Rethinking Tools for the Morphosyntactic Analysis of Underdocumented Languages

Sara Court[1] Maria Copot[2]

[1]The Ohio State University [2]Université 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 plan

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

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

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

Language documentation - why do it

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

Language documentation - why do it

- 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

Language documentation - why care

- 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**

*X*S = X.PLURAL

Documenting and analysing morphosyntactic structure

1. Eliciting basic **vocabulary**

squirrels

squids

cats

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

*X*S = 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

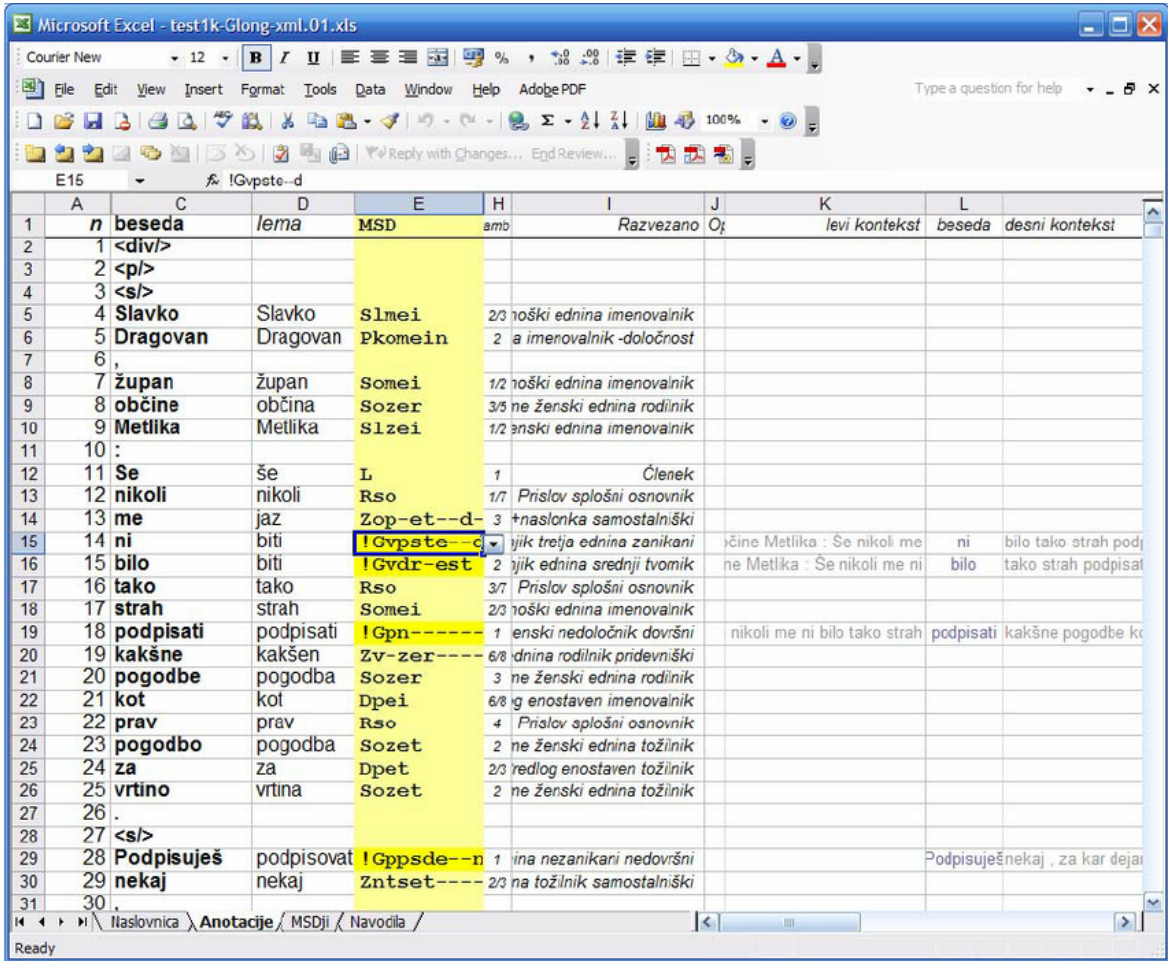
squirrel-s	would=’ve	chase-d	mice
squirrel-PL	COND=PERF	chase-PST	PL\mouse

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

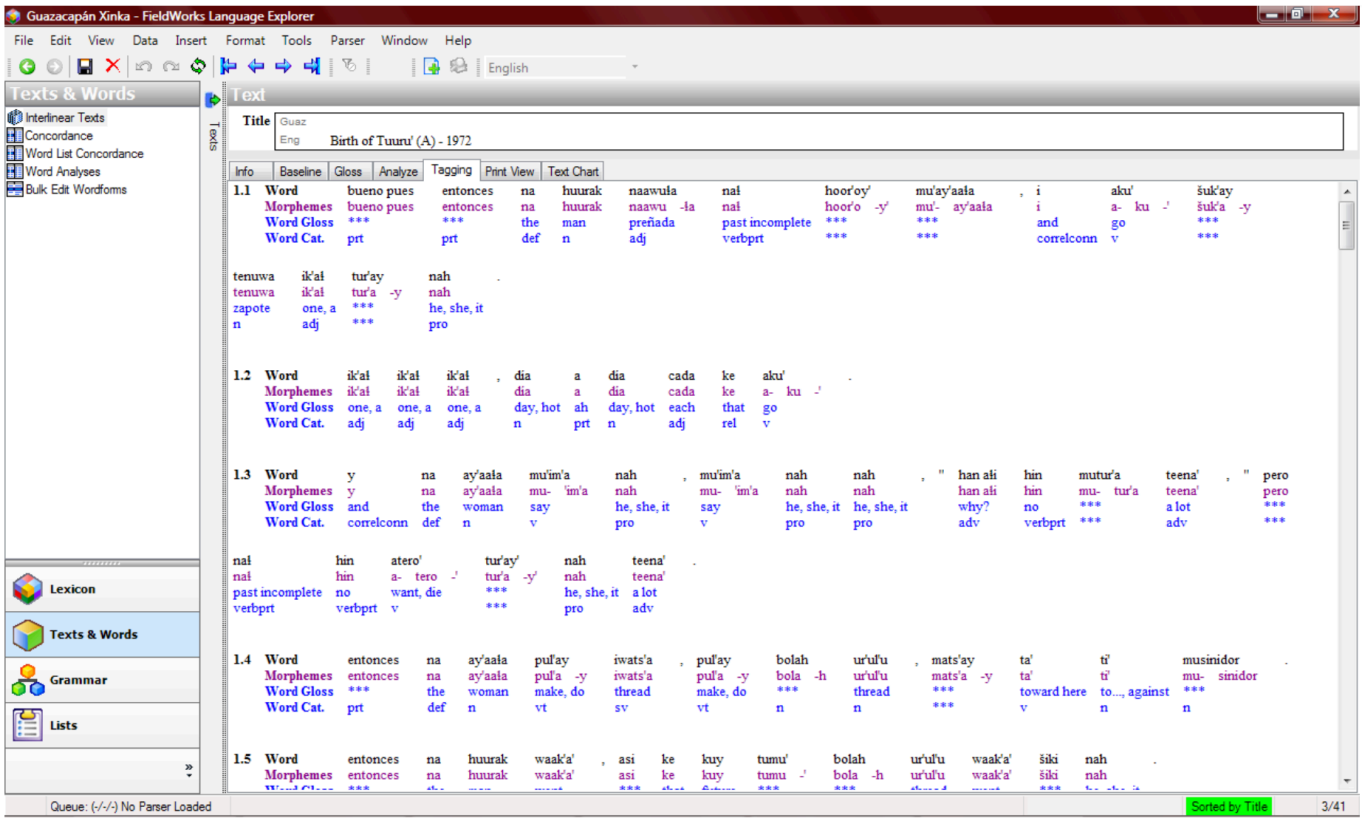
- Excel is a popular choice - a dire situation



	A	C	D	E	H	I	J	K	L	
1	*n*	**beseda**	*lema*	**MSD**	*amb*	*Razvezano*	*O*	*levi kontekst*	*beseda*	*desni kontekst*
2	1	** **								
3	2	**<p/>**								
4	3	**<s/>**								
5	4	**Slavko**	Slavko	Slmei	2/3	oški ednina imenovalnik				
6	5	**Dragovan**	Dragovan	Pkomein	2	a imenovalnik -določnost				
7	6	**,**								
8	7	**župan**	župan	Somei	1/2	oški ednina imenovalnik				
9	8	**občine**	občina	Sozer	3/5	ne ženski ednina rodilnik				
10	9	**Metlika**	Metlika	Slzei	1/2	nski ednina imenovalnik				
11	10	**:**								
12	11	**Se**	še	L	1	Členek				
13	12	**nikoli**	nikoli	Rso	1/7	Prislov splošni osnovnik				
14	13	**me**	jaz	Zop-et--d-	3	+naslonka samostalniški				
15	14	**ni**	biti	!Gvpste--d		jik tretja ednina zanikani		činе Metlika : Še nikoli me	ni	bilo tako strah pod
16	15	**bilo**	biti	!Gvdr-est	2	jik ednina srednji tvomik		ne Metlika : Še nikoli me ni	bilo	tako strah podpisat
17	16	**tako**	tako	Rso	3/7	Prislov splošni osnovnik				
18	17	**strah**	strah	Somei	2/3	oški ednina imenovalnik				
19	18	**podpisati**	podpisati	!Gpn------	1	enski nedoločnik dovršni		nikoli me ni bilo tako strah	podpisati	kakšne pogodbe k
20	19	**kakšne**	kakšen	Zv-zer----	6/8	dnina rodilnik pridevniški				
21	20	**pogodbe**	pogodba	Sozer	3	ne ženski ednina rodilnik				
22	21	**kot**	kot	Dpei	6/8	g enostaven imenovalnik				
23	22	**prav**	prav	Rso	4	Prislov splošni osnovnik				
24	23	**pogodbo**	pogodba	Sozet	2	ne ženski ednina tožilnik				
25	24	**za**	za	Dpet	2/3	redlog enostaven tožilnik				
26	25	**vrtino**	vrtina	Sozet	2	ne ženski ednina tožilnik				
27	26	**.**								
28	27	**<s/>**								
29	28	**Podpisuješ**	podpisovat	!Gppsde--n	1	ina nezanikani nedovršni			Podpisuješ	nekaj , za kar dejal
30	29	**nekaj**	nekaj	Zntset----	2/3	na tožilnik samostalniški				
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



Existing software for annotation - not so FLEx-ible

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

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

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

     All forms must be entered as ”variants” - can’t describe systematic relationships
     - Ablaut:

mice-Ø
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

Doing better

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

Doing better

- 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

Doing better

- 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 $\neq$ *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**:

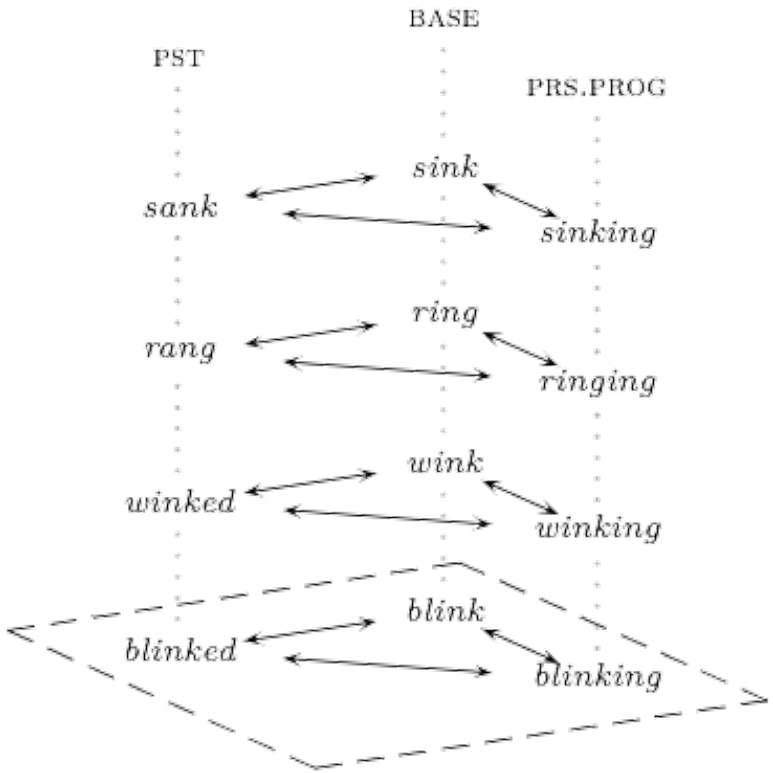
    natur-al *sens-u-al*, *fact-u-al*
  - Some bits of meaning have **no form attached**:

    sheep (sg) vs *sheep* (pl)
  - The whole is often **more than the sum of its parts**:

    glasses $\neq$ *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

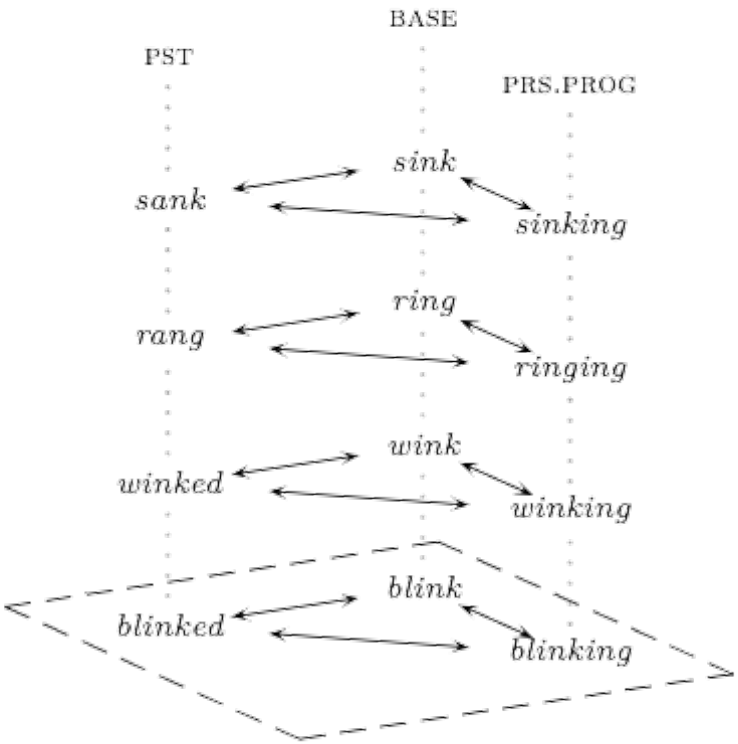
These relationships can span all the lexicon

sink ∼ sunk ∼ sinkable ∼ ...
ring ∼ rung ∼ ringable ∼ ...

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

Word and Paradigm morphology

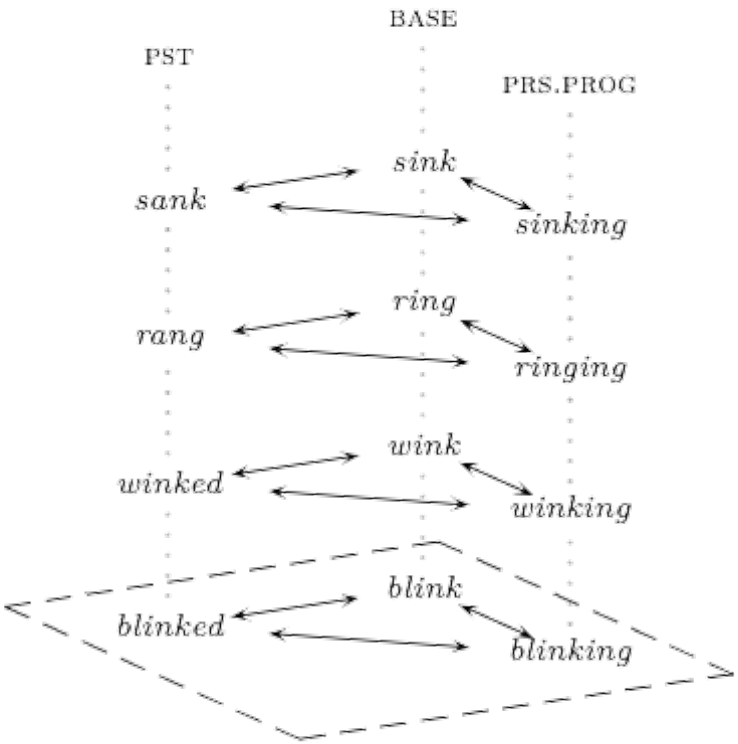
- Establishing parallel relationships of form and meaning between words



- 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

A machine-aided workflow for morphological analysis

Paradigmatic morphological analysis from documentary corpora

A machine-aided workflow for morphological analysis

Paradigmatic morphological analysis from documentary corpora

- Computational automation of the initial steps of the analysis

A machine-aided workflow for morphological 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?

A machine-aided workflow for morphological 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?
- 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

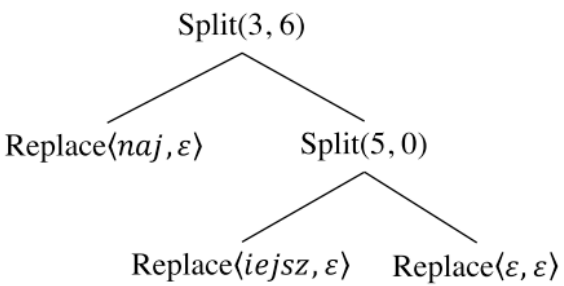


Figure 2: Visualization of the EDIT TREE constructed from *najtrudniejszy* to *trudny* (Chrupała, 2008).

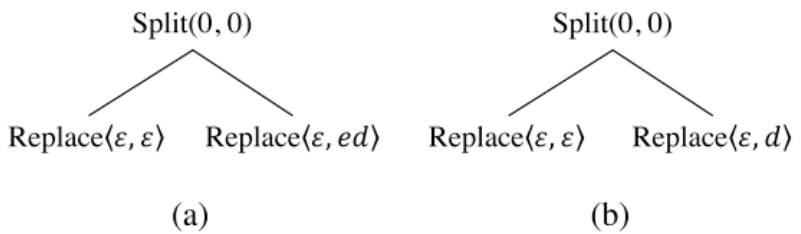


Figure 3: Visualization of the EDIT TREES representing (a) *work* ↦ *worked* and (b) *continue* ↦ *continued*.

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) ∼ *danced* (V.PST)
    - **Derivation:** *dance* (V.PRS) ∼ *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) ∼ *danced* (V.PST)
    - **Derivation:** *dance* (V.PRS) ∼ *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) ∼ *danced* (V.PST)
   - **Derivation:** *dance* (V.PRS) ∼ *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) ∼ *danced* (V.PST)
   - **Derivation:** *dance* (V.PRS) ∼ *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

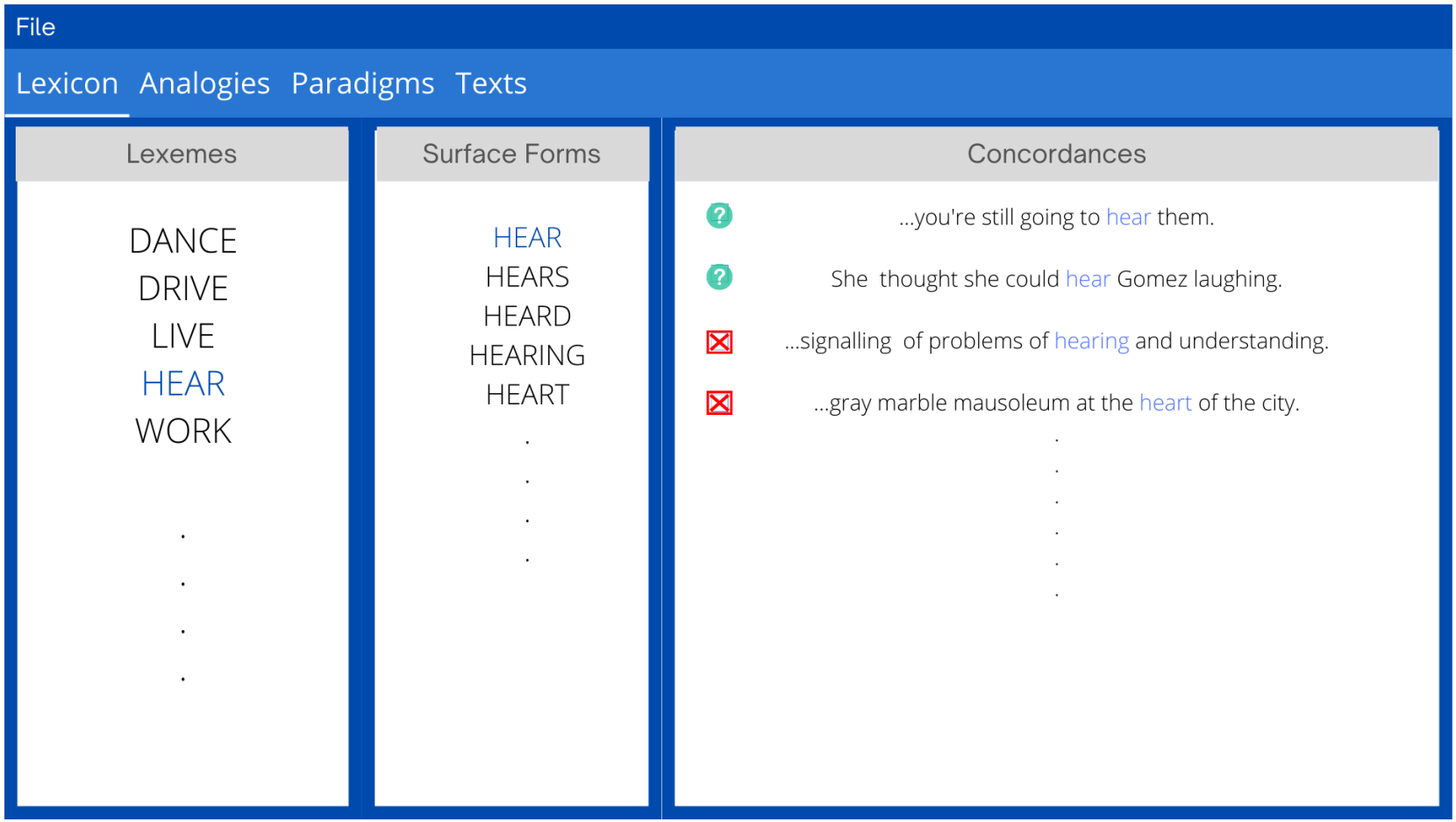
The model's output:

	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 :)

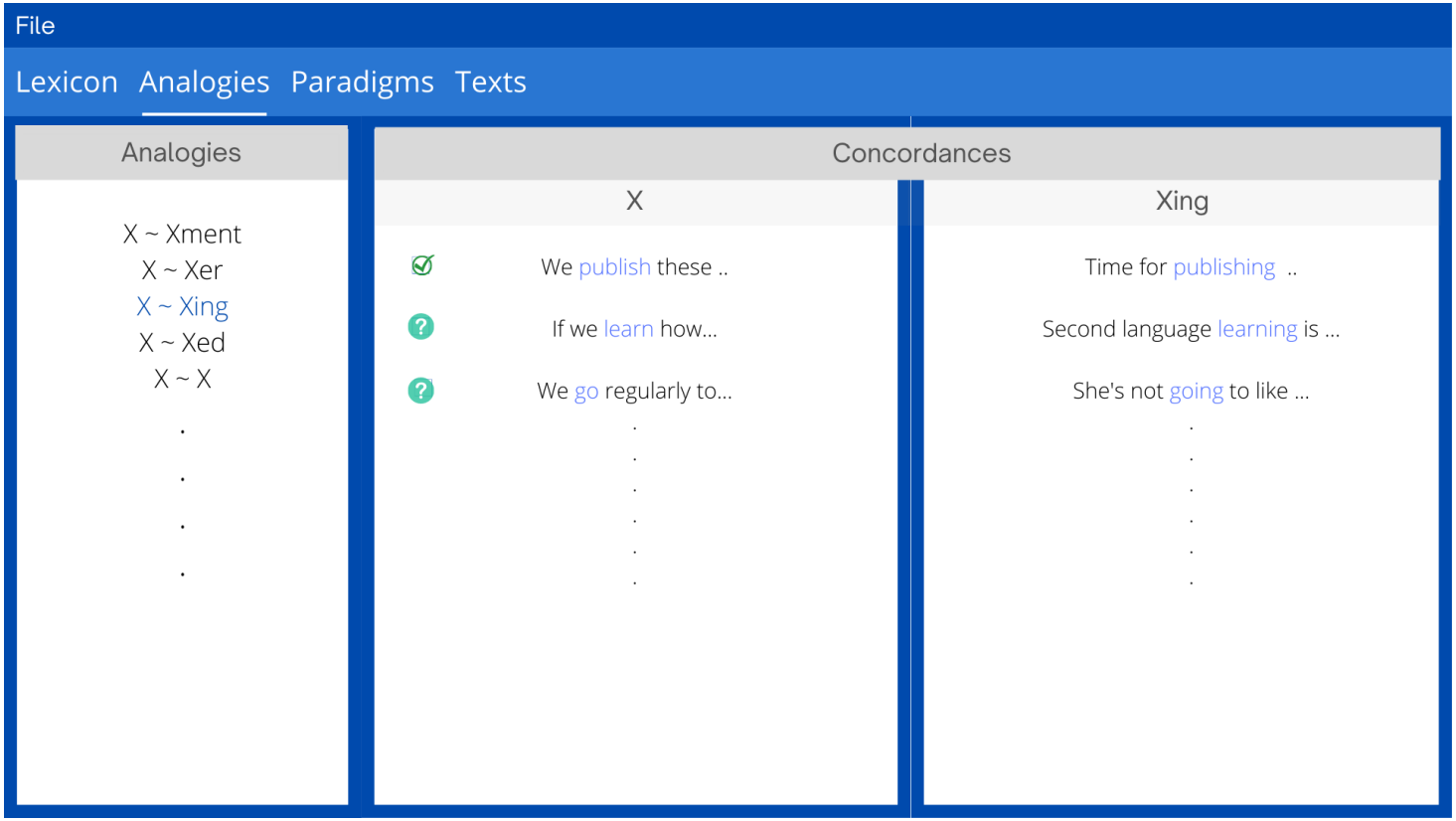
Step 2: Same or different? (Lexemes)

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



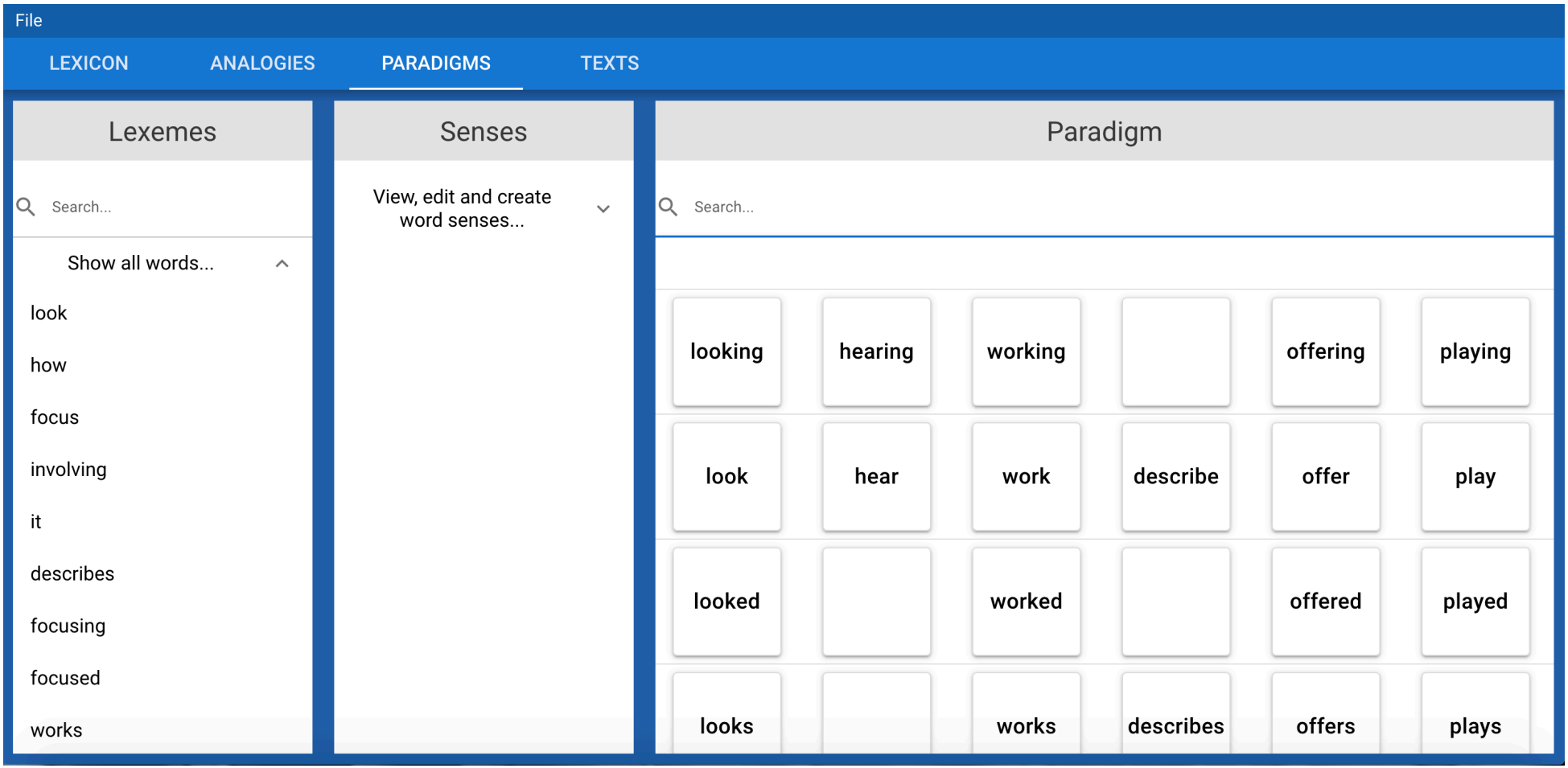
Step 3: Same or different? (Analogies)

- Pairwise analogy groups forms instantiating the same paradigm cell



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

The result: Unlabeled paradigms



Experiments and results

Experiments: English & Croatian

- 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

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

Cell

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

English & Croatian Results

Lexeme

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

Cell

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

English & Croatian Results

Lexeme

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

Cell

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

Case Study: Wao Terero

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)

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

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

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

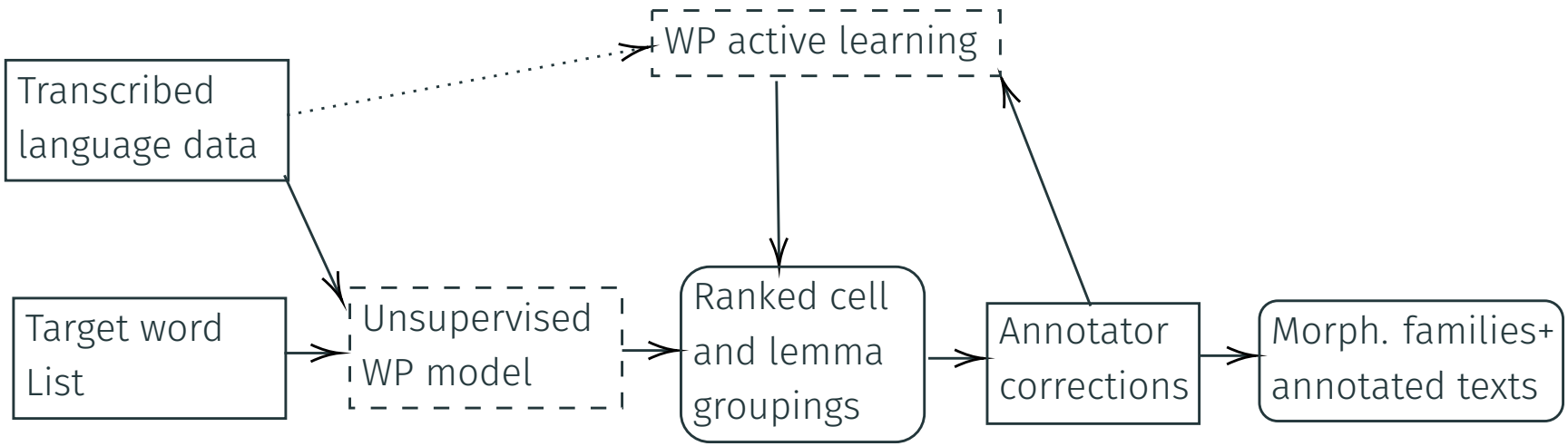
Proof of concept

Copot et al. (2022)

A Word-and-Paradigm Workflow for Fieldwork Annotation

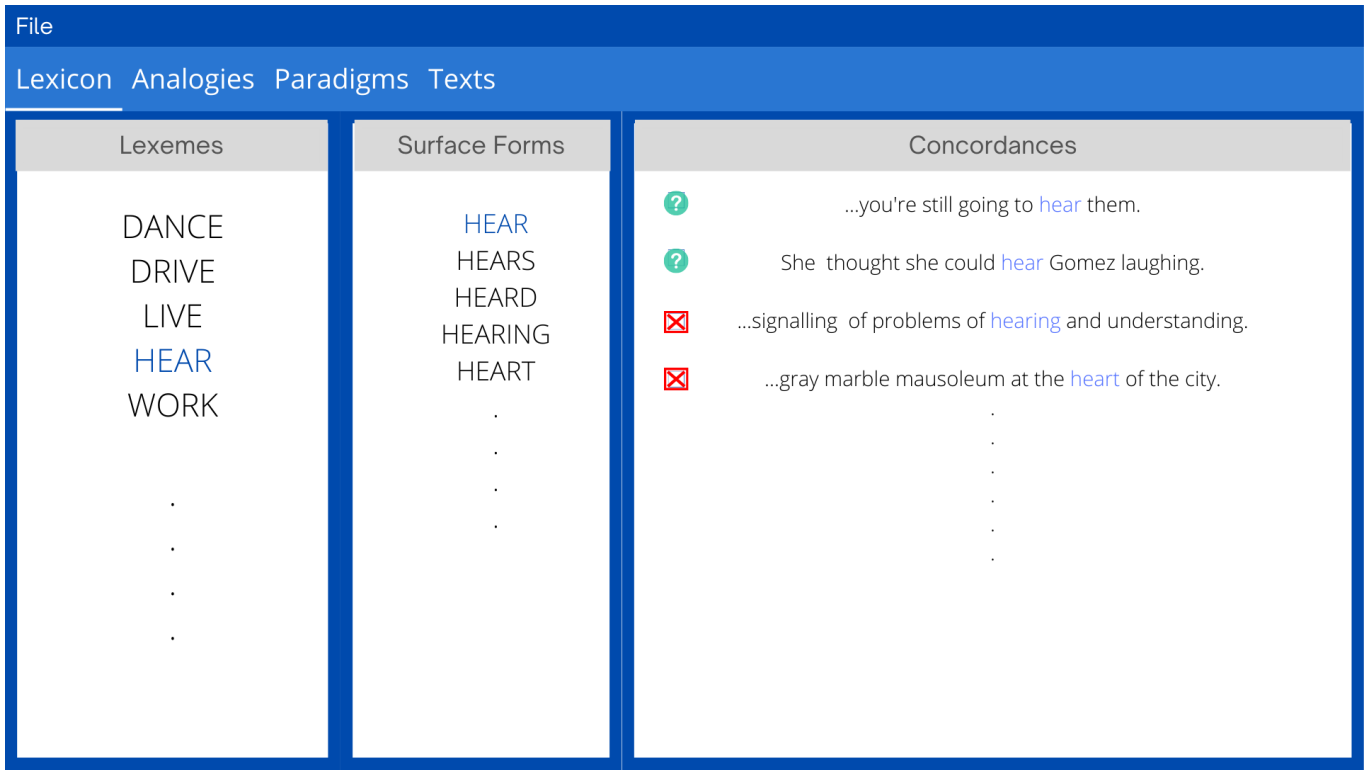
In Development...

Implementing the Full Workflow



Ranking + Active Learning

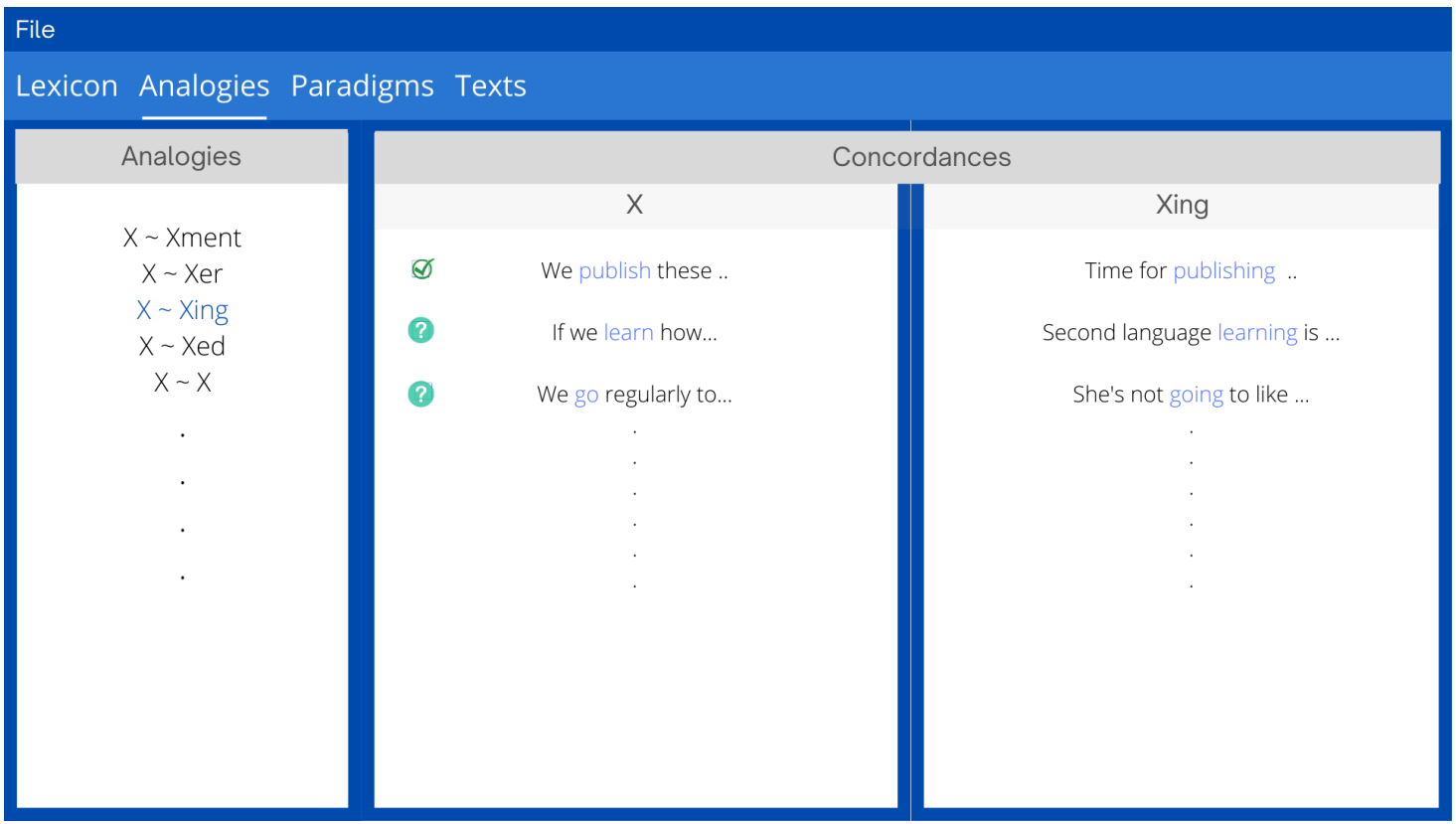
- Warm start a supervised classifier using the unsupervised model's output as silver data
- System uses annotator's corrections for active learning



- 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



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

Benefits of the Workflow for Linguistic Fieldwork

Word-and-Paradigm annotation makes direct comparisons in context

- Intuitive for untrained consultants
  - Increases community participation

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

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!