

Rethinking Tools for the Morphosyntactic Analysis of Underdocumented Languages

Sara Court¹

Maria Copot²

¹The Ohio State University ²Université Paris Cité, LLF, CNRS

Collaboration with...



Noah Diewald



Stephanie Antetomaso



Micha Elsner

The thing that this is

- A **workflow for morphosyntactic annotation and analysis** of underdocumented languages.

The thing that this is

- A **workflow for morphosyntactic annotation and analysis** of underdocumented languages.
 - Dealing with data scarcity through heavy use of **ML, NLP and human-in-the-loop** methods.

The thing that this is

- A **workflow for morphosyntactic annotation and analysis** of underdocumented languages.
 - Dealing with data scarcity through heavy use of **ML, NLP and human-in-the-loop** methods.
 - Different **theoretical framework** from previous tools.

The thing that this is

- A **workflow for morphosyntactic annotation and analysis** of underdocumented languages.
 - Dealing with data scarcity through heavy use of **ML, NLP and human-in-the-loop** methods.
 - Different **theoretical framework** from previous tools.
 - Designed to increase **community engagement** with linguistic fieldwork.

- The **problem**
 - Fieldwork on underdocumented languages - challenges and stakeholders
 - Morphosyntactic description and analysis

- The **problem**
 - Fieldwork on underdocumented languages - challenges and stakeholders
 - Morphosyntactic description and analysis
- The **solution**
 - The theory: Word-and-Paradigm morphology
 - The implementation: software and piloting

The plan

- The **problem**
 - Fieldwork on underdocumented languages - challenges and stakeholders
 - Morphosyntactic description and analysis
- The **solution**
 - The theory: Word-and-Paradigm morphology
 - The implementation: software and piloting
- What's **next?**

The problem

- 50-90% of world's languages estimated to be severely endangered or dead by 2100 (Austin & Sallabank, 2011)

- 50-90% of world's languages estimated to be severely endangered or dead by 2100 (Austin & Sallabank, 2011)
- Communities shift to speaking majority languages
 - out of stigmatisation
 - as a means of seeking out opportunities

- Affected communities are losing part of their **identity**
- Humanity is losing access to **knowledge**
- Researchers are experiencing **artificially reduced variation** in the object of study

Bridging the divide between researchers and community

- A growing trend to have **speakers** take an **active role** in fieldwork.

Bridging the divide between researchers and community

- A growing trend to have **speakers** take an **active role** in fieldwork.
- Positive because:
 - The speaker is **aware of the community's wants and needs** - fieldwork can be steered to target them.

Bridging the divide between researchers and community

- A growing trend to have **speakers** take an **active role** in fieldwork.
- Positive because:
 - The speaker is **aware of the community's wants and needs** - fieldwork can be steered to target them.
 - **Diminishes power differential** between the researcher and the community, making fieldwork more ethical.

Bridging the divide between researchers and community

- A growing trend to have **speakers** take an **active role** in fieldwork.
- Positive because:
 - The speaker is **aware of the community's wants and needs** - fieldwork can be steered to target them.
 - **Diminishes power differential** between the researcher and the community, making fieldwork more ethical.
 - The speaker has **intuitive knowledge** of the object of study, narrowing the hypothesis space.

Bridging the divide between researchers and community

- A growing trend to have **speakers** take an **active role** in fieldwork.
- Positive because:
 - The speaker is **aware of the community's wants and needs** - fieldwork can be steered to target them.
 - **Diminishes power differential** between the researcher and the community, making fieldwork more ethical.
 - The speaker has **intuitive knowledge** of the object of study, narrowing the hypothesis space.
- **Common tasks:**
 - Collect raw data (recordings of their community, writing up stories)
 - Data processing and analysis

Barriers for further active involvement of speakers

- Collecting raw data requires **mastering recording equipment/digitising notes**, which may already be a challenge

Barriers for further active involvement of speakers

- Collecting raw data requires **mastering recording equipment/digitising notes**, which may already be a challenge
- The real **bottleneck** is involvement in **data processing and analysis**
 - **Technical barrier** to use existing software
 - Need for **linguistic training** for e.g. applying morphological labels

Documenting and analysing morphosyntactic structure

1. Eliciting basic vocabulary



squirrels



squids



cats

Documenting and analysing morphosyntactic structure

1. Eliciting basic vocabulary



squirrels



squids



cats

2. Understanding the meaning of recurrent substrings

$XS = X.PLURAL$

Documenting and analysing morphosyntactic structure

1. Eliciting basic **vocabulary**



squirrels



squids



cats

2. Understanding the **meaning of recurrent substrings**

$XS = X.PLURAL$

Morphosyntactic analysis is

- a crucial part of **describing** the linguistic system
- the basis for **glossing** – a way to convey linguistic structure for other purposes

Current standard morphosyntactic documentation practices

squirrel-s	would='ve	chase-d	mice
squirrel-PL	COND=PERF	chase-PST	PL\mouse

Current standard morphosyntactic documentation practices

squirrel-s	would='ve	chase-d	mice
squirrel-PL	COND=PERF	chase-PST	PL\mouse

Tricky for understudied languages

Current standard morphosyntactic documentation practices

squirrel-s	would='ve	chase-d	mice
squirrel-PL	COND=PERF	chase-PST	PL\mouse

Tricky for understudied languages

- **Theoretical issues**
 - Early commitment to an analysis
 - Assumption that all morphological patterns are easily described in concatenative terms

Current standard morphosyntactic documentation practices

squirrel-s	would='ve	chase-d	mice
squirrel-PL	COND=PERF	chase-PST	PL\mouse

Tricky for understudied languages

- **Theoretical issues**

- Early commitment to an analysis
- Assumption that all morphological patterns are easily described in concatenative terms

- **Practical issues**

- Suboptimal use of human time
- Requires linguistic training

Existing software for annotation

- **Excel** is a popular choice - a dire situation

Microsoft Excel - test1k-Glong.xml.01.xls

	A	C	D	E	H	I	J	K	L	
1	n	beseda	lema	MSD	amb	Razvezano	Oz	levi kontekst	beseda	desni kontekst
2		<div>								
3		<p>								
4		<s>								
5	4	Slavko	Slavko	slmei	2/2	hoški	ednina	imenovnik		
6	5	Dragovan	Dragovan	pkomein	2	a	imenovnik	-določnost		
7	6	,								
8	7	župan	župan	Somei	1/2	hoški	ednina	imenovnik		
9	8	občine	občina	Sozer	3/5	ne ženski	ednina	rodnik		
10	9	Metlika	Metlika	Sizei	1/2	ženske	ednina	imenovnik		
11	10	:								
12	11	Se	še	L	1			Členek		
13	12	nikoli	nikoli	Rso	1/7	Pristav	splošni	osnovnik		
14	13	me	jaz	Zop-et---	3	nastavka	samoostalniški			
15	14	ni	biti	IGvpat-e-d	1	lik	trejra	ednina	zanimani	občina Metlika : Se nikoli me ni bilo tako strah pod
16	15	bilo	biti	IGvdr-est-	2	lik	ednina	srednji	človek	ne Metlika : Se nikoli me ni bilo tako strah podpisat
17	16	tako	tako	Rso	3/7	Pristav	splošni	osnovnik		
18	17	strah	strah	Somei	2/2	hoški	ednina	imenovnik		
19	18	podpisati	podpisati	IGpn-----	1	enski	nedoločnik	dovršni	nikoli me ni bilo tako strah	podpisati
20	19	kakšne	kakšen	Zv-zer----	6/8	dnina	rodnik	pridevniki		kakšne pogodbe ku
21	20	pogodbe	pogodba	Sozer	3	ne ženski	ednina	rodnik		
22	21	kot	kot	Dpel	6/8	g	enostaven	imenovnik		
23	22	prav	práv	Rso	4	Pristav	splošni	osnovnik		
24	23	pogodbo	pogodba	Sozet	2	ne ženski	ednina	tožnik		
25	24	za	za	Dpet	2/3	redlog	enostaven	tožnik		
26	25	virtino	virtina	Sozet	2	ne ženski	ednina	tožnik		
27	26	.								
28	27	<s>								
29	28	Podpisuješ	podpisovati	IGvpsde--n	1	ina	nezanikani	nedovršni		Podpisuješ nekaj, za kar dejaj
30	29	nekaj	nekaj	Zntset----	2/2	na	tožnik	samoostalniški		
31	30	.								

Existing software for annotation

- **Excel** is a popular choice
- **FLEX**: proprietary software built and owned by SIL
 - Automates some parsing and tagging, links cultural/semantic information to annotated corpora, can extract concordances

The screenshot displays the 'Text' window in the 'Guuzacápin Xinka - FieldWorks Language Explorer' application. The window title is 'Birth of Tuuzi' (A) - 1972. The main area shows a list of words and their corresponding morphemes and glosses. The interface includes a menu bar (File, Edit, View, Data, Insert, Format, Tools, Parser, Window, Help) and a toolbar. On the left, there is a sidebar with 'Texts & Words' and 'Lexicon' sections. The main text area contains the following data:

Info	Baseline	Class	Analysis	Tagging	Part View	Test Chart								
1.1	Word	bunoo	puus	entences	na	huuzak	naavula	nal	hooyoy'	na'y/aa'a	i	aku'	ku'ky	
	Morphemes	bunoo	puus	entences	na	huuzak	naavuu	-la	hooy'o'-y'	na'y/aa'a	i	aku'	ku'ky	
	Word Gloss	+++			the	man	pretends		+++	+++	and	go	+++	
	Word Cst.	prt			def	n	adj				conjunction	v		
	teruuva	á'al	tu'xy	nah										
	teruuva	á'al	tu'a'-y'	nah										
	uqoye	om.a	+++	he, she, it										
	n	adj	+++	pro										
1.2	Word	á'al	á'al	á'al	da	a	da	cada	ke	aku'				
	Morphemes	á'al	á'al	á'al	da	a	da	cada	ke	a-ku'				
	Word Gloss	om.a	om.a	om.a	day, hot	ah	day, hot	each	that	go				
	Word Cst.	adj	adj	adj	adj	prt	n	adj	rel	v				
1.3	Word	y	na	ay'aa'a	mu'na	nah	na	na	na	han ali	hin	mu'ta'a	teenu'	pero
	Morphemes	y	na	ay'aa'a	mu'na	nah	na	na	na	han ali	hin	mu'ta'a	teenu'	pero
	Word Gloss	and	the	woman	say	he, she, it	say	he, she, it	he, she, it	why?	no	+++	a lot	+++
	Word Cst.	conjunction	def	n	v	pro	v	pro	adv	verbprt	+++	adv		+++
	nal	hin	atew'	tu'xy'	nah	teenu'								
	nal	hin	a-teeo'	tu'a'-y'	nah	teenu'								
	past incomplete	no	want, de	+++	he, she, it	a lot								
	verbprt	verbprt	v	+++	pro	adv								
1.4	Word	entences	na	ay'aa'a	pu'xy	wa'ta'a	pu'xy	boloh	u'u'lu	na'txy	ta'	u'	mu'xinder	
	Morphemes	entences	na	ay'aa'a	pu'a'-y'	wa'ta'a	pu'a'-y'	boloh	u'u'lu	na'txy	ta'	u'	mu'xinder	
	Word Gloss	+++	the	woman	make, do	thread	make, do	+++	thread	+++	toward here	to	against	
	Word Cst.	prt	def	n	vt	vt	vt	+++	n	+++	v	n	+++	
1.5	Word	entences	na	huuzak	waak'i	an	ke	kuy	humi'	boloh	u'u'lu	waak'i	iki	nah
	Morphemes	entences	na	huuzak	waak'i	an	ke	kuy	humi'	boloh	u'u'lu	waak'i	iki	nah
	Word Gloss	+++	the	man	is	is	is	is	is	+++	+++	+++	+++	+++
	Word Cst.	+++	the	man	is	is	is	is	is	+++	+++	+++	+++	+++

1. Requires **non-trivial ease with technology**
 - ...you are absolutely **overestimating the technological ability** of researchers, let alone of speakers.

Existing software for annotation - not so FLEx-ible

1. Requires **non-trivial technological ability**
2. Researchers often want to use the software in ways it was **not built for**
 - Multilingual/multimodal/multispeaker data

Existing software for annotation - not so FLEx-ible

1. Requires **non-trivial technological ability**
2. Researchers often want to use the software in ways it was **not built for**
 - Multilingual/multimodal/multispeaker data
 - Non-concatenative morphology

Existing software for annotation - not so FLEx-ible

1. Requires **non-trivial technological ability**
2. Researchers often want to use the software in ways it was **not built for**
 - Multilingual/multimodal/multispeaker data
 - Non-concatenative morphology
 - Templatic morphology:

k-t-b katabtu aktubu kātibun

WRITE *I wrote* *I write* *he/she who writes*

All forms must be entered as "variants" - can't describe systematic relationships

Existing software for annotation - not so FLEx-ible

1. Requires **non-trivial technological ability**
2. Researchers often want to use the software in ways it was **not built for**

- Multilingual/multimodal/multispeaker data
- Non-concatenative morphology

- Templatic morphology:

k-t-b	katabtu	aktubu	kātibun
WRITE	<i>I wrote</i>	<i>I write</i>	<i>he/she who writes</i>

All forms must be entered as "variants" - can't describe systematic relationships

- Ablaut:

mice-∅
mouse-PL

Existing software for annotation - not so FLEx-ible

1. Requires **non-trivial technological ability**
2. Researchers often want to use the software in ways it was **not built for**

- Multilingual/multimodal/multispeaker data
- Non-concatenative morphology

- Templatic morphology:

k-t-b	katabtu	aktubu	kātibun
WRITE	<i>I wrote</i>	<i>I write</i>	<i>he/she who writes</i>

All forms must be entered as "variants" - can't describe systematic relationships

- Ablaut:

mice-∅
mouse-PL

Learning and creating these workarounds requires **time and knowledge**.

Existing software for annotation - not so FLEx-ible

1. Requires **non-trivial technological ability**
2. Researchers often want to use the software in ways it was **not built for**

- Multilingual/multimodal/multispeaker data
- Non-concatenative morphology

- Templatic morphology:

k-t-b	katabtu	aktubu	kātibun
WRITE	<i>I wrote</i>	<i>I write</i>	<i>he/she who writes</i>

All forms must be entered as "variants" - can't describe systematic relationships

- Ablaut:

mice-∅
mouse-PL

Learning and creating these workarounds requires **time and knowledge**.

Often relies on **exporting and re-importing** to Python, ELAN, LaTeX, raising the technical barrier

Existing software for annotation - not so FLEx-ible

1. Requires **non-trivial technological ability**
2. Researchers often want to use the software in ways it was **not built for**
3. **Closed source proprietary software**: technically capable people can't implement or share improvements.
 - Particularly regrettable: hard to take advantage of **NLP and ML** technology built for aiding work on underdescribed languages.

The idea behind the solution

- **Word-and-Paradigm Morphology**
 - A more **intuitive** annotation process and software interface, allowing for increased community involvement

- **Word-and-Paradigm Morphology**
 - A more **intuitive** annotation process and software interface, allowing for increased community involvement
- **Computational methods and machine learning**
 - **Automating** the initial steps of analysis
 - **Suggesting** most informative data points to analyse next
 - Automatically **extending** the annotation and analysis to new data

- **Word-and-Paradigm Morphology**
 - A more **intuitive** annotation process and software interface, allowing for increased community involvement
- **Computational methods and machine learning**
 - **Automating** the initial steps of analysis
 - **Suggesting** most informative data points to analyse next
 - Automatically **extending** the annotation and analysis to new data
- Software that is **modular and open source**

Morphemic approaches to morphology

- Glossing and traditional morphosyntactic analysis are based on a **morphemic conception** of language

Morphemic approaches to morphology

- Glossing and traditional morphosyntactic analysis are based on a **morphemic conception** of language
 - Morphology is about **carving words up**
 - Describing a language's morphology amounts to making an **inventory of its FORM = MEANING pairings**.

Morphemic approaches to morphology

- Glossing and traditional morphosyntactic analysis are based on a **morphemic conception** of language
 - Morphology is about **carving words up**
 - Describing a language's morphology amounts to making an **inventory of its FORM = MEANING pairings**.

-s = PLURAL; -ed = PAST; -er = AGENT

workers = work- + -er + -s



+ AGENT + PLURAL

Morphemic approaches to morphology - limitations

- For several reasons, these reductionist approaches are **empirically inadequate**

Morphemic approaches to morphology - limitations

- For several reasons, these reductionist approaches are **empirically inadequate**
 - Not always possible or easy to establish **morpheme boundaries**:
driv-er? drive-er? drive-r?

Morphemic approaches to morphology - limitations

- For several reasons, these reductionist approaches are **empirically inadequate**
 - Not always possible or easy to establish **morpheme boundaries**:
driv-er? drive-er? drive-r?
 - Some bits of form have **no meaning**:
natur-al *sens-u-al*, *fact-u-al*

Morphemic approaches to morphology - limitations

- For several reasons, these reductionist approaches are **empirically inadequate**
 - Not always possible or easy to establish **morpheme boundaries**:
driv-er? drive-er? drive-r?
 - Some bits of form have **no meaning**:
natur-al *sens-u-al*, *fact-u-al*
 - Some bits of meaning have **no form attached**:
sheep (sg) vs *sheep* (pl)

Morphemic approaches to morphology - limitations

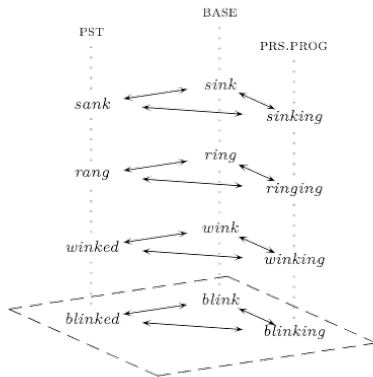
- For several reasons, these reductionist approaches are **empirically inadequate**
 - Not always possible or easy to establish **morpheme boundaries**:
driv-er? drive-er? drive-r?
 - Some bits of form have **no meaning**:
natur-al *sens-u-al*, *fact-u-al*
 - Some bits of meaning have **no form attached**:
sheep (sg) vs *sheep* (pl)
 - The whole is often **more than the sum of its parts**:
glasses ≠ *glass*+PLURAL

Morphemic approaches to morphology - limitations

- For several reasons, these reductionist approaches are **empirically inadequate**
 - Not always possible or easy to establish **morpheme boundaries**:
driv-er? drive-er? drive-r?
 - Some bits of form have **no meaning**:
natur-al *sens-u-al*, *fact-u-al*
 - Some bits of meaning have **no form attached**:
sheep (sg) vs *sheep* (pl)
 - The whole is often **more than the sum of its parts**:
glasses ≠ *glass*+PLURAL
- **Word-based** approaches to morphology see the word as the **smallest unit of meaning**, rather than the morpheme, for the reasons above.

Doing morphology with word-based units

Define a word's meaning by the **place it occupies in the system**, relative to other words.



Morphology is about establishing **parallel analogical relationships between words**, and looking at the system as a whole.

Building up a picture of the system

Collect sets with **parallel relationships** of form and meaning

sink ~ sunk

ring ~ rung

*silk ~ *sulk

Building up a picture of the system

Collect sets with **parallel relationships** of form and meaning

sink ~ sunk

ring ~ rung

*silk ~ *sulk

These relationships can **span all the lexicon**

sink ~ sunk ~ sinkable ~ ...

ring ~ rung ~ ringable ~ ...

Building up a picture of the system

Collect sets with **parallel relationships** of form and meaning

sink ~ sunk

ring ~ rung

*silk ~ *sulk

These relationships can **span all the lexicon**

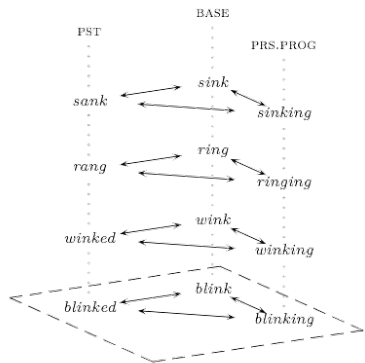
sink ~ sunk ~ sinkable ~ ...

ring ~ rung ~ ringable ~ ...

Morphological families are built up and aligned, starting from pairwise relationships

Word and Paradigm morphology

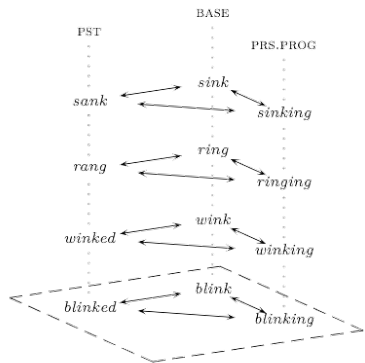
- Establishing **parallel relationships of form and meaning** between words



- The **word is the smallest unit**
 - Defined by its place in a system of contrasts, not by its component parts

Word and Paradigm morphology

- Establishing **parallel relationships of form and meaning** between words



- The **word is the smallest unit**
 - Defined by its place in a system of contrasts, not by its component parts
- Concepts like **paradigm cell** or **lexeme** are emergent
 - The result of establishing contrasts and similarities between words along different dimensions

The goal

Paradigmatic morphological analysis from documentary corpora

Paradigmatic morphological analysis from documentary corpora

- Computational **automation of the initial steps** of the analysis

Paradigmatic morphological analysis from documentary corpora

- Computational **automation of the initial steps** of the analysis
- The **annotator corrects** the initial analysis
 - Simple task: same or different?

Paradigmatic morphological analysis from documentary corpora

- Computational **automation of the initial steps** of the analysis
- The **annotator corrects** the initial analysis
 - Simple task: same or different?
- **Active learning**
 - Updates the analysis after each annotator correction
 - Directs the annotator's attention to the most informative data points

The workflow

Step 1: Automated paradigm discovery

- Corpus of collected texts
 - + list of target lemmas
 - + unsupervised model (Jin et al. 2020)
 - = **initial unlabeled paradigms**

Step 1: Automated paradigm discovery

- Corpus of collected texts
 - + list of target lemmas
 - + unsupervised model (Jin et al. 2020)
 - = **initial unlabeled paradigms**
- System searches a documentary corpus to identify related forms for each lexeme and **group surface forms into paradigms**

	Cell					
Lexeme	1	2	3	4	5	6
HEAR	hear	heard	-	hearing	heart	-
HELP	help	-	helped	helping	-	helps
			...			

Step 1: Automated paradigm discovery

The Model:

Unsupervised Morphological Paradigm Completion (Jin et al., 2020)

- Official baseline for **SIGMORPHON 2020** shared task (Task 2)

Step 1: Automated paradigm discovery

The Model:

Unsupervised Morphological Paradigm Completion (Jin et al., 2020)

- Official baseline for **SIGMORPHON 2020** shared task (Task 2)
- Uses **LONGEST COMMON SUBSTRING (LCS)** to identify paradigm candidates for each lemma input

Step 1: Automated paradigm discovery

The Model:

Unsupervised Morphological Paradigm Completion (Jin et al., 2020)

- Official baseline for **SIGMORPHON 2020** shared task (Task 2)
- Uses **LONGEST COMMON SUBSTRING (LCS)** to identify paradigm candidates for each lemma input
- Computes **EDIT TREES** to identify recurrent changes in surface forms across paradigms and defines paradigm cells accordingly

Step 1: Automated paradigm discovery

The Model:

Unsupervised Morphological Paradigm Completion (Jin et al., 2020)

- Official baseline for **SIGMORPHON 2020** shared task (Task 2)
- Uses **LONGEST COMMON SUBSTRING (LCS)** to identify paradigm candidates for each lemma input
- Computes **EDIT TREES** to identify recurrent changes in surface forms across paradigms and defines paradigm cells accordingly

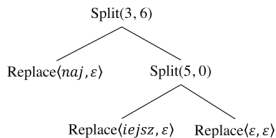


Figure 2: Visualization of the EDIT TREE constructed from *najtrudniejszy* to *trudny* (Chrupała, 2008).

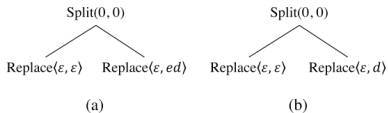


Figure 3: Visualization of the EDIT TREES representing (a) *work* \mapsto *worked* and (b) *continue* \mapsto *continued*.

Step 1: Automated paradigm discovery

The model makes a number of simplifying assumptions:

Step 1: Automated paradigm discovery

The model makes a number of simplifying assumptions:

1. Analyzes **inflection** as distinct from **derivation**
 - **Inflection:** *dance* (V.PRS) \sim *danced* (V.PST)
 - **Derivation:** *dance* (V.PRS) \sim *dancer* (N.AGENT)

Step 1: Automated paradigm discovery

The model makes a number of simplifying assumptions:

1. Analyzes **inflection** as distinct from **derivation**
 - **Inflection:** *dance* (V.PRS) \sim *danced* (V.PST)
 - **Derivation:** *dance* (V.PRS) \sim *dancer* (N.AGENT)
2. Assumes exactly **one form per paradigm cell**
 - But variation is common across languages!
e.g., English *dreamed/dreamt*

Step 1: Automated paradigm discovery

The model makes a number of simplifying assumptions:

1. Analyzes **inflection** as distinct from **derivation**
 - **Inflection:** *dance* (V.PRS) \sim *danced* (V.PST)
 - **Derivation:** *dance* (V.PRS) \sim *dancer* (N.AGENT)
2. Assumes exactly **one form per paradigm cell**
 - But variation is common across languages!
e.g., English *dreamed/dreamt*
3. Assumes exactly **one paradigm cell per form**
 - This is also often not the case!
e.g., English *read* can indicate 1SG.PRS, 1SG.PST, 2SG.PRS, 1PL.PRS...

Step 1: Automated paradigm discovery

The model makes a number of simplifying assumptions:

1. Analyzes **inflection** as distinct from **derivation**
 - **Inflection:** *dance* (V.PRS) \sim *danced* (V.PST)
 - **Derivation:** *dance* (V.PRS) \sim *dancer* (N.AGENT)
2. Assumes exactly **one form per paradigm cell**
 - But variation is common across languages!
e.g., English *dreamed/dreamt*
3. Assumes exactly **one paradigm cell per form**
 - This is also often not the case!
e.g., English *read* can indicate 1SG.PRS, 1SG.PST, 2SG.PRS, 1PL.PRS...
4. Assumes **concatenative** relationships and **consistent affix ordering**

Step 1: Automated paradigm discovery

The model's output:

	Cell					
Lexeme	1	2	3	4	5	6
HEAR	hear	heard	-	hearing	heart	-
HELP	help	-	helped	helping	-	helps
			...			

Step 1: Automated paradigm discovery

The model's output:

	Cell					
Lexeme	1	2	3	4	5	6
HEAR	hear	heard	-	hearing	heart	-
HELP	help	-	helped	helping	-	helps
			...			

... it's a start! **Humans can help** :)

Step 2: Same or different? (Lexemes)

- Automatically extract examples of each **form in context** from the corpus
- The annotator marks **items that don't belong with the others**

File		
Lexicon Analogies Paradigms Texts		
Lexemes	Surface Forms	Concordances
DANCE	HEAR	? ...you're still going to hear them.
DRIVE	HEARS	? She thought she could hear Gomez laughing.
LIVE	HEARD	✗ ...signalling of problems of hearing and understanding.
HEAR	HEARING	✗ ...gray marble mausoleum at the heart of the city.
WORK	HEART	.
.	.	.
.	.	.
.	.	.
.	.	.

Step 3: Same or different? (Analogies)

- **Pairwise analogy** groups forms instantiating the same **paradigm cell**

The screenshot shows a software interface with a menu bar (File, Lexicon, Analogies, Paradigms, Texts) and a main workspace. The workspace is divided into three panels: 'Analogies', 'Concordances', and 'Xing'. The 'Analogies' panel lists pairs like 'X ~ Xment', 'X ~ Xer', 'X ~ Xing', 'X ~ Xed', and 'X ~ X'. The 'Concordances' panel is split into two columns: 'X' and 'Xing'. The 'X' column shows examples like 'We publish these ..', 'If we learn how...', and 'We go regularly to...'. The 'Xing' column shows examples like 'Time for publishing ..', 'Second language learning is ...', and 'She's not going to like ...'. Green checkmarks and question marks are used to mark the concordance pairs.

Analogies	Concordances	
	X	Xing
X ~ Xment		
X ~ Xer	✓ We publish these ..	Time for publishing ..
X ~ Xing	? If we learn how...	Second language learning is ...
X ~ Xed	? We go regularly to...	She's not going to like ...
X ~ X	.	.
.	.	.
.	.	.
.	.	.
.	.	.

- The annotator's task is the same: mark pairs that don't belong, and confirm those that do

The result: Unlabeled paradigms

File

LEXICON ANALOGIES PARADIGMS TEXTS

Lexemes

Search...

Show all words... ^

- look
- how
- focus
- involving
- it
- describes
- focusing
- focused
- works

Senses

View, edit and create word senses... v

Paradigm

Search...

looking	hearing	working		offering	playing
look	hear	work	describe	offer	play
looked		worked		offered	played
looks		works	describes	offers	plays

Experiments and results

- **Universal Dependencies** datasets for **English** and **Croatian** provide a gold standard for evaluation
- **Annotators**: 4 linguists (2 per language), fluent English speakers
 - English: **upper estimate** of model + annotator performance
 - Croatian: **unfamiliar language**
- 30 minutes per task: **lexeme groupings** + **cell groupings**

English & Croatian Results

Lexeme				Cell			
	Acc.	Marked	Corr.		Acc.	Marked	Corr.
English				English			
Base	81%	-	-	Base	67%	-	-
A1	84%	58	50	A1	97%	129	120
A2	83%	43	33	A2	94%	119	108
Croatian				Croatian			
Base	66%	-	-	Base	90%	-	-
A3	67%	19	19	A3	90%	8	-1
A4	66%	12	12	A4	90%	28	16

English & Croatian Results

Lexeme				Cell			
	Acc.	Marked	Corr.		Acc.	Marked	Corr.
English				English			
Base	81%	-	-	Base	67%	-	-
A1	84%	58	50	A1	97%	129	120
A2	83%	43	33	A2	94%	119	108
Croatian				Croatian			
Base	66%	-	-	Base	90%	-	-
A3	67%	19	19	A3	90%	8	-1
A4	66%	12	12	A4	90%	28	16

English & Croatian Results

	Lexeme			Cell			
	Acc.	Marked	Corr.		Acc.	Marked	Corr.
English				English			
Base	81%	-	-	Base	67%	-	-
A1	84%	58	50	A1	97%	129	120
A2	83%	43	33	A2	94%	119	108
Croatian				Croatian			
Base	66%	-	-	Base	90%	-	-
A3	67%	19	19	A3	90%	8	-1
A4	66%	12	12	A4	90%	28	16

Case Study: Wao Terero

Wao Terero provides a demonstration of this workflow in the field.

- Linguistic isolate spoken in **Ecuadorian Amazon**
 - Estimated 1,200-3,000 speakers
 - No standard orthography
- **Collaboration** with native speakers (Spanish-Wao bilinguals)

- Two **native speaker consultants** from the Wao community of Geyepade served as annotators.
 - Neither consultant had taken a course in linguistics

- Two **native speaker consultants** from the Wao community of Geyepade served as annotators.
 - Neither consultant had taken a course in linguistics
- 10 minutes of training, with Spanish verbal paradigms
 - annotate as many items (lexemes and paradigm cells) as possible within **1 hour**

Case Study: Wao Terero

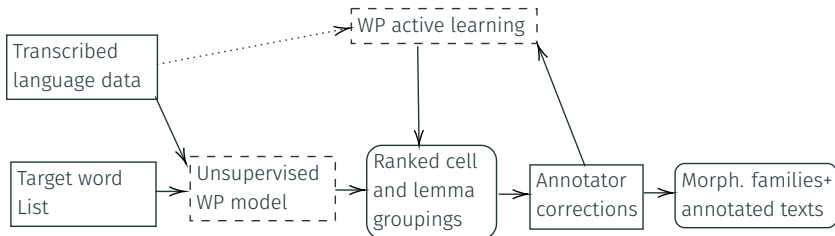
- Two **native speaker consultants** from the Wao community of Geyepade served as annotators.
 - Neither consultant had taken a course in linguistics
- 10 minutes of training, with Spanish verbal paradigms
 - annotate as many items (lexemes and paradigm cells) as possible within **1 hour**
- Annotators found the task **understandable** and **interesting**, with high inter-annotator agreement across annotated examples



Copot et al. (2022)
A Word-and-Paradigm Workflow for Fieldwork Annotation

In Development...

Implementing the Full Workflow



Ranking + Active Learning

- Warm start a supervised classifier using the unsupervised model's output as **silver data**
- System uses annotator's corrections for **active learning**

Lexemes	Surface Forms	Concordances
DANCE	HEAR	? ...you're still going to hear them.
DRIVE	HEARS	? She thought she could hear Gomez laughing.
LIVE	HEARD	X ...signalling of problems of hearing and understanding.
HEAR	HEARING	X ...gray marble mausoleum at the heart of the city.
WORK	HEART	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.

- Items are reordered in real time for **efficient use of annotator time**

Ranking + Active Learning

- Analysis is pairwise-relational over sets of **formal**, **structural**, and **semantic** properties
- Lexeme and cell groupings **emerge** from the existence of shared relationships

The screenshot shows a software interface with a menu bar at the top containing 'File', 'Lexicon', 'Analogies', 'Paradigms', and 'Texts'. The 'Analogies' menu is currently selected, and the interface is divided into two main panels. The left panel, titled 'Analogies', lists several pairs: 'X ~ Xment', 'X ~ Xer', 'X ~ Xing', 'X ~ Xed', and 'X ~ X', with vertical ellipses below. The right panel, titled 'Concordances', is further divided into two sub-columns: 'X' and 'Xing'. Under the 'X' column, there are three entries: 'We **publish** these ..' (with a green checkmark icon), 'If we **learn** how...' (with a green question mark icon), and 'We **go** regularly to...' (with a green question mark icon), followed by vertical ellipses. Under the 'Xing' column, there are three entries: 'Time for **publishing** ..', 'Second language **learning** is ...', and 'She's not **going** to like ...', followed by vertical ellipses.

Improving Unsupervised Paradigm Discovery

- Current methods almost exclusively rely on **formal** relationships
 - We can **incorporate context features** from the corpus to bolster **semantic** representations

Improving Unsupervised Paradigm Discovery

- Current methods almost exclusively rely on **formal** relationships
 - We can **incorporate context features** from the corpus to bolster **semantic** representations
- Still highly biased towards **concatenative** relationships
 - Can we leverage initial output to identify additional **non-concatenative** alternations?

Improving Unsupervised Paradigm Discovery

- Current methods almost exclusively rely on **formal** relationships
 - We can **incorporate context features** from the corpus to bolster **semantic** representations
- Still highly biased towards **concatenative** relationships
 - Can we leverage initial output to identify additional **non-concatenative** alternations?
- Want to incorporate derivational and agglutinative relationships to establish networks of **morphological families**
 - **Derivational:** *build ~ rebuild; build ~ builder; rebuild ~ rebuilder*
 - **Agglutinative:** epäjärjestelmällistyttämättömyydellänsäkäänköhän
"I wonder if – even with his/her quality of not having been made unsystematized"

Conclusion

Word-and-Paradigm annotation **makes direct comparisons in context**

- **Intuitive** for untrained consultants
 - Increases **community participation**

Word-and-Paradigm annotation **makes direct comparisons in context**

- **Intuitive** for untrained consultants
 - Increases **community participation**
- Defers difficult decisions about segmentation and labeling
 - Paradigmatic analysis of **morphological system as a whole**

Benefits of the Workflow for Linguistic Fieldwork

Word-and-Paradigm annotation **makes direct comparisons in context**

- **Intuitive** for untrained consultants
 - Increases **community participation**
- Defers difficult decisions about segmentation and labeling
 - Paradigmatic analysis of **morphological system as a whole**
- **Modular architecture:**
 - Future improvements in state of the art machine learning can immediately benefit annotator

Benefits of the Workflow for Linguistic Fieldwork

Word-and-Paradigm annotation **makes direct comparisons in context**

- **Intuitive** for untrained consultants
 - Increases **community participation**
- Defers difficult decisions about segmentation and labeling
 - Paradigmatic analysis of **morphological system as a whole**
- **Modular architecture:**
 - Future improvements in state of the art machine learning can immediately benefit annotator
- Annotation output may be used for **linguistic analysis** as well as **community resource development**

Many thanks to our consultants,
Flora and Alberto Boyotai!

Thank you!
