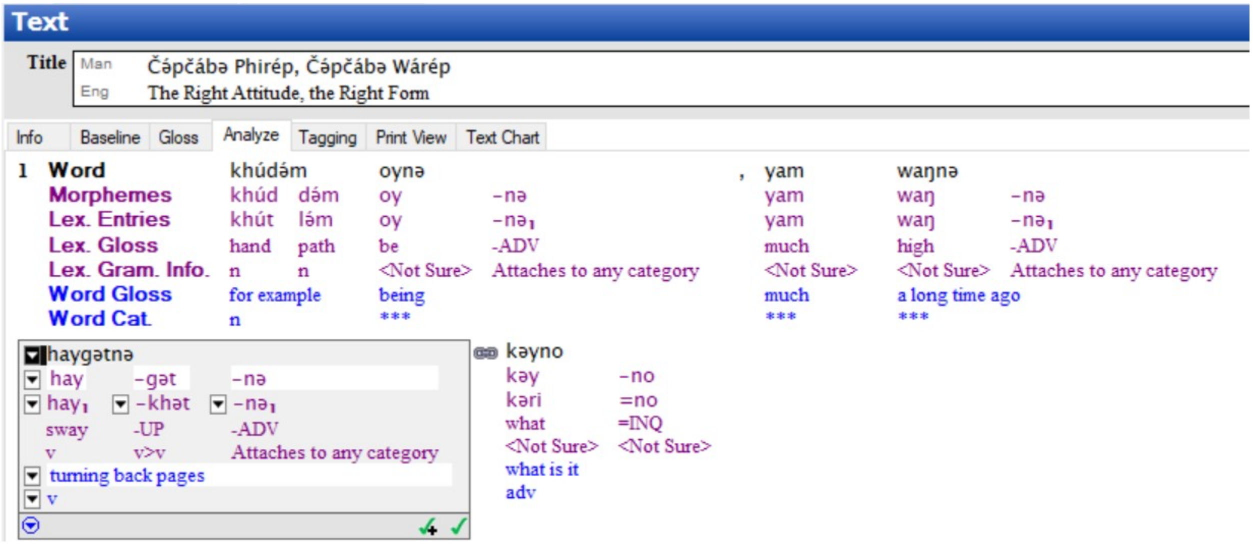


1 Transcribed sentence	Maria	ama		las		manzanas	
2 Morphemes	Maria	am	-a	l	-as	manzana	-s
3 Gloss	Mary	love	3.SG. PRESENT	the	PL. FEMININE	apple	PL
4 Lexical categories (morpheme)	Proper Noun	Verb	Verb Agreement	DEF	Noun Agreement	Noun	Number
5 Part of speech (word)	Proper Noun	Verb		Definite article		Noun	
6 *...any number of other lines...*	…	…		…		…	
7 Free translation	'Mary loves apples.'						

Interlinearizing data uncovers the rarer and unique linguistic phenomena. Interlinearization opens the door for deeper linguistic analysis and lays the foundation for reference grammars, dictionaries, and language learning materials, but interlinearization is not sufficient to create complete grammars, dictionaries, etc. One additional descriptive task is often included: the collection of morphological inflection patterns, or paradigms, for several lemmata (task 4). Inflectional paradigms are elicited because complete paradigms are rarely found in natural language. Complete paradigms are needed to infer general rules of inflection.

Without translations, morpheme segmentation, and glossing, the data is understandable only to someone who already speaks the language. If no speakers are left, the data is mostly inaccessible, much like Egyptian hieroglyphics before the Rosetta Stone was discovered. A few specially designed software tools provide limited automated assistance. The two most popular are ELAN (Auer et al. 2010) and FLEx (Rogers 2010). Examples of their interlinearization interfaces are shown in Figures 2 and 3. These tools implement hand-constructed, rule-based computational morphological parsers but rule-based parsers do not generalize to new data. FLEx also copies morpheme boundaries and glosses onto other words if they are identical to words that were previously annotated by hand. Neither tool incorporates machine learning.

FIGURE 2. USER INTERFACE FOR INTERLINEARIZATION
IN FIELDWORKS LANGUAGE EXPLORER (FLEX) DISPLAYING A MANIPURI [MNI] TEXT

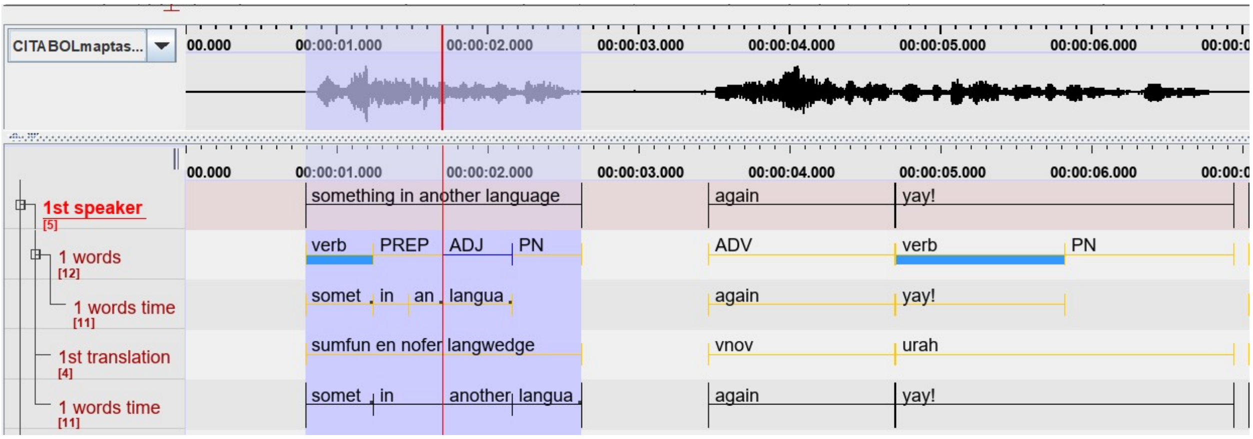


Title
Man Čəpčábə Phirép, Čəpčábə Wárép
Eng The Right Attitude, the Right Form

1 **Word**	khúdəm		oynə		,	yam	waŋnə	
Morphemes	khúd	dəm	oy	-nə		yam	waŋ	-nə
Lex. Entries	khút	ləm	oy	-nə$_1$		yam	waŋ	-nə$_1$
Lex. Gloss	hand	path	be	-ADV		much	high	-ADV
Lex. Gram. Info.	n	n	<Not Sure>	Attaches to any category		<Not Sure>	<Not Sure>	Attaches to any category
Word Gloss	for example		being			much	a long time ago	
Word Cat.	n		***			***	***	

haygətnə			kəyno	
hay	-gət	-nə	kəy	-no
hay$_1$	-khət	-nə$_1$	kəri	=no
sway	-UP	-ADV	what	=INQ
v	v>v	Attaches to any category	<Not Sure>	<Not Sure>
turning back pages			what is it	
v			adv	

FIGURE 3. USER INTERFACE FOR INTERLINEARIZATION
IN ELAN SHOWING A PRACTICE SESSION IN ENGLISH



3. NATURAL LANGUAGE PROCESSING (NLP) FOR LOW-RESOURCE LANGUAGES

Though the line between documentation and description may not be clear, one thing is clear: current methods cannot easily process large amounts of data. Most archived corpora are only partly annotated because funding and time constraints do not allow complete interlinearization (Cox, Bouliame & Alam 2019). Methods currently that are today used commonly in LDD rely primarily on hand annotation which is extremely inefficient. The typical strategy of annotating texts from top to bottom is non-optimal for training a supervised machine learning model (Baldridge & Osborne 2008; Baldridge & Palmer 2009; Palmer 2009). Since naturally occurring speech contains many repeated linguistic structures, manual annotation has been described as repetitive, monotonous, costly, and time-consuming (Duong 2017; He et al. 2016). It can take anywhere from 20 to 100 hours to transcribe (task 2a) a single hour of speech (Seifart et al. 2018) and it is reasonable to assume that interlinearization (tasks 2b and 3) and eliciting morphological paradigms (task 4) require significantly more time.

A recent growth of NLP interest in low-resource languages (LRL) has brought machine learning models and methods that achieve good results even with LDD field data. A notable example is ELPIS (Foley et al. 2018), an online tool that includes a user interface accessible to those with no programming background. Machine translation (MT) has also been applied to documentary data, using the output of an automatic speech recognition system as input to the MT system (Anastasopoulos, Chiang & Duong 2016; Duong et al. 2016).

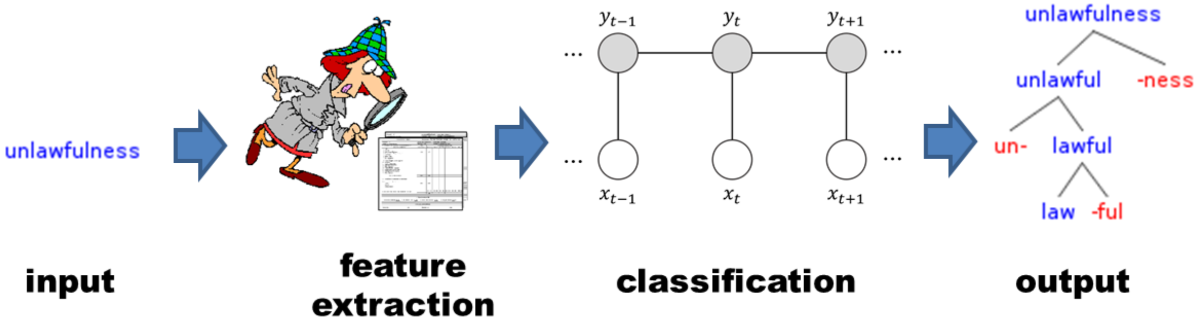
The potential for machine learning to perform morphological analysis during interlinearization has been clearly demonstrated (Baldridge & Palmer 2009; Palmer 2009; Palmer et al. 2010; Xia et al. 2016). For example, Felt (2012) found that when a round of annotation is done automatically by a machine learning model and then corrected by the human annotators, the annotators' accuracy is improve if the machine learning model achieves at least 60% accuracy and significantly speeds manual annotation if it achieves an accuracy of 80%. In the area of morphological paradigm learning, the annual SIGMORPHON and CoNLL-SIGMORPHON shared tasks (Cotterell et al. 2016; Cotterell et al. 2017; Cotterell et al. 2018; McCarthy et al. 2019; Nicolai, Gorman & Cotterell 2020) have developed successful methods with limited training data.

Although NLP interest in LRL has grown noticeably in the past few years, it is not a new area of research. Since the late 20th century, NLP has taken several approaches to low-resource languages that can be classified as either rule-based (i.e., finite state transducers) (e.g., Cotterell et al. 2015; Forsberg & Hulden 2016; Moeller et al. 2018; Moeller et al. 2019) or machine learning models that "learn" rules from data. Machine learning approaches to LRL can be classified according to whether the training data was annotated completely (supervised) (e.g., Bergmanis et al. 2017; Sudhakar & Singh 2017; Makarov, Ruzsics & Clematide 2017; Liu et al. 2018; Makarov & Clematide 2018a), partially (semi-supervised) (e.g., Ahlberg, Forsberg & Hulden 2014), or not at all (unsupervised) (e.g., Moon, Erk & Baldridge 2009; Palmer et al. 2010; Kirschenbaum, Wittenburg & Heyer 2012; Soricut & Och 2015).

At first glance, unsupervised and semi-supervised learning seem most promising for LDD because they do not require as much manually annotated data. Supervised learning is trained on "gold standard" annotated data. However, even though supervised learning requires annotation, it needs much less data than unsupervised learning and almost always yields better results (Ruokolainen et al. 2013; Cotterell et al. 2015). Additionally, without annotated labels, unsupervised learning can only really cluster data by the latent patterns in the data. Discovering latent patterns might be quite useful for linguists when first exploring the data; for example, frequent character patterns and substrings that a model discovers could provide an initial hypothesis to the linguist about the language's morphological structure. However, no matter how accurate an unsupervised model may be, it cannot substitute the valuable process of manually analyzing and discovering patterns in the data. Detailed analysis of new data is vital for linguists because through that process the linguist becomes familiar with the data and begins to absorb an intuitive knowledge of the language. Nevertheless, the latent patterns discovered by unsupervised models can have many uses such as being leveraged in a semi-supervised approach. Semi-supervised learning combines some supervised data with a larger set of unsupervised data (Kohonen, Virpioja & Lagus 2010; Poon, Cherry & Toutanova 2009). This approach is suitable if available annotated data is not adequate to effectively train a supervised model and it may be ideal for LDD because having substantial amounts of unannotated data with a small amount of annotated data is a common situation. Unfortunately, real applications of semi-supervised learning, specifically for computational morphology, are relatively rare, particularly with neural networks. There are exceptions, such as Ahlberg et al. (2014), where semi-supervised learning was used to induce morphological paradigms in low-resource settings.

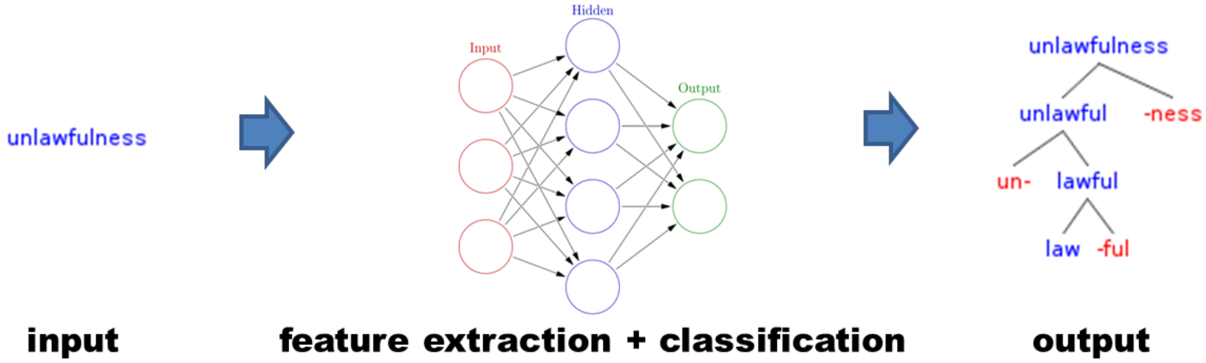
Until the 2010s, most machine learning models were feature-based with hand-designed features, illustrated in Figure 4. The input would be a hand-designed feature function that for morphological analysis might include 1) the whole word, 2) the position of the word in the sentence, 3) surrounding words or morphemes, 4) the POS tag of the previous morpheme/word. Features are assigned weights by the model during training to achieve optimal performance according to some objective function such as classification accuracy. For example, in a morpheme segmentation task where one chosen feature is the previous word and the previous word is some form of the English “to be” verb, and the target word ends in “ing”, then the model might give a high weight to the previous word so that the model pays attention to it when deciding how to segment a word ending in “ing”. The performance of feature-based models, such as Conditional Random Fields (CRF) and Support Vector Machines (SVM), relies heavily on the manual choice of features. This could be a drawback for under-described languages, because if little linguistic description is available, how does one know which features are optimal for that language? Fortunately, some feature-based models have been shown to perform reasonably well using language-independent features such as length of word or placement of letter in word (Ruokolainen et al. 2016; Moeller & Hulden 2018).

FIGURE 4. FEATURE-BASED MACHINE LEARNING REQUIRES A HUMAN TO IDENTIFY AND EXTRACT FEATURES THAT A FEATURE-BASED CLASSIFICATION MODEL SUCH AS A CRF USES TO PROVIDE THE CORRECT OUTPUT



Currently, neural networks models, or deep learning models, are dominating NLP (Goldberg 2017). Even though they outperform older, feature-based models on almost all tasks, they did not become popular until the mid-2010s because they require greater computing power and, for some tasks, train more slowly (Cotterell & Heigold 2017). Neural networks, illustrated in Figure 5, refers to a family of supervised machine learning models that are composed of layers of statistical units. The layers essentially substitute the feature engineering needed in non-neural machine learning. Multiple embedded layers allow the model to look at an exponential number of "semantically" neighboring instances of each training instance it encounters (Bengio et al. 2003). The layers create intermediate representations of the data that allow the model to "learn" a distributed representation of elements within each instance (e.g., a distributed representation of words within a sentence). This ability of the model to learn requires no (or at most, quite simple) manual feature design. Each unit in each layer is connected to each unit in the adjacent layers. Vector representations of the data are received by an input layer and transformed in "hidden" layers. The hidden layers feed into a final logistic function layer (i.e., softmax) that outputs a prediction of each possible class as a probability between 0 and 1. The connections between layers are represented by learnable weights; the higher the weight the more influence a unit has on the result. Since deep learning is supervised the weights are adjusted with feedback from the gold standard.[2] This is done via stochastic gradient descent or some similar optimization algorithm (Goldberg 2017) and backpropagation, which tells the model how to change the parameters which build the representation of each layer from the previous layer (LeCun, Bengio & Hinton 2015).

FIGURE 5. NEURAL NETWORKS, OR DEEP LEARNING, MODELS LEARN WHAT FEATURES IN THE DATA ARE IMPORTANT FOR GIVING THE CORRECT OUTPUT



Until recently, neural networks had the same great disadvantage that unsupervised learning has – superior performance required a great deal of data. Data from LDD would have been considered inadequate to train neural networks (Duong 2017). Even now, a non-neural model can outperform any given neural model that is not tuned to low-resource settings and neural models can be difficult to optimize and tune for low-resource settings (Popel & Bojar 2018).

New methods are being explored to overcome neural models' dependence on large corpora. Examples include fine-tuning a model to the specific task and input data, training intermediate steps, or augmenting the training data. van Biljon et al. (2020) looked at fine-tuning a model and determined that shallow- or medium-depth size Transformer models, for example only 3 encoder and 3 decoder layers, give better results with limited training data. An example of an intermediate training step would be first training a segmentation model to produce surface segments (morphs) and from them to learn underlying forms of morphemes (e.g., "impossible" → "in-possible" → "neg-possible") (Cotterell, Vieira & Schütze 2016; Liu et al. 2018; Moeller et al. 2019). The third successful method is augmenting training data. Augmentation can be done with artificial word forms (Liu et al. 2018) or with information extracted from other resources such as grammars and dictionaries. These are just a few of techniques that have been investigate; there are probably many more that we have not yet discovered.

Although NLP research in LRL has been growing since the mid-2010's, very little of it has been applied to linguistic on under-documented languages. One exception is the AGGREGATION project (Bender 2014) which has used IGT to automatically infer grammatical structure for multiple languages (Lepp, Zamaraeva & Bender 2019; Wax 2014). Much of their data comes from the Online Database of INterlinear Text (Lewis & Xia 2010, ODIN) which is a collection extracted from published linguistic articles or books. These IGT excerpts differ from IGTs produced by field linguists in at least one important way. Noise (i.e., typos, inconsistencies, etc.) is generally removed before publication, so that ODIN does not have the level of noise that field IGT does which simplifies pre-processing and does not distract machine learning models with spurious patterns.

4. MORPHOLOGICAL ANALYSIS

Morphological analysis is a key activity in LDD. Morphological analysis is particularly important when working with morphologically complex languages. Languages that build words from multiple morphemes or via significant morphophonological changes produce a high number of inflected and compound words which appear to the machine as brand new, unrelated words (Dreyer & Eisner 2011; Goldsmith, Lee & Xanthos 2017; Hammarström & Borin 2011; Kann, Cotterell & Schütze 2016; Ruokolainen et al. 2013). NLP systems that account for morphology can reduce data sparsity caused by an abundance of individual word forms (McCarthy et al. 2019; Vylomova et al. 2020) and help mitigate bias in training data (Zmigrod et al. 2019). Computational morphological systems have often been limited to languages with publicly available structured data, for example, tables of inflectional patterns in online dictionaries like Wiktionary. Unfortunately, complete inflectional tables are not easily available for many of the world's languages.

Morphological analysis can be separated into two core tasks (Cotterell et al. 2015; Hammarström & Borin 2011; Nicolai & Kondrak 2017; Palmer 2009). The first task is identifying morphemes by determining their shapes and marking boundaries between them, as was done for the Lezgi noun in example 1b below. This is known as (unlabeled) morpheme segmentation (Creutz & Lagus 2007; Snyder & Barzilay 2008). The second task is deducing each morpheme's meaning, which is known as parsing, or sometimes called morphological analysis by itself.[3] This single step is known in linguistics as glossing, and in computational linguists as labeled morpheme segmentation or, merely, labeling, or tagging.

Together segmentation and glossing make up a significant part of interlinearization in documentary and descriptive linguistics. These two tasks (step 3 of Bird and Chiang's workflow on) are often the most detailed analytical tasks undertaken while still in the field. They are also perhaps the most time-consuming tasks, requiring at least as much, and probably more, time than transcription which can take up to 100 hours for each hour of recorded speech. The linguistic information provided by morpheme segments and glosses lays a vital foundation for subsequent descriptive work.

Many NLP models have been applied to morpheme segmentation and glossing. Automatic morpheme segmentation is commonly traced to the early work of Harris (1955) and much segmentation research since then has implemented unsupervised learning which he inspired (Goldsmith 2001; Creutz & Lagus 2002; Poon, Cherry & Toutanova 2009). The preponderance of unsupervised models was probably motivated by the difficulty of finding the high quantity and quality manually segmented data needed to train supervised models. Lack of sufficient training data is illustrated by a recent supervised segmentation experiment (Ansari et al. 2019) which needed to manually segment a corpus before conducting the experiment.

In LDD, segmentation and glossing are typically tackled simultaneously. Segmentation finds breaks between morphemes as for the Lezgi noun in 1b. Glossing labels morphemes with their meaning or function, as in 1c. Glossing does not require segmentation, and if done independently is sometimes referred to as parsing. Parsing by itself would only provide the information in 1c without indication of morpheme boundaries.

(1) a. paçahdin
b. paçah-di-n
c. king-obl-gen
d. ‘king’s’

NLP experiments with LRL often treat segmentation and glossing as separate tasks. Other NLP works have taken a tip from LDD and joined the two tasks. Joint learning of segmentation and glossing, or labeled segmentation, is less common but has been successful for LRL (Cotterell et al., 2015; Moeller & Hulden, 2018). In general, joint learning is characterized by training on different types of information and is based on the intuition that one type of linguistic knowledge (e.g., syntax) can improve results in another domain (e.g., morphology) (Goldsmith et al., 2017). Much NLP work focuses on glossing only, which assumes that the data is already is, or does not need to be, segmented into morphemes. McMillan-Major (2020) trained systems to produce a gloss line by incorporating predictions made from existing segmentations and from the free translation enriched with INTENT (Georgi 2016). Both Samardzic et al. (2015) also used information from other IGT lines such as translation and part-of-speech tags to train a system to gloss.

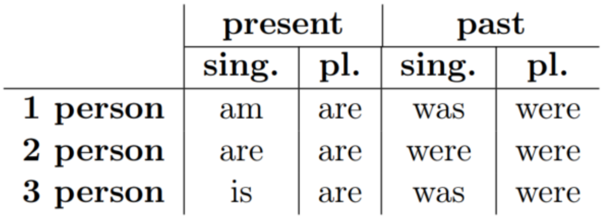
5. INFLECTIONAL PARADIGM LEARNING

Morphology includes the inference of rules that govern word building strategies and the discovery of how word forms are systematically related (Roark & Sproat 2007). Therefore, Virpioja et al. (2011) add a third task to morphological analysis: identification of morphologically related words through patterns of inflection.

Durrett and DeNero (2013) claim that the inference of inflectional patterns must be based on three assumptions. First, each lexical category is dictated by a subsystem of rules. Russian nouns, for example, can be generalized into three simplified patterns of inflection that are usually labeled "masculine", "feminine", and "neuter". Lexemes that adhere to the same pattern are grouped into inflectional classes (sometimes called "declensions" for nouns and adjectives and "conjugations" for verbs). The patterns themselves are known as inflectional paradigms. Second, inflectional changes are triggered by context and, therefore, the patterns can be inferred from context. Descriptive studies look to phonology or else to both phonological structure and the semantic content of the lexeme for the triggering context. Computational models, due to the nature of their input, look to orthographic context. The third assumption is that each stem morpheme is inflected consistently according to the inflectional class it belongs to with any idiosyncrasies of the stem itself.

Monson et al. (2007) give two guiding principles for computational paradigm learning. One is that inflected forms of a lemma look similar to each other. This principle holds well enough to serve as a solid working assumption, although languages abound with exceptions and inflection can even be suppletive (e.g., *go* vs. *went*, etc.). The second principle is that "in any given corpus, a particular lexeme will likely not occur in all possible inflected forms". Inflectional paradigms can be quite large. Languages may have hundreds or even thousands of forms per lemma (Corbett 2013). Even with a large corpus, attempts to learn paradigms, like the one illustrated in Figure 6, by only using the corpus will leave empty cells in the paradigm's table. It is possible that certain forms may never occur in natural language even though they are grammatically possible (Silfverberg & Hulden 2018). Additionally, frequent words often follow irregular patterns, as does the English verb *be*. For these reasons, the LDD workflow includes elicitation of morphological paradigms (Lupke 2010; Boerger et al. 2016).

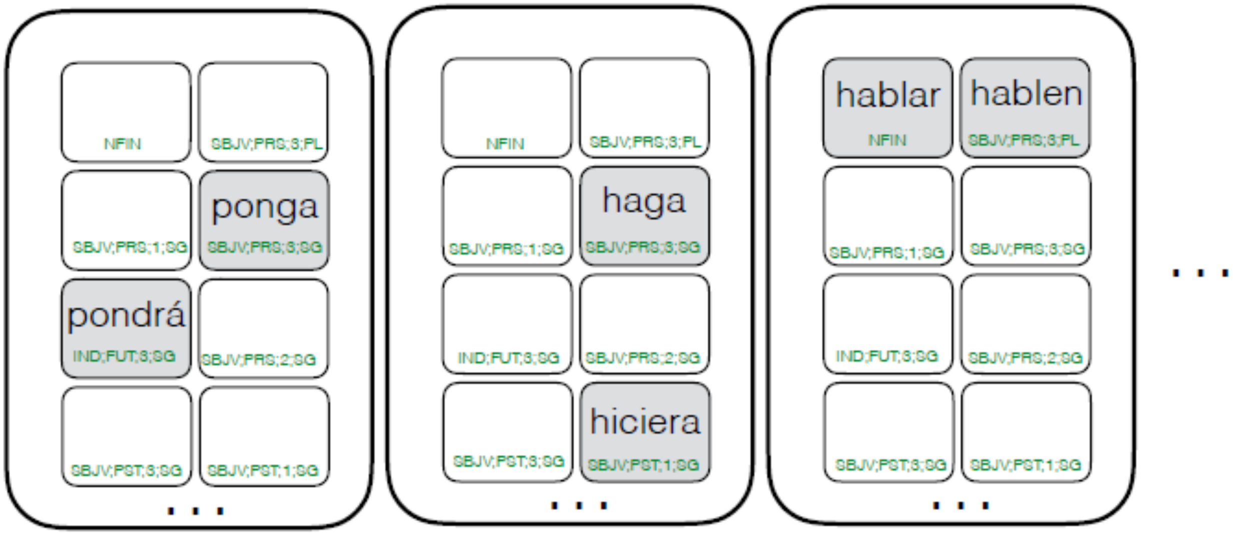
FIGURE 6. INFLECTIONAL PARADIGM OF THE ENGLISH VERB "TO BE"



	present		**past**	
	sing.	**pl.**	**sing.**	**pl.**
1 person	am	are	was	were
2 person	are	are	were	were
3 person	is	are	was	were

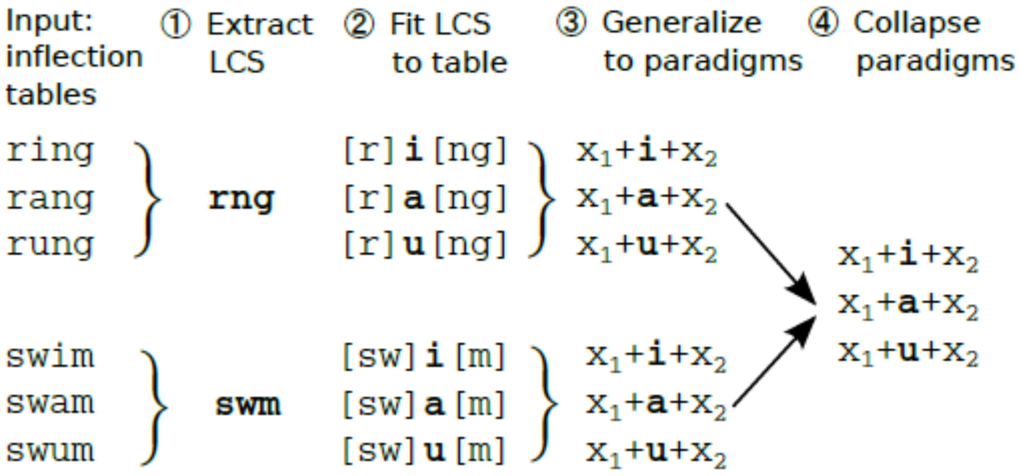
Computational models can learn frequent and regular paradigmatic patterns with over 90% accuracy even in low-resource settings (Hammarström & Borin 2011; Durrett & DeNero 2013; Ahlberg, Forsberg & Hulden 2014). Most early work on paradigm induction applied unsupervised learning to concatenative morphology (Goldsmith 2001; Chan 2006; Monson et al. 2007). The unsupervised version of the paradigm completion task (Jin et al. 2020) has been the subject of a recent shared task (Kann et al. 2020), with the conclusion that it is extremely challenging for current state-of-the-art systems. Semi-supervised models have been more recently applied on concatenative and non-concatenative languages (Dreyer & Eisner 2011; Durrett & DeNero 2013). Supervised learning has also been applied to inflectional morphology. Some work focuses on generating inflected forms, including work motivated by the Paradigm Cell Filling Problem (PCFP), illustrated in Figure 7 (Ackerman, Blevins & Malouf 2009). The PCFP is framed as an attempt to model how new speakers (e.g., young children) infer the inflected forms they have not yet encountered (Dreyer & Eisner 2011; Ahlberg, Forsberg & Hulden 2015; Malouf 2016; Silfverberg & Hulden 2018).

FIGURE 7. ILLUSTRATION OF THE PARADIGM CELL FILLING PROBLEM (SILFVERBERG AND HULDEN, 2018) WITH SPANISH VERB PARADIGMS



Other work with supervised learning has attempted to induce inflectional paradigms from text. With this method, paradigms are completed by finding overlapping patterns from several incomplete paradigms in text. One method does this by abstracting the longest common subsequence of characters in inflected forms of the same lexeme and then clustering words with same or similar patterns (Ahlberg, Forsberg & Hulden 2014; Ahlberg, Forsberg & Hulden 2015). This is illustrated in Figure 8. Exceptions or irregularities in the paradigms can be accounted for by collapsing the similar patterns. The experiment has been quite successful for a few Indo-European languages (German, Spanish, Catalan, French, Galician, Italian, Portuguese, Russian), as well as Maltese and Finnish. Kann et al. (2017a) differed from other approaches in that they encoded multiple inflected forms of a lemma to provide complementary information in order to generate unknown forms. Cotterell et al. (2017) introduced neural graphical models which completed paradigms based on principal parts.

FIGURE 8. AHLBERG ET AL. (2015): INDUCING PARADIGMS



The longest common subsequences (LCS) *rng* or *swm* are extracted (step 1) and represented as x_1 and x_2 which replace the LCS (step 2). Words with the same inflectional patterns will be identical (step 3) and can be generalized into paradigms (step 4). The remaining characters *i*, *a*, *u* are assumed to be inflectional affixes.

Most recent work in paradigm induction has been concerned with generation (as opposed to analysis) of inflected words and has focused on morphological inflection or reinflection (Durrett & DeNero 2013; Nicolai, Cherry & Kondrak 2015; Faruqui et al. 2016; Kann & Schütze 2016; Aharoni & Goldberg 2017). Partially building on these, other research has developed machine learning models which are more suitable for LRL and perform well with limited data (Kann, Cotterell & Schütze 2017b; Sharma, Katrapati & Sharma 2018; Makarov & Clematide 2018b; Wu & Cotterell 2019; Kann, Bowman & Cho 2020; Wu, Cotterell & Hulden 2021).

6. CONCLUSION

Morphology comprises word-building properties in human languages and their accompanying (morpho)syntactic phenomena. Historically, NLP and "paper-and-pencil linguists" have taken different and sometimes seemingly incompatible approaches to morphology (Karttunen & Beesley 2005). Despite their out-of-sync approaches, both benefit from morphological analysis (Cotterell et al. 2015). Morphological analysis includes morpheme segmentation, glossing, and learning inflectional paradigmatic patterns. This paper presented the history of computational models for morphological analysis and looked specifically at their application and success when limited training data is available. The work discussed in this paper demonstrate that computational models and methods can both successfully perform morphological analysis. This success may be of great benefit for the documentation and description of endangered or under-documented languages.

REFERENCES

Ackerman, Farrell, James P. Blevins & Robert Malouf. 2009. Parts and wholes: Implicative patterns in inflectional paradigms. In *Analogy in Grammar*. Oxford: Oxford University Press.

Aharoni, Roee & Yoav Goldberg. 2017. Morphological Inflection Generation with Hard Monotonic Attention. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 1. 2004–2015. https://aclanthology.coli.uni-saarland.de/papers/P17-1183/p17-1183 (16 January, 2018).

Ahlberg, Malin, Markus Forsberg & Mans Hulden. 2014. Semi-supervised learning of morphological paradigms and lexicons. In *Proceedings of 14th Conference of the European Chapter of the Association for Computational Linguistics*, 569–578. Gothenburg, Sweden: Association for Computational Linguistics. http://www.aclweb.org/anthology/E/E14/E14-1.pdf#page=569 (1 November, 2016).

Ahlberg, Malin, Markus Forsberg & 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*, 1024–1029. Denver, Colorado: Association for Computational Linguistics. http://www.aclweb.org/anthology/N15-1107 (18 January, 2019).

Anastasopoulos, Antonios, David Chiang & Long Duong. 2016. An Unsupervised Probability Model for Speech-to-Translation Alignment of Low-Resource Languages. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1255–1263. Austin, Texas: Association for Computational Linguistics. https://aclweb.org/anthology/D16-1133 (2 July, 2018).

Ansari, Ebrahim, Zdeněk Žabokrtský, Mohammad Mahmoudi, Hamid Haghdoost & Jonáš Vidra. 2019. Supervised Morphological Segmentation Using Rich Annotated Lexicon. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, 52–61. Varna, Bulgaria: INCOMA Ltd. https://www.aclweb.org/anthology/R19-1007 (25 January, 2020).

Auer, Eric, Albert Russel, Han Sloetjes, Peter Wittenburg, Oliver Schreer, S. Masnieri, Daniel Schneider & Sebastian Tschöpel. 2010. ELAN as Flexible Annotation Framework for Sound and Image Processing Detectors. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *European Language Resources Association LREC 2010: Proceedings of the 7th International Language Resources and Evaluation*, 890–893. Paris: ELRA: European Language Resources Association. http://dblp.uni-trier.de/db/conf/lrec/lrec2010.html#AuerRSWSMST10 (30 January, 2014).

Baldridge, Jason & Miles Osborne. 2008. Active learning and logarithmic opinion pools for HPSG parse selection. *Natural Language Engineering* 14(2). 191–222.

Baldridge, Jason & Alexis Palmer. 2009. How well does active learning actually work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 296–305. Singapore. http://www.aclweb.org/anthology/D/D09/D09-1031.pdf.

Bender, Emily M. 2014. Language CoLLAGE: Grammatical Description with the LinGO Grammar Matrix. In *Proceedings of the Ninth International Conference of Language Resources and Evaluation (LREC-2014)*, 2447–2451. http://www.lrec-conf.org/proceedings/lrec2014/pdf/639_Paper.pdf.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent & Christian Jauvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research* 3. 1137–1155.

Bergmanis, Toms, Katharina Kann, Hinrich Schütze & Sharon Goldwater. 2017. Training Data Augmentation for Low-Resource Morphological Inflection. *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection* 31–39. https://aclanthology.coli.uni-saarland.de/papers/K17-2002/k17-2002 (16 January, 2018).

Biljon, Elan van, Arnu Pretorius & Julia Kreutzer. 2020. On Optimal Transformer Depth for Low-Resource Language Translation. *arXiv:2004.04418 [cs]*. http://arxiv.org/abs/2004.04418 (14 May, 2020).

Bird, Steven. 2009. Natural Language Processing and Linguistic Fieldwork. *Computational Linguistics* 35(3). 469–474. http://dx.doi.org/10.1162/coli.35.3.469 (16 March, 2015).

Bird, Steven & David Chiang. 2012. Machine Translation for Language Preservation. In *Proceedings of COLING 2012*, 125–134. Mumbai.

Boerger, Brenda H., Sarah Ruth Moeller, Will Reiman & Stephen Self. 2016. *Language and Culture Documentation Manual*. Leanpub. https://leanpub.com/languageandculturedocumentationmanual (22 May, 2020).

Bowern, Claire. 2008. *Linguistic fieldwork: a practical guide*. Houndmills, Basingstoke, Hampshire [England]; New York: Palgrave Macmillan.

Chan, Erwin. 2006. Learning Probabilistic Paradigms for Morphology in a Latent Class Model. In *Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology and Morphology* (SIGPHON '06), 69–78. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1622165.1622174.

Corbett, Greville G. 2013. The unique challenge of the Archi paradigm. In Chundra Cathcart, Shinae Kang & Clare S. Sandy (eds.), *Proceedings of the 37th Annual Meeting of the Berkeley Linguistics Society:Special Session on Languages of the Caucasus*, 52–67. Berkeley, CA. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.657.9753&rep=rep1&type=pdf (12 January, 2021).

Cotterell, Ryan & Georg Heigold. 2017. Cross-lingual Character-Level Neural Morphological Tagging. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* 748–759. https://aclanthology.coli.uni-saarland.de/papers/D17-1078/d17-1078 (16 January, 2018).

Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, et al. 2018. The CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection. In *Proceedings of the CoNLL SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, 1–27. Brussels: Association for Computational Linguistics. http://www.aclweb.org/anthology/K18-3001 (2 November, 2018).

Cotterell, Ryan, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, et al. 2017. CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection in 52 Languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, 1–30.

Vancouver: Association for Computational Linguistics. http://www.aclweb.org/anthology/K17-2001.

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

Cotterell, Ryan, Thomas Müller, Alexander M. Fraser & Hinrich Schütze. 2015. Labeled Morphological Segmentation with Semi-Markov Models. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, 164–174. Beijing, China: Association for Computational Linguistics.

Cotterell, Ryan, John Sylak-Glassman & Christo Kirov. 2017. Neural Graphical Models over Strings for Principal Parts Morphological Paradigm Completion. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* 2. 759–765. https://aclanthology.coli.uni-saarland.de/papers/E17-2120/e17-2120 (16 January, 2018).

Cotterell, Ryan, Tim Vieira & Hinrich Schütze. 2016. A Joint Model of Orthography and Morphological Segmentation. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 664–669. https://aclanthology.coli.uni-saarland.de/papers/N16-1080/n16-1080 (16 January, 2018).

Cox, Christopher, Gilles Bouliame & Jahangir Alam. 2019. Taking aim at the transcription bottleneck: Integrating speech technology into language documentation and conservation. Slideshow presented at the 6th International Conference on Language Documentation and Conservation (ICDLC), Honolulu, HI. https://scholarspace.manoa.hawaii.edu/handle/10125/44841.

Creutz, Mathias & Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 workshop on Morphological and phonological learning-Volume 6*, 21–30. Philadelphia, PA: Association for Computational Linguistics.

Creutz, Mathias & Krista Lagus. 2007. Unsupervised Models for Morpheme Segmentation and Morphology Learning. *ACM Trans. Speech Lang. Process.* 4(1). 3:1-3:34. http://users.ics.aalto.fi/krista/papers/creutz07acmtslp.pdf.

Czaykowska-Higgins, Ewa. 2009. Research Models, Community Engagement, and Linguistic Fieldwork: Reflections on Working within Canadian Indigenous Communities. *Language Documentation & Conservation* 3(1). 15–50. http://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/4423/czaykowskahiggins.pdf?sequence=1.

Dreyer, Markus & Jason Eisner. 2011. Discovering morphological paradigms from plain text using a Dirichlet process mixture model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 616–627. Association for Computational Linguistics.

Duong, Long. 2017. *Natural language processing for resource-poor languages*. Melbourne, Australia: University of Melbourne PhD Thesis. http://minerva-access.unimelb.edu.au/handle/11343/192938 (29 June, 2018).

Duong, Long, Antonios Anastasopoulos, David Chiang, Steven Bird & Trevor Cohn. 2016. An Attentional Model for Speech Translation Without Transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 949–959. San Diego, California: Association for Computational Linguistics. http://www.aclweb.org/anthology/N16-1109 (6 July, 2018).

Durrett, Greg & 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*, 1185–1195. Atlanta, Georgia: Association for Computational Linguistics. https://research.google.com/pubs/pub41850.html (2 January, 2018).

Faruqui, Manaal, Yulia Tsvetkov, Graham Neubig & 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*, 634–643. San Diego, California: Association for Computational Linguistics. https://www.aclweb.org/anthology/N16-1077 (28 April, 2021).

Felt, Paul. 2012. *Improving the Effectiveness of Machine-Assisted Annotation*. Brigham Young University MA Thesis. https://scholarsarchive.byu.edu/etd/3214.

Foley, Ben, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath, et al. 2018. Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System. In *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2018)*. https://www.isca-speech.org/archive/SLTU_2018/pdfs/Ben.pdf (7 March, 2019).

Forsberg, Markus & Mans Hulden. 2016. Learning Transducer Models for Morphological Analysis from Example Inflections. In *Proceedings of the SIGFSM Workshop on Statistical NLP and Weighted Automata*, 42–50. Berlin, Germany: Association for Computational Linguistics. http://anthology.aclweb.org/W16-2405 (18 January, 2019).

Georgi, Ryan Alden. 2016. *From Aari to Zulu: Massively Multilingual Creation of Language Tools using Interlinear Glossed Text*. PhD Thesis. https://digital.lib.washington.edu:443/researchworks/handle/1773/37168 (28 September, 2020).

Goldberg, Yoav. 2017. *Neural Network Methods for Natural Language Processing* (Synthesis Lectures on Human Language Technologies 37). Morgan & Claypool. http://www.morganclaypool.com/doi/10.2200/S00762ED1V01Y201703HLT037 (12 January, 2019).

Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. *Computational linguistics* 27(2). 153–198.

Goldsmith, John, Jackson Lee & Aris Xanthos. 2017. Computational Learning of Morphology. *Annual Review of Linguistics* 3. 85–106.

Hammarström, Harald & Lars Borin. 2011. Unsupervised learning of morphology. *Computational Linguistics* 37(2). 309–350. http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00050 (29 August, 2017).

Harris, Zellig. 1955. From phoneme to morpheme. *Language* 31(2). 190–222.

He, Luheng, Julian Michael, Mike Lewis & Luke Zettlemoyer. 2016. Human-in-the-Loop Parsing. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2337–2342. Austin, Texas: Association for Computational Linguistic.

Himmelmann, Nikolaus P. 1998. Documentary and Descriptive Linguistics. *Linguistics* (36). 161–195.

Jin, Huiming, Liwei Cai, Yihui Peng, Chen Xia, Arya McCarthy & Katharina Kann. 2020. Unsupervised Morphological Paradigm Completion. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6696–6707. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.598. https://www.aclweb.org/anthology/2020.acl-main.598 (8 April, 2021).

Kann, Katharina, Samuel R. Bowman & Kyunghyun Cho. 2020. Learning to Learn Morphological Inflection for Resource-Poor Languages. *Proceedings of the AAAI Conference on Artificial Intelligence* 34(05). 8058–8065. https://ojs.aaai.org/index.php/AAAI/article/view/6316 (26 January, 2021).

Kann, Katharina, Ryan Cotterell & Hinrich Schütze. 2016. Neural Morphological Analysis: Encoding-Decoding Canonical Segments. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 961–967. Austin, Texas: Association for Computational Linguistics. http://aclweb.org/anthology/D16-1097 (6 November, 2020).

Kann, Katharina, Ryan Cotterell & Hinrich Schütze. 2017a. Neural Multi-Source Morphological Reinflection. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* 514–524.

Kann, Katharina, Ryan Cotterell & Hinrich Schütze. 2017b. One-Shot Neural Cross-Lingual Transfer for Paradigm Completion. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1993–2003. Vancouver, Canada: Association for Computational Linguistics. https://www.aclweb.org/anthology/P17-1182 (26 January, 2021).

Kann, Katharina, Arya D. McCarthy, Garrett Nicolai & Mans Hulden. 2020. 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*, 51–62. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.sigmorphon-1.3 (8 April, 2021).

Kann, Katharina & 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)*, 555–560. Berlin, Germany: Association for Computational Linguistics. http://anthology.aclweb.org/P16-2090 (5 November, 2018).

Karttunen, Lauri & Kenneth R. Beesley. 2005. Twenty-five years of finite-state morphology. In *Inquiries Into Words, a Festschrift for Kimmo Koskenniemi on his 60th Birthday*, 71–83. CSLI Publications.

Kirschenbaum, Amit, Peter Wittenburg & Gerhard Heyer. 2012. Unsupervised morphological analysis of small corpora: First experiments with Kilivila. *Language Documentation & Conservation Special Publication* (Potentials of Language Documentation: Methods, Analyses, and Utilization) 3. 25–31. http://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/4513/04kirschenbaumetal.pdf.

Kohonen, Oskar, Sami Virpioja & Krista Lagus. 2010. Semi-supervised learning of concatenative morphology. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, 78–86. Association for Computational Linguistics.

LeCun, Yann, Yoshua Bengio & Geoffrey Hinton. 2015. Deep learning. *Nature* 521(7553). 436–444. https://doi.org/10.1038/nature14539. https://doi.org/10.1038/nature14539.

Lepp, Haley, Olga Zamaraeva & Emily M. Bender. 2019. Visualizing Inferred Morphotactic Systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 127–131. Minneapolis, Minnesota: Association for Computational Linguistics. https://www.aclweb.org/anthology/N19-4022 (6 April, 2020).

Lewis, William D. & Fei Xia. 2010. Developing ODIN: A Multilingual Repository of Annotated Language Data for Hundreds of the World's Languages. *Literary and Linguistic Computing* 25(3). 303–319. http://llc.oxfordjournals.org/content/25/3/303 (29 January, 2014).

Liu, Ling, Ilamvazhuthy Subbiah, Adam Wiemerslage, Jonathan Lilley & Sarah Moeller. 2018. Morphological Reinflection in Context: CU Boulder's Submission to CoNLL-

SIGMORPHON 2018 Shared Task. In *Proceedings of the CoNLL SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, 86–92. Brussels: Association for Computational Linguistics. http://www.aclweb.org/anthology/K18-3010 (2 November, 2018).

Lupke, Friederike. 2010. Data collection methods for field-based language documentation. *Language Documentation and Description* 7. 55–104.

Makarov, Peter & Simon Clematide. 2018a. UZH at CoNLL–SIGMORPHON 2018 Shared Task on Universal Morphological Reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, 69–75. Brussels: Association for Computational Linguistics. https://www.aclweb.org/anthology/K18-3008 (16 May, 2019).

Makarov, Peter & Simon Clematide. 2018b. Imitation Learning for Neural Morphological String Transduction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2877–2882. Brussels, Belgium: Association for Computational Linguistics. https://www.aclweb.org/anthology/D18-1314 (28 April, 2021).

Makarov, Peter, Tatiana Ruzsics & Simon Clematide. 2017. Align and Copy: UZH at SIGMORPHON 2017 Shared Task for Morphological Reinflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, 49–57. Vancouver: Association for Computational Linguistics. http://www.aclweb.org/anthology/K17-2004 (3 November, 2018).

Malouf, Robert. 2016. Generating morphological paradigms with a recurrent neural network. *San Diego Linguistic Papers* 6. 122–129.

McCarthy, Arya D, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, et al. 2019. The SIGMORPHON 2019 Shared Task: Crosslinguality and Context in Morphology. In *Proceedings of the 16th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*. Florence, Italy: Association for Computational Linguistics.

McMillan-Major, Angelina. 2020. Automating Gloss Generation in Interlinear Glossed Text. In *Proceedings of the Society for Computation in Linguistics*, vol. 3, 338–349. https://doi.org/10.7275/tsmk-sa32. https://scholarworks.umass.edu/scil/vol3/iss1/33.

Moeller, Sarah & Mans Hulden. 2018. Automatic Glossing in a Low-Resource Setting for Language Documentation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, 84–93. Santa Fe, New Mexico, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/W18-4809 (22 August, 2018).

Moeller, Sarah, Ghazaleh Kazeminejad, Andrew Cowell & Mans Hulden. 2018. A Neural Morphological Analyzer for Arapaho Verbs Learned from a Finite State Transducer. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, 12–20. Santa Fe, New Mexico, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/W18-4802 (22 August, 2018).

Moeller, Sarah, Ghazaleh Kazeminejad, Andrew Cowell & Mans Hulden. 2019. Improving Low-Resource Morphological Learning with Intermediate Forms from Finite State Transducers. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, vol. 1. Honolulu, HI. https://www.aclweb.org/anthology/W19-6011/.

Monson, Christian, Jaime Carbonell, Alon Lavie & Lori Levin. 2007. ParaMor: Finding Paradigms across Morphology. In *Advances in Multilingual and Multimodal Information Retrieval* (Lecture Notes in Computer Science), 900–907. Springer, Berlin, Heidelberg. https://link.springer.com/chapter/10.1007/978-3-540-85760-0_115 (27 February, 2018).

Moon, Taesun, Katrin Erk & Jason Baldridge. 2009. Unsupervised morphological segmentation and clustering with document boundaries. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2*, 668–677. Association for Computational Linguistics.

Nicolai, Garrett, Colin Cherry & Grzegorz Kondrak. 2015. Inflection Generation as Discriminative String Transduction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 922–931. Denver, Colorado: Association for Computational Linguistics. https://doi.org/10.3115/v1/N15-1093. https://www.aclweb.org/anthology/N15-1093 (26 January, 2021).

Nicolai, Garrett, Kyle Gorman & Ryan Cotterell (eds.). 2020. *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and*

Morphology. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.sigmorphon-1.0 (8 April, 2021).

Nicolai, Garrett & Grzegorz Kondrak. 2017. Morphological Analysis without Expert Annotation. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* 2. 211–216. https://aclanthology.coli.uni-saarland.de/papers/E17-2034/e17-2034 (16 January, 2018).

Palmer, Alexis Mary. 2009. *Semi-automated annotation and active learning for language documentation*. University of Texas at Austin PhD Thesis.

Palmer, Alexis, Taesun Moon, Jason Baldridge, Katrin Erk, Eric Campbell & Telma Can. 2010. Computational strategies for reducing annotation effort in language documentation. *Linguistic Issues in Language Technology* 3(4). 1–42. http://journals.linguisticsociety.org/elanguage/lilt/article/view/663.html (24 January, 2014).

Poon, Hoifung, Colin Cherry & Kristina Toutanova. 2009. Unsupervised morphological segmentation with log-linear models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 209–217. Association for Computational Linguistics.

Popel, Martin & Ondřej Bojar. 2018. Training Tips for the Transformer Model. *The Prague Bulletin of Mathematical Linguistics* 110(1). 43–70. http://content.sciendo.com/view/journals/pralin/110/1/article-p43.xml (13 May, 2020).

Rice, Sally & Dorothy Thunder. 2017. Community-based corpus-building: Three case studies. Presented at the 5th International Conference on Language Documentation and Conservation (ICLDC), Honolulu, HI. http://scholarspace.manoa.hawaii.edu/handle/10125/42052 (6 June, 2017).

Roark, Brian & Richard William Sproat. 2007. *Computational approaches to morphology and syntax*. Oxford; New York: Oxford University Press.

Rogers, Chris. 2010. Review of Fieldworks Language Explorer (FLEx) 3.0. *Language Documentation & Conservation* 4. 78–84. http://scholarspace.manoa.hawaii.edu/handle/10125/4471 (13 March, 2018).

Ruokolainen, Teemu, Oskar Kohonen, Kairit Sirts, Stig-Arne Grönroos, Mikko Kurimo & Sami Virpioja. 2016. A Comparative Study of Minimally Supervised Morphological

Segmentation. *Computational Linguistics* 42(1). 91–120. http://www.mitpressjournals.org/doi/10.1162/COLI_a_00243 (17 January, 2018).

Ruokolainen, Teemu, Oskar Kohonen, Sami Virpioja & Mikko Kurimo. 2013. Supervised Morphological Segmentation in a Low-Resource Learning Setting using Conditional Random Fields. In *CoNLL*, 29–37.

Samardzic, Tanja, Robert Schikowski & Sabine Stoll. 2015. Automatic interlinear glossing as two-level sequence classification. In *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Herita ge, Social Sciences, and Humanities (LaTeCH)*, 68–72. Beijing, China: Association for Computational Linguistics. http://aclweb.org/anthology/W15-3710 (25 May, 2020).

Seifart, Frank, Nicholas Evans, Harald Hammarström & Stephen C. Levinson. 2018. Language documentation twenty-five years on. *Language* 94(4). e324–e345. https://muse.jhu.edu/article/712110 (8 October, 2019).

Sharma, Abhishek, Ganesh Katrapati & Dipti Misra Sharma. 2018. IIT(BHU)–IIITH at CoNLL–SIGMORPHON 2018 Shared Task on Universal Morphological Reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, 105–111. Brussels: Association for Computational Linguistics. https://www.aclweb.org/anthology/K18-3013 (26 January, 2021).

Silfverberg, Miikka & 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*, 2883–2889. Brussels, Belgium: Association for Computational Linguistics. https://www.aclweb.org/anthology/D18-1315 (25 April, 2019).

Snyder, Benjamin & Regina Barzilay. 2008. Unsupervised Multilingual Learning for Morphological Segmentation. In *ACL*, 737–745.

Soricut, Radu & Franz Och. 2015. Unsupervised Morphology Induction Using Word Embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1627–1637. Denver, Colorado.

Sudhakar, Akhilesh & Anil Kumar Singh. 2017. Experiments on Morphological Reinflection: CoNLL-2017 Shared Task. *Proceedings of the CoNLL SIGMORPHON 2017 Shared*

Task: Universal Morphological Reinflection 71–78. https://doi.org/10.18653/v1/K17-2007. https://aclanthology.coli.uni-saarland.de/papers/K17-2007/k17-2007 (16 January, 2018).

Vallejos, Rosa. 2014. Integrating Language Documentation, Language Preservation, and Linguistic Research: Working with the Kokamas from the Amazon. *Language Documentation & Conservation* 8. 38–65. http://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/4618/vallejos.pdf?sequence=1.

Virpioja, Sami, Ville Turunen, Sebastian Spiegler, Oskar Kohonen & Mikko Kurimo. 2011. Empirical Comparison of Evaluation Methods for Unsupervised Learning of Morphology. *Trait. Autom. des Langues* 52(2). 45–90.

Vylomova, Ekaterina, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, et al. 2020. SIGMORPHON 2020 Shared Task 0: Typologically Diverse Morphological Inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 1–39. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.sigmorphon-1.1 (27 April, 2021).

Wang, Linlin, Zhu Cao, Yu Xia & Gerard de Melo. 2016. Morphological Segmentation with Window LSTM Neural Networks. In *AAAI'16: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2842–2848.

Wax, David Allen. 2014. *Automated Grammar Engineering for Verbal Morphology*. Thesis. https://digital.lib.washington.edu:443/researchworks/handle/1773/25373 (6 April, 2020).

Woodbury, Tony. 2003. Defining Documentary Linguistics. *Language Documentation and Description* 1. 35–51.

Wu, Shijie & Ryan Cotterell. 2019. Exact Hard Monotonic Attention for Character-Level Transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1530–1537. Florence, Italy: Association for Computational Linguistics. https://www.aclweb.org/anthology/P19-1148 (26 January, 2021).

Wu, Shijie, Ryan Cotterell & Mans Hulden. 2021. Applying the Transformer to Character-level Transduction. In *Proceedings of the 16th Conference of the European Chapter of the*

Association for Computational Linguistics: Main Volume, 1901--1907. Association for Computational Linguistics. https://aclanthology.org/2021.eacl-main.163.

Xia, Fei, William D. Lewis, Michael Wayne Goodman, Glenn Slayden, Ryan Georgi, Joshua Crowgey & Emily M. Bender. 2016. Enriching a massively multilingual database of interlinear glossed text. *Language Resources and Evaluation* 50(2). 321–349. https://doi.org/10.1007/s10579-015-9325-4 (14 January, 2020).

Zmigrod, Ran, Sabrina J. Mielke, Hanna Wallach & Ryan Cotterell. 2019. Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1651–1661. Florence, Italy: Association for Computational Linguistics. https://www.aclweb.org/anthology/P19-1161 (27 April, 2021).

ENDNOTES

[1] According to LORELEI (https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents), ``low-resource" refers to languages for which no automated human language technology exists. This is due to a lack of linguistic resources. \citet{szymanski_morphological_2012} estimates that 99\% of the world's languages are ``resource-poor". In linguistics, it is more common to hear other terms. The current work uses the term ``under-described languages" to refer to languages with minimal published linguistic resources; these have been called ``very scarce-resource language" \citep{duong_natural_2017}. The term ``under-documented languages" (Duong's ``extremely scarce-resource languages") refer to languages that lack sufficient raw or annotated data to write a full reference grammar. The term ``endangered languages" refers to languages that are predicted to have no native speakers within a generation or two. Most endangered languages are under-documented and/or under-described, as well as fitting the definition of low-resource languages. The distinctions between the terms are rarely crucial in the current work. In practice, these terms can be used almost interchangeably.

[2] Deep learning morphological segmentation has been performed on unsupervised texts with some success (Wang et al. 2016).

[3]Nicolai and Kondrak (2017) subdivide morphological analysis slightly differently, making a distinction between morphological “analysis” and morphological tagging. They describe morphological analysis as a combination of segmentation and labeling, though they later state that “morphological tagging can be performed as a downstream application of morphological analysis” (p. 211), thereby adhering to the same two distinctions described in this section.