Past, Present, Future

A Computational Investigation of the Typology of Tense in **1000** Languages

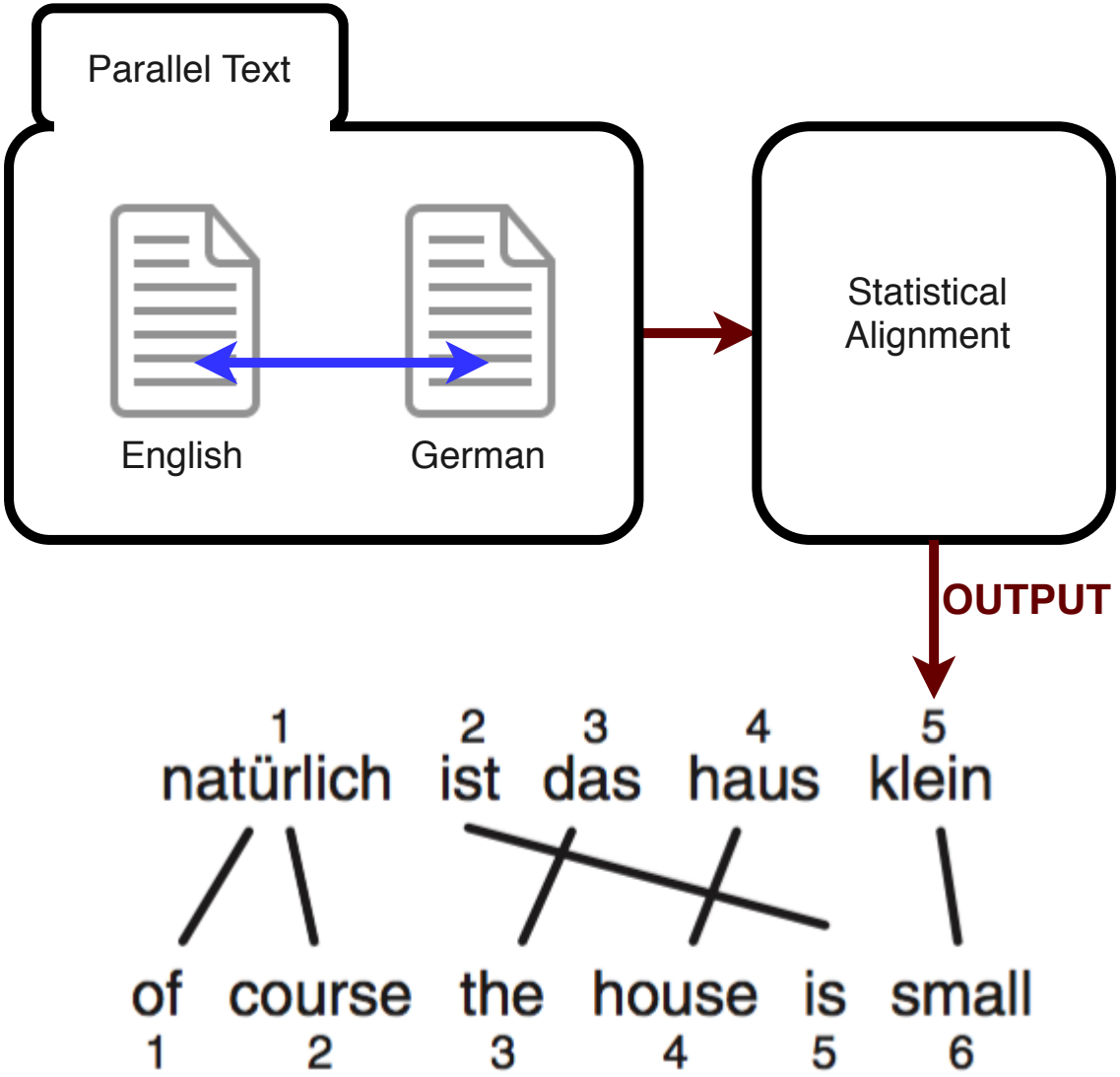
Ehsaneddin Asgari[1,2] Hinrich Schütze[1]

[1]Center for information and language processing
University of Munich - LMU

[2]Applied Science and Technology
University of California, Berkeley

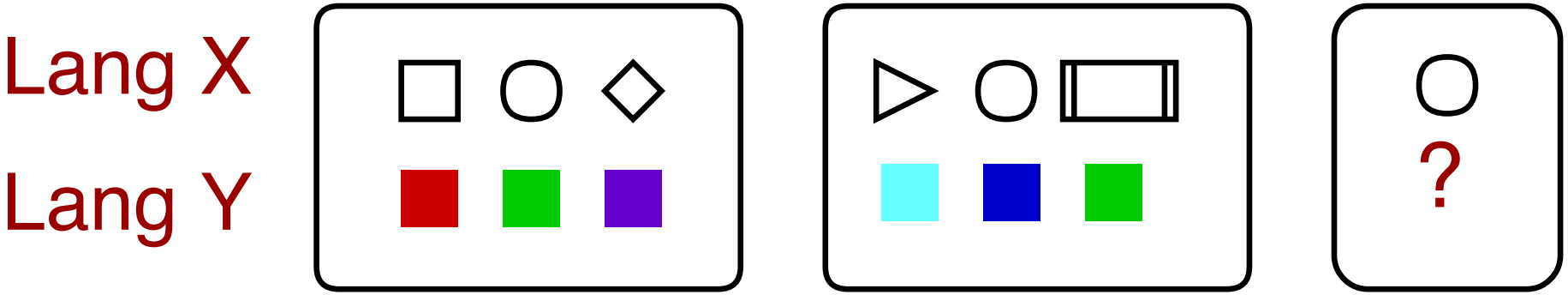
LMU, Fall 2018

Review: Statistical alignment using parallel corpus



Review: Statistical alignment using parallel corpus

Parallel text is sufficient and no linguistic knowledge is needed!



Applications: Machine translation, **linguistic resource creation**, etc.

Outline

1. Problem definition
2. Linguistic Inspiration
3. Methods and Results
   - Data set
   - SuperPivot pipeline
     - Head pivot set
     - Subword-level marker finding
     - Linguistic Analysis
   - Language clustering based on tense marking
4. Contributions

Annotation Projection: project labels via alignment

Important area of NLP research: Yarowsky et al. (2001); Spreyer and Frank (2008); Padó & Lapata (2009); Das and Petrov (2011); Agić et al. (2016).

Annotation Projection: project labels via alignment

Important area of NLP research: Yarowsky et al. (2001); Spreyer and Frank (2008); Padó & Lapata (2009); Das and Petrov (2011); Agić et al. (2016).

English Jesus knew what they pondered

French Jésus discerna ce qu'ils pensaient

Annotation Projection: project labels via alignment

Important area of NLP research: Yarowsky et al. (2001); Spreyer and Frank (2008); Padó & Lapata (2009); Das and Petrov (2011); Agić et al. (2016).

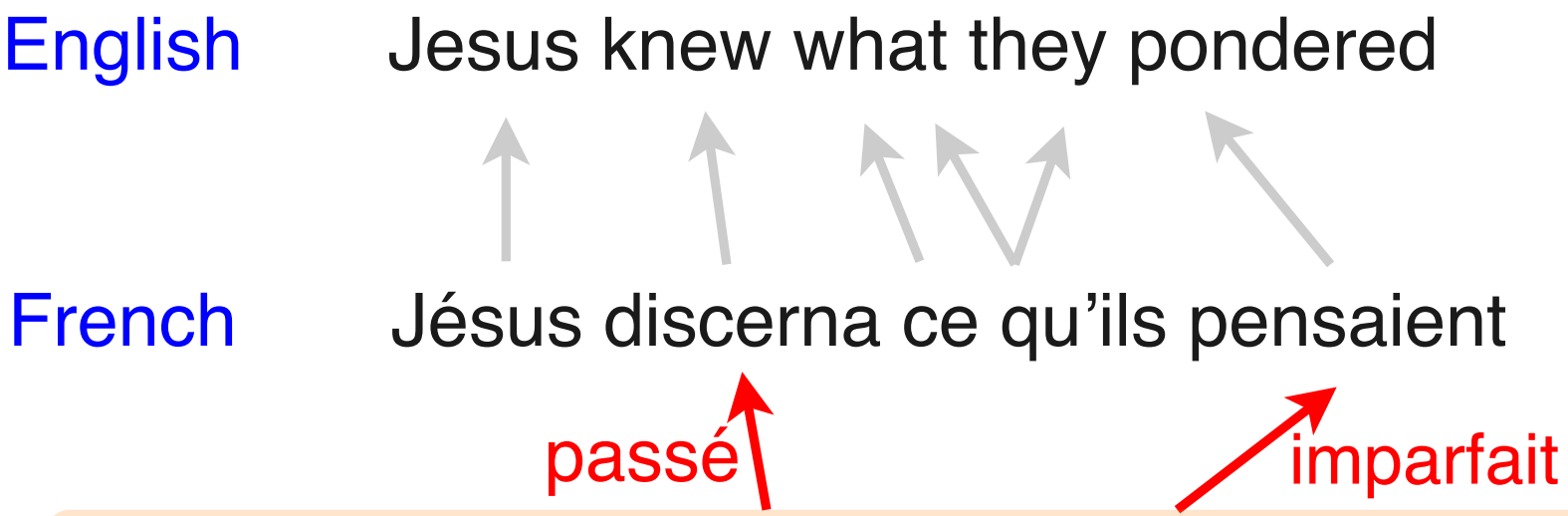
English Jesus knew what they pondered

French Jésus discerna ce qu'ils pensaient

Linguistic Annotation E.g. Part-of-speech tagging

Annotation Projection: project labels via alignment

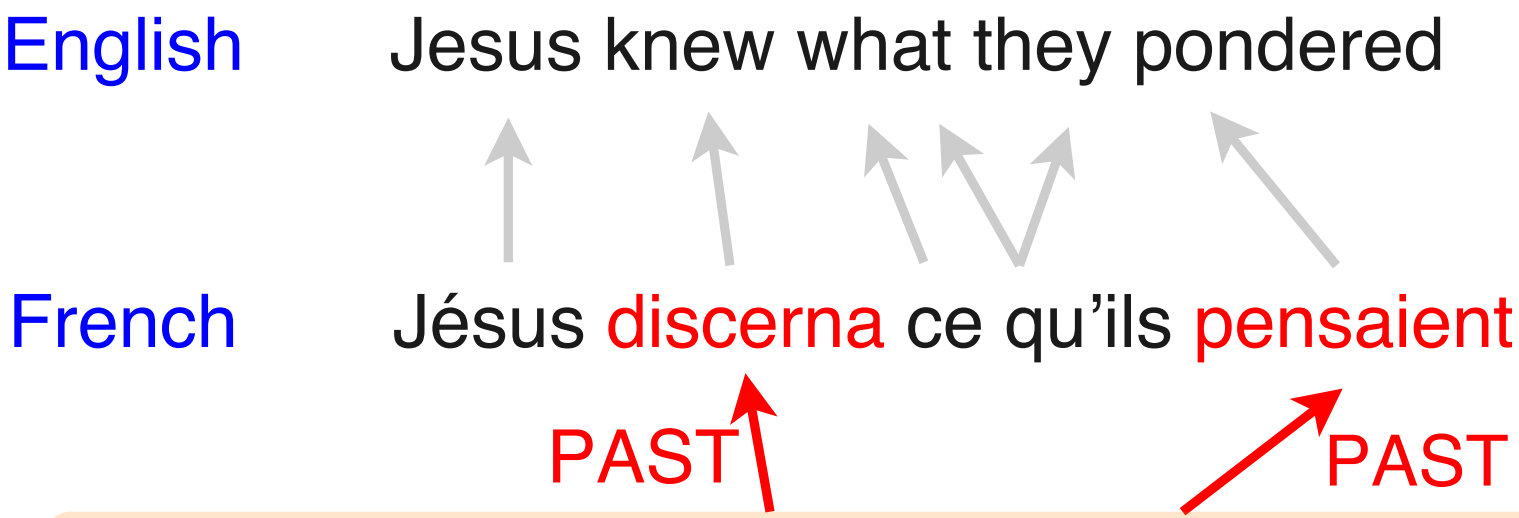
Important area of NLP research: Yarowsky et al. (2001); Spreyer and Frank (2008); Padó & Lapata (2009); Das and Petrov (2011); Agić et al. (2016).



Linguistic Annotation E.g. Part-of-speech tagging

Annotation Projection: project labels via alignment

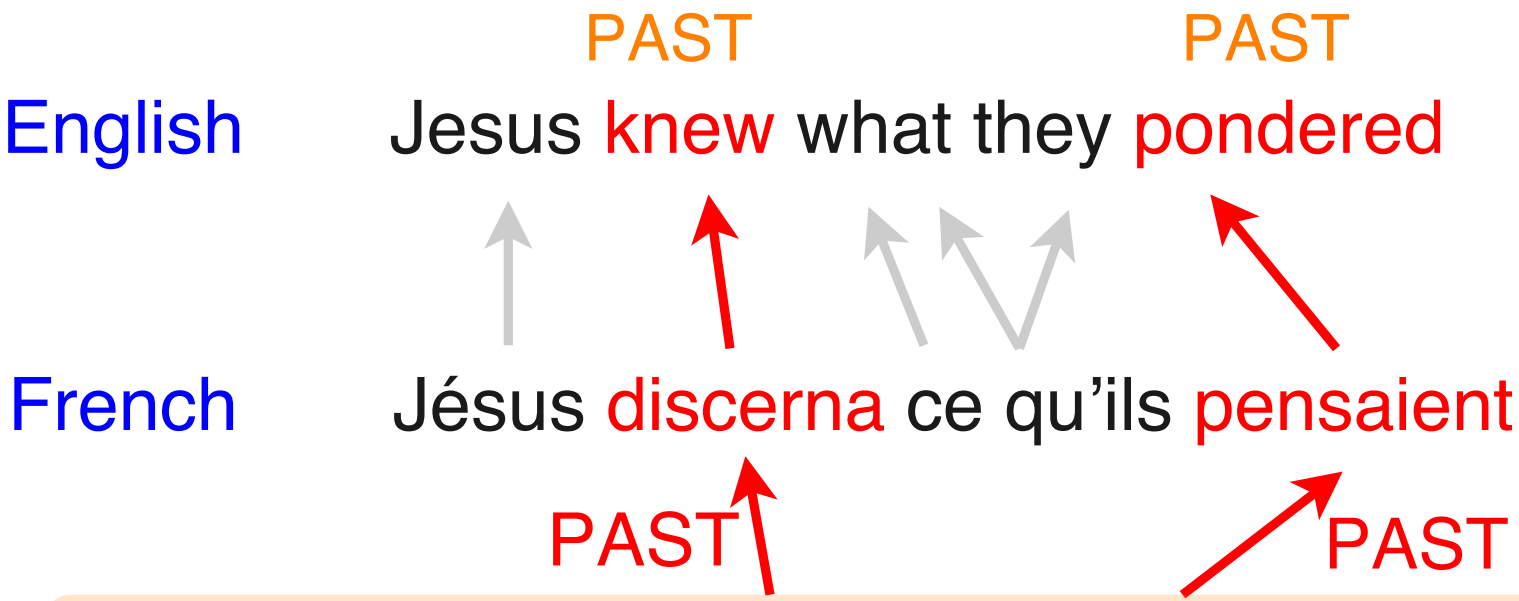
Important area of NLP research: Yarowsky et al. (2001); Spreyer and Frank (2008); Padó & Lapata (2009); Das and Petrov (2011); Agić et al. (2016).



Linguistic Annotation E.g. Part-of-speech tagging

Annotation Projection: project labels via alignment

Important area of NLP research: Yarowsky et al. (2001); Spreyer and Frank (2008); Padó & Lapata (2009); Das and Petrov (2011); Agić et al. (2016).



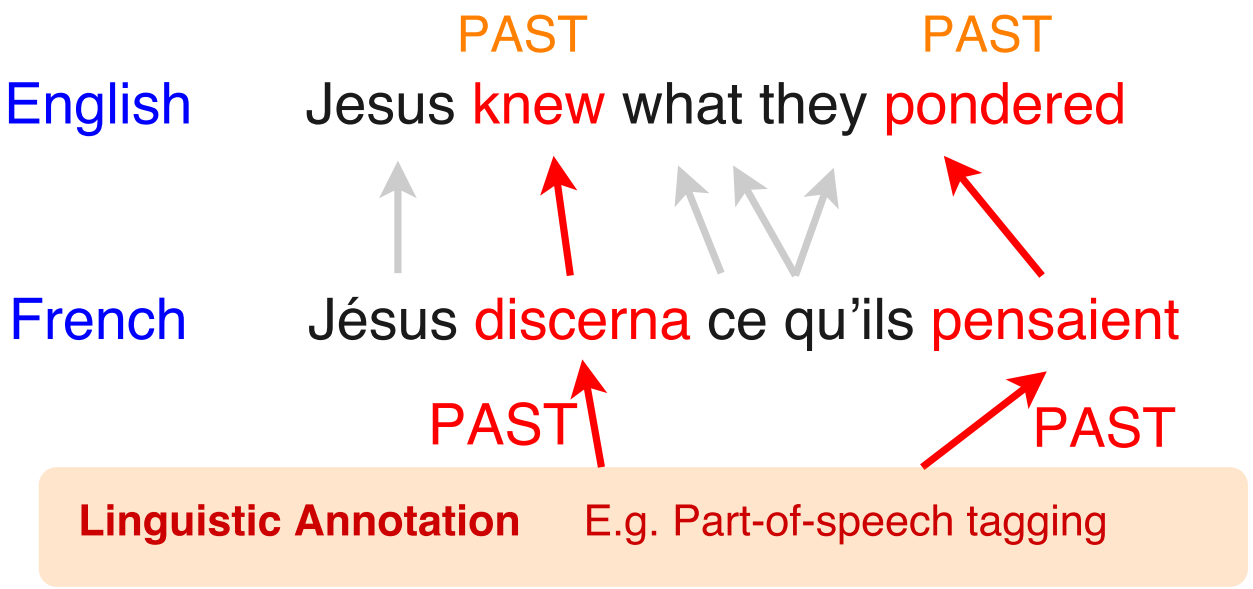
Linguistic Annotation E.g. Part-of-speech tagging

Practical: Introducing an alignment tool

Practical session with FastAlign: Let's align German and English

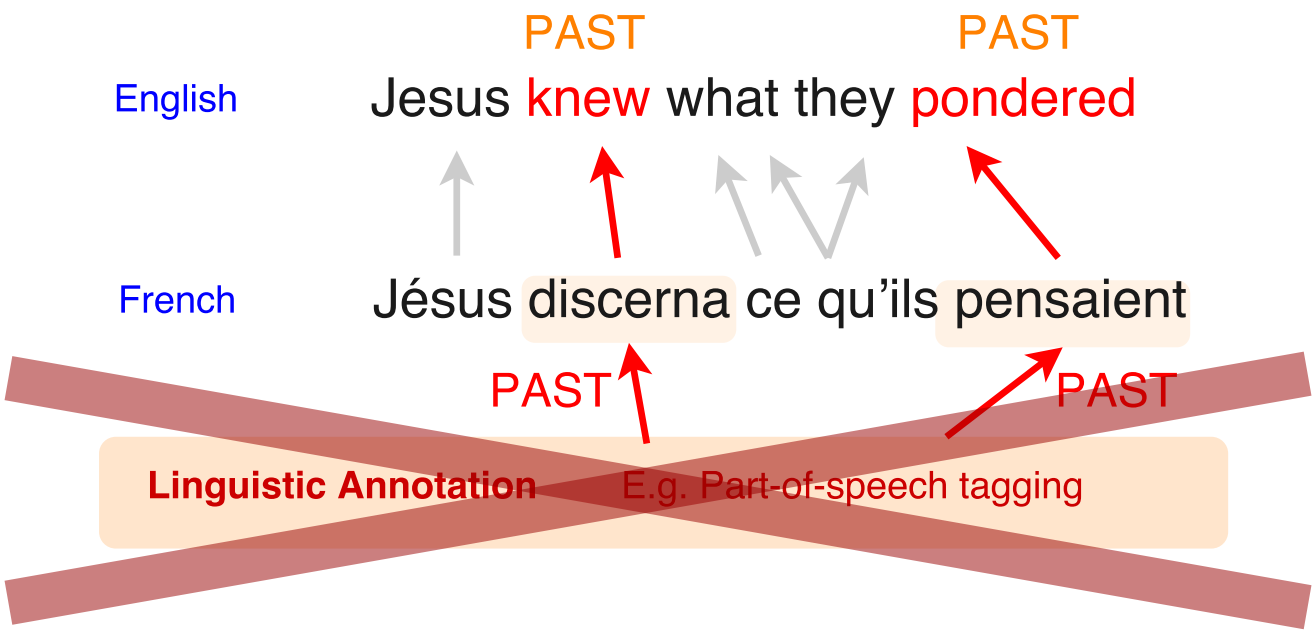
Basic insight of this work

Use words (not human labels) in one language
as annotation labels , for words in another language.



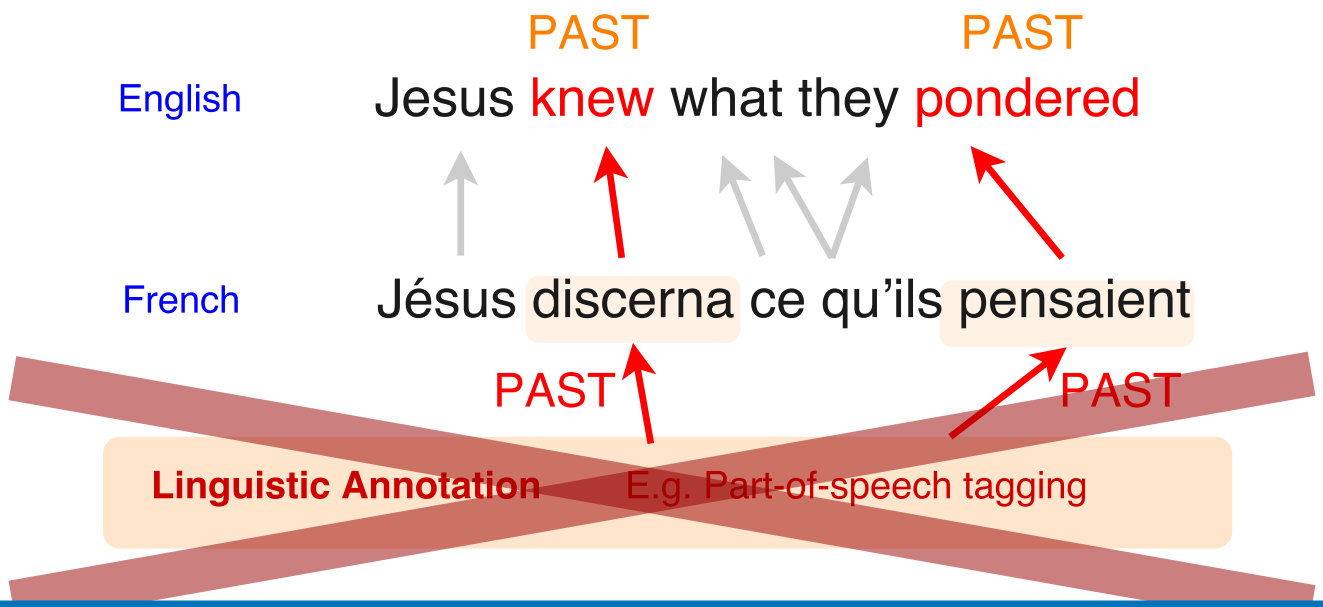
Basic insight of this work

Use words (not human labels) in one language
as annotation labels , for words in another language.



Basic insight of this work

Use words (not human labels) in one language
as annotation labels , for words in another language.



Example

- We use "ti" in Seychelles Creole (CRS) as an annotation label

Basic insight of this work

Use words (not human labels) in one language
as annotation labels , for words in another language.

English: Jesus knew what they pondered

Seychelles Creole: Me Zezi ti konn tou sa ki ti pe pase dan zot lespri

No annotation!

Example

- We use "ti" in Seychelles Creole (CRS) as an annotation label

Basic insight of this work

Use words (not human labels) in one language
as annotation labels , for words in another language.

	PAST	PAST
English	Jesus knew	what they pondered
Seychelles Creole	Me Zezi ti konn	tou sa ki ti pe pase dan zot lespri

No annotation!

Example

- We use "ti" in Seychelles Creole (CRS) as an annotation label

Basic insight of this work

Use words (not human labels) in one language
as annotation labels , for words in another language.

	PAST	PAST
English	Jesus knew	what they pondered
Seychelles Creole	Me Zezi ti konn	tou sa ki ti pe pase dan zot lespri

No annotation!

Example

- We use "ti" in Seychelles Creole (CRS) as an annotation label
- "ti" in CRS is a past tense marker.

Basic insight of this work

Use words (not human labels) in one language
as annotation labels , for words in another language.

	PAST	PAST
English	Jesus knew	what they pondered
Seychelles Creole	Me Zezi ti konn	tou sa ki ti pe pase dan zot lespri

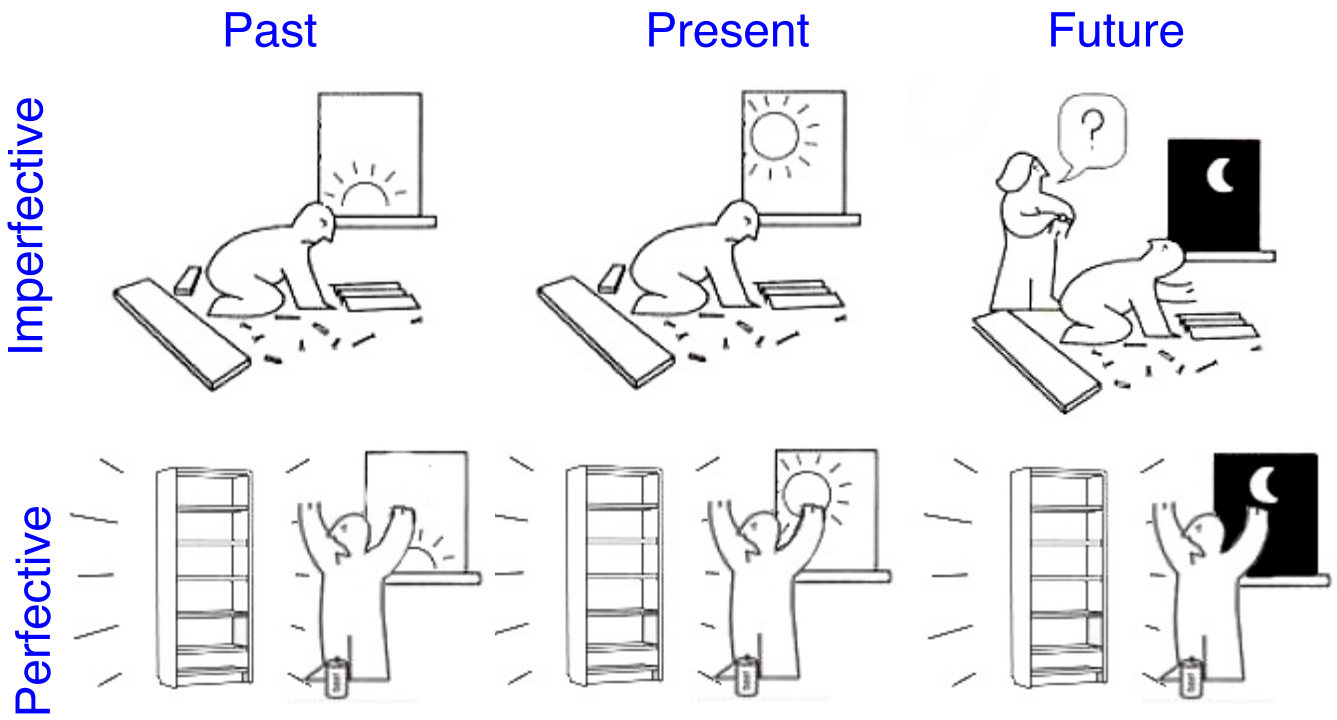
No annotation!

Example

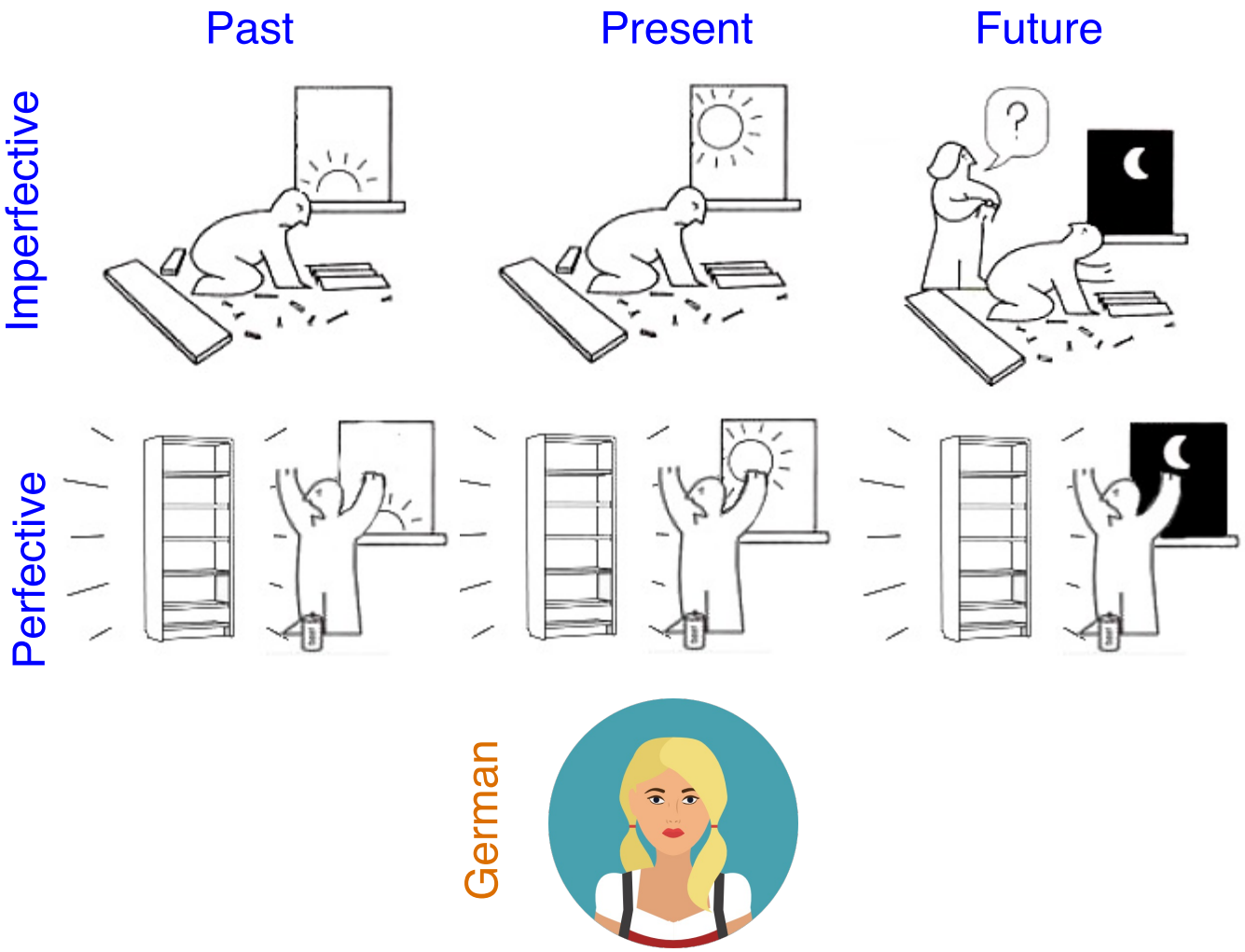
- We use "ti" in Seychelles Creole (CRS) as an annotation label
- "ti" in CRS is a past tense marker.
- "ti" has 100% recall & 100% precision for past tense .

Tense/aspect example: Assembling Ikea shelves

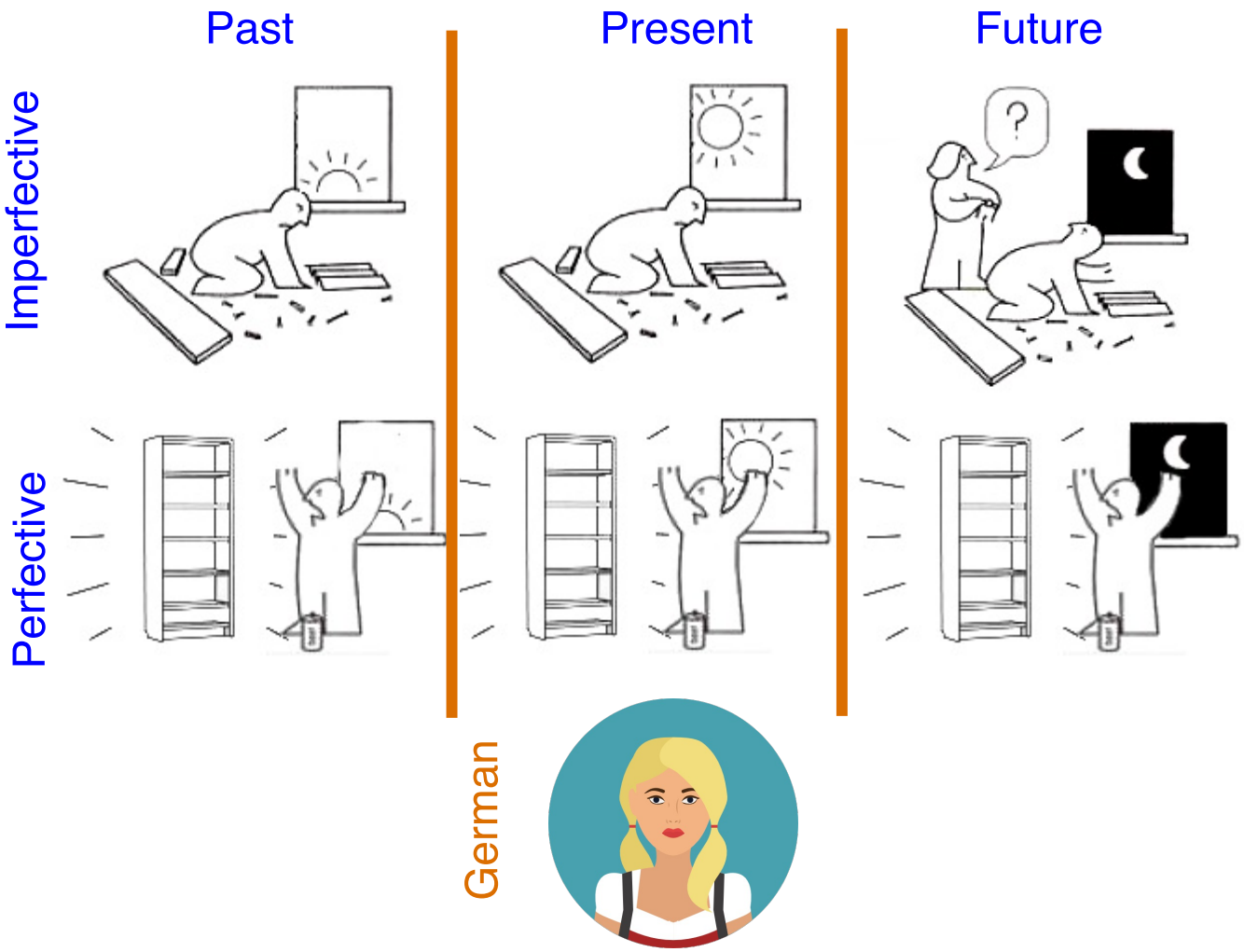
Tense/aspect example: Assembling Ikea shelves



Tense/aspect example: Assembling Ikea shelves

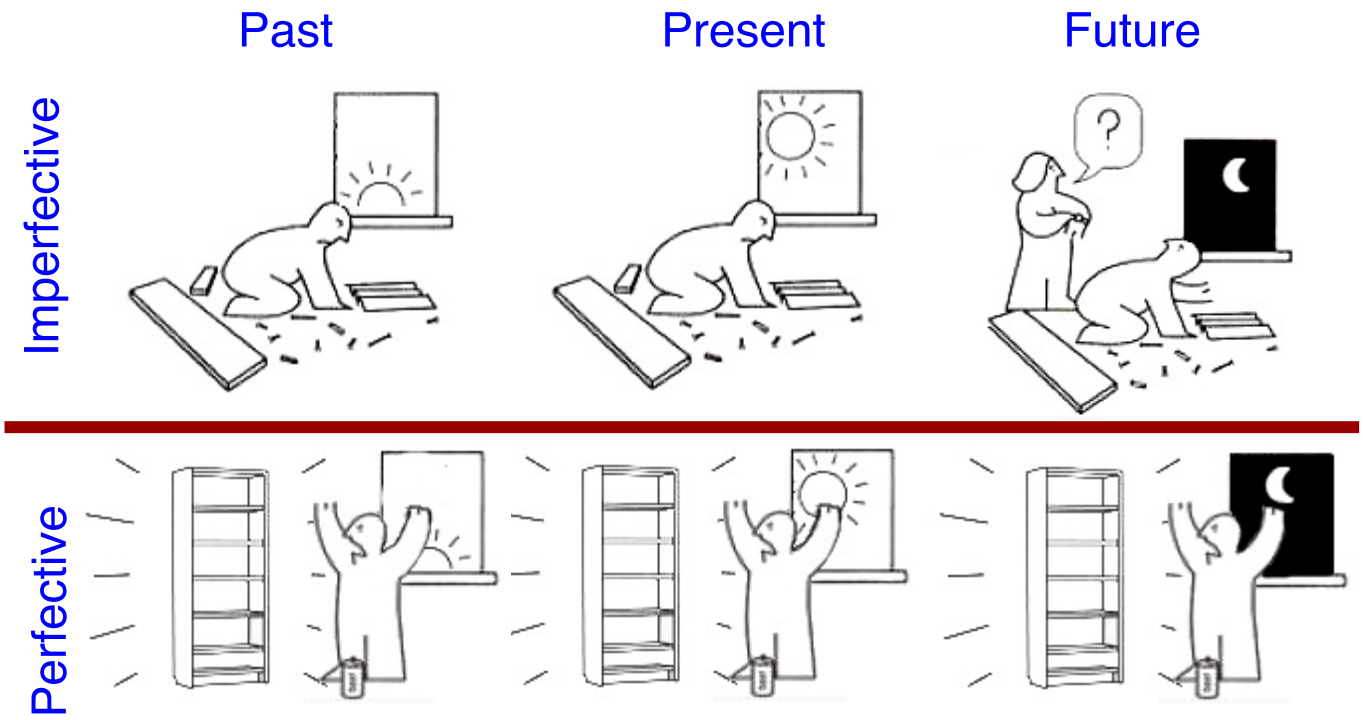


Tense/aspect example: Assembling Ikea shelves

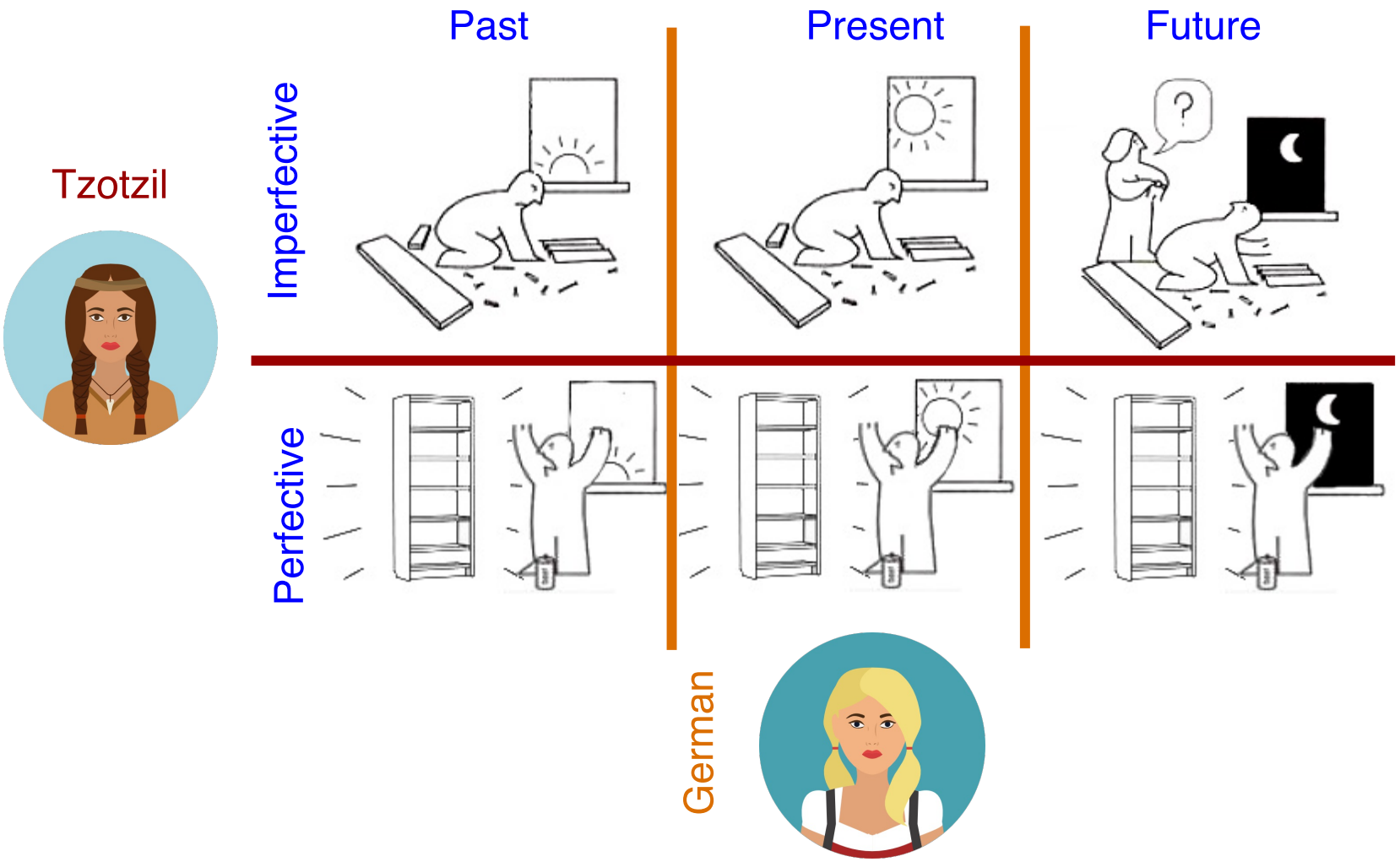


Tense/aspect example: Assembling Ikea shelves

Tzotzil



Tense/aspect example: Assembling Ikea shelves



Summary of our task

Project annotations of tense in 1000 languages

- Don't use linguistic annotations
- but use words as annotations

Summary of our task

Project annotations of tense in 1000 languages

- Don't use linguistic annotations
- but use words as annotations

	PAST	PAST
English	Jesus knew	what they pondered
Seychelles Creole	Me Zezi ti konn	tou sa ki ti pe pase dan zot lespri

No annotation!

Summary of our task

Project annotations of tense in 1000 languages

- Don't use linguistic annotations
- but use words as annotations

	PAST	PAST
English	Jesus knew	what they pondered
Seychelles Creole	Me Zezi ti konn	tou sa ki ti pe pase dan zot lespri

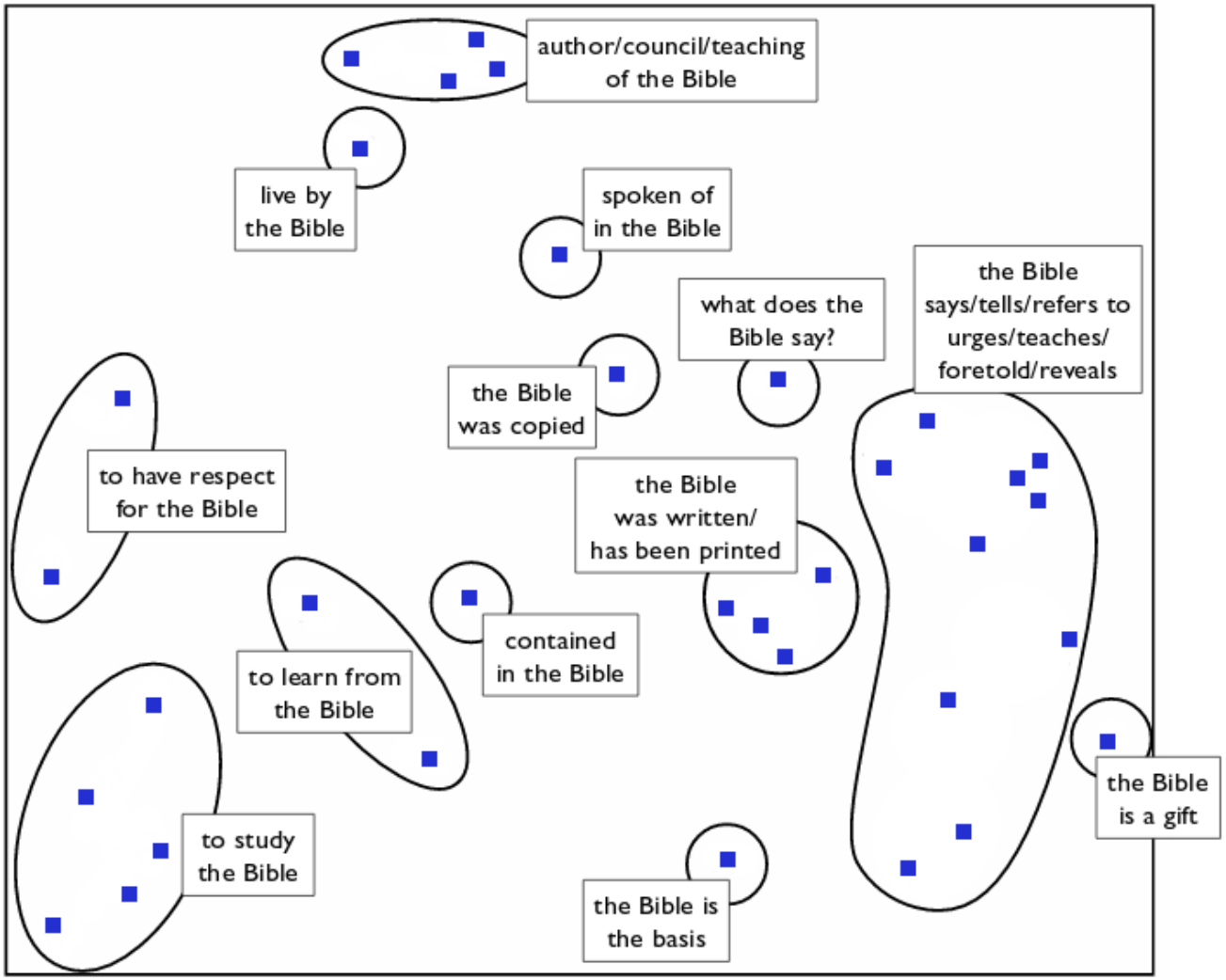
No annotation!

Example

- In particular: past, present, and future tenses

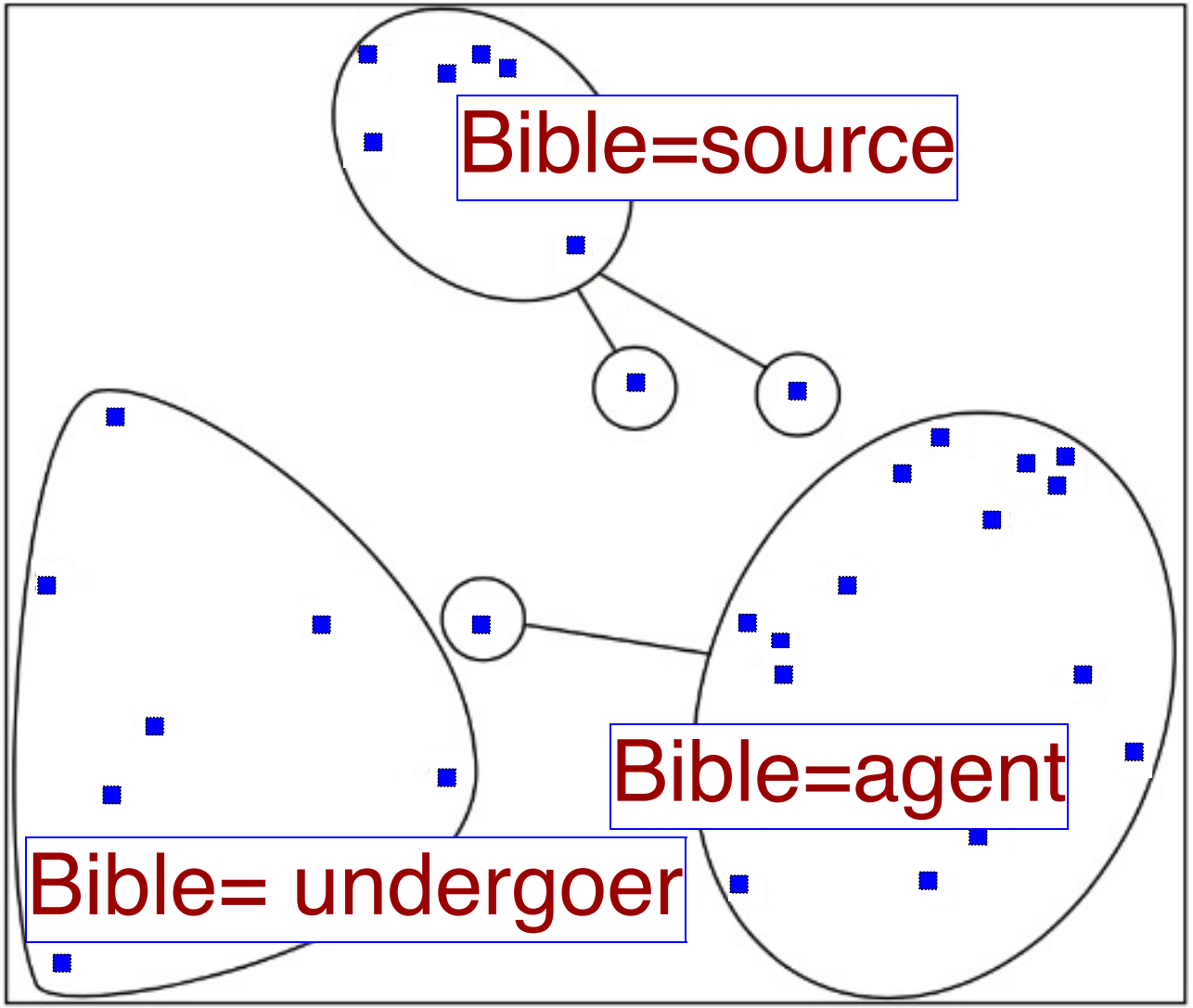
Linguistic Inspiration: Cysouw's typological map

Empirical & universal map of semantic roles based on parallel text. Each point = one sentence. Sentences with similar semantic roles are close.



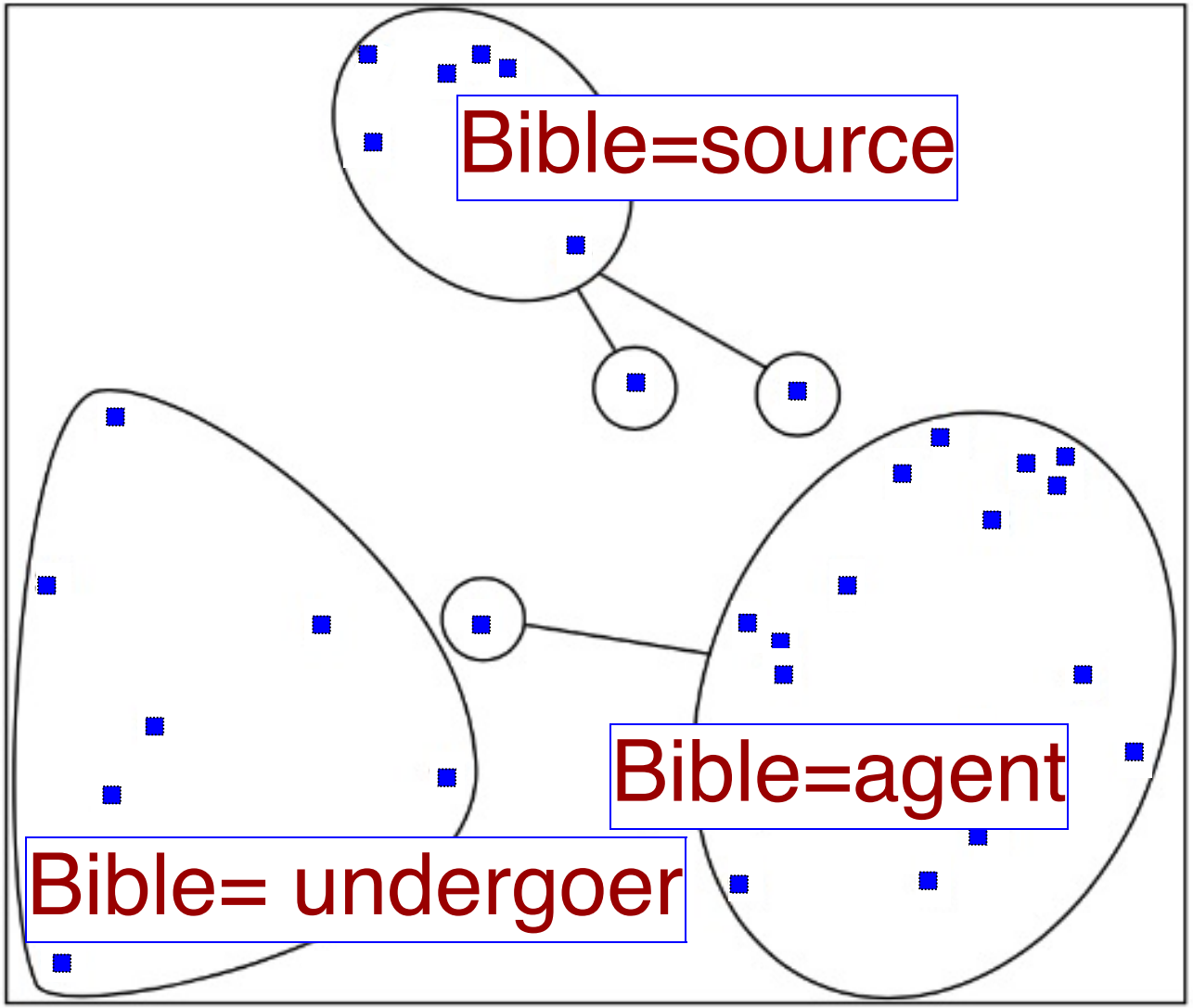
Linguistic Inspiration: Cysouw's typological map

Empirical & universal map of semantic roles based on parallel text. Each point = one sentence. Sentences with similar semantic roles are close.



Linguistic Inspiration: Cysouw's typological map

Empirical & universal map of semantic roles based on parallel text. Each point = one sentence. Sentences with similar semantic roles are close.



Our contribution

more features
more languages,
& automatically!

Motivation for NLP

Low-resource languages

- Many of 7000 world's languages are low resource

Motivation for NLP

Low-resource languages

- Many of 7000 world's languages are low resource
- Even machine translation is challenging for low-resource languages

Motivation for NLP

Low-resource languages

- Many of 7000 world's languages are low resource
- Even machine translation is challenging for low-resource languages
- Typological information is useful for low-resource NLP

Motivation for NLP

Low-resource languages

- Many of 7000 world's languages are low resource
- Even machine translation is challenging for low-resource languages
- Typological information is useful for low-resource NLP
- Completion of world atlas of language structures

Motivation for NLP

Low-resource languages

- Many of 7000 world's languages are low resource
- Even machine translation is challenging for low-resource languages
- Typological information is useful for low-resource NLP
- Completion of world atlas of language structures

How can we create typological resources?

Motivation for NLP

Low-resource languages

- Many of 7000 world's languages are low resource
- Even machine translation is challenging for low-resource languages
- Typological information is useful for low-resource NLP
- Completion of world atlas of language structures

How can we create typological resources?

- **Collaborative efforts on manual annotation creation**

Motivation for NLP

Low-resource languages

- Many of 7000 world's languages are low resource
- Even machine translation is challenging for low-resource languages
- Typological information is useful for low-resource NLP
- Completion of world atlas of language structures

How can we create typological resources?

- **Collaborative efforts on manual annotation creation**
  - e.g. UniMorph or World Atlas of Language Structures

Motivation for NLP

Low-resource languages

- Many of 7000 world's languages are low resource
- Even machine translation is challenging for low-resource languages
- Typological information is useful for low-resource NLP
- Completion of world atlas of language structures

How can we create typological resources?

- **Collaborative efforts on manual annotation creation**
  - e.g. UniMorph or World Atlas of Language Structures
- **Annotation projection**
  - That's what we do, but without human annotation!

Outline

1. Problem definition
2. Linguistic Inspiration
3. Methods and Results
   - Data set
   - SuperPivot pipeline
     - Head pivot set
     - Subword-level marker finding
     - Linguistic Analysis
   - Language clustering based on tense marking
4. Contributions

Super-parallel corpora

- 1556 translations of the Bible / 1169 unique languages
- ≈7000 aligned verses per language

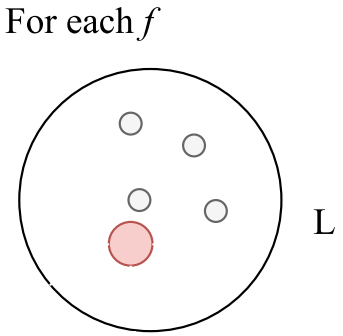
Outline

1. Problem definition
2. Linguistic Inspiration
3. Methods and Results
   - Data set
   - SuperPivot pipeline
     - Head pivot set
     - Subword-level marker finding
     - Linguistic Analysis
   - Language clustering based on tense marking
4. Contributions

SuperPivot hypotheses

H1 Overt encoding exists.

$\forall$ linguistic distinction $f \rightarrow \exists$ a few languages that encode f overtly



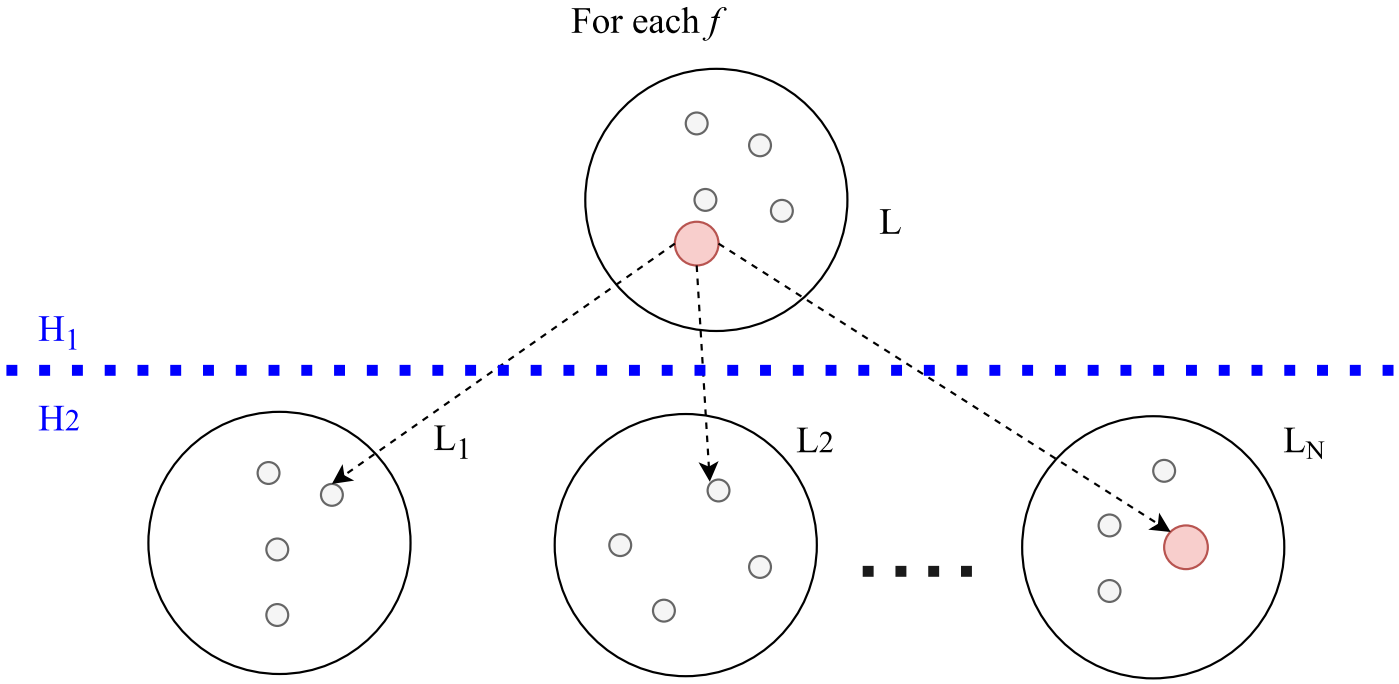
SuperPivot hypotheses

H1 Overt encoding exists.

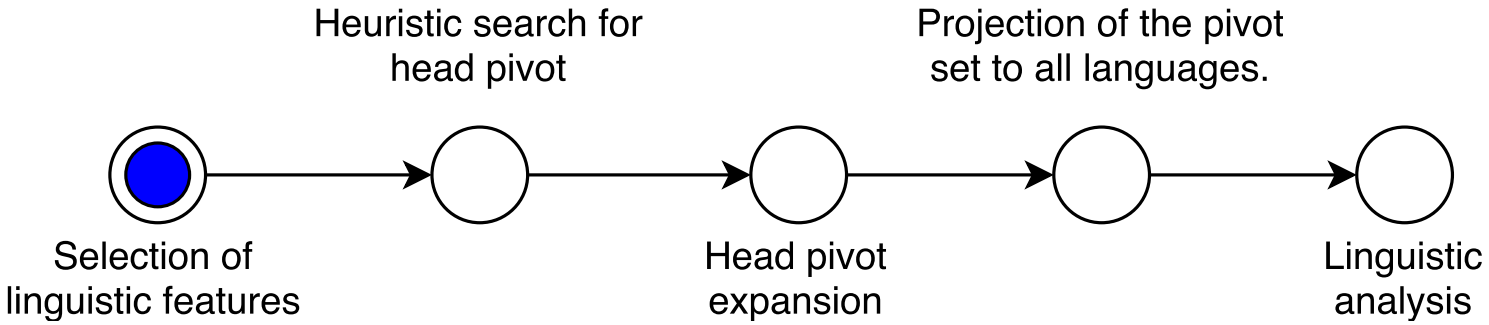
$\forall$ linguistic distinction $f \rightarrow \exists$ few languages that encode f overtly

H2 Overt encoding can be projected.

Projection of f to a language $l' \rightarrow$ either overt or non-overt markers in l'.



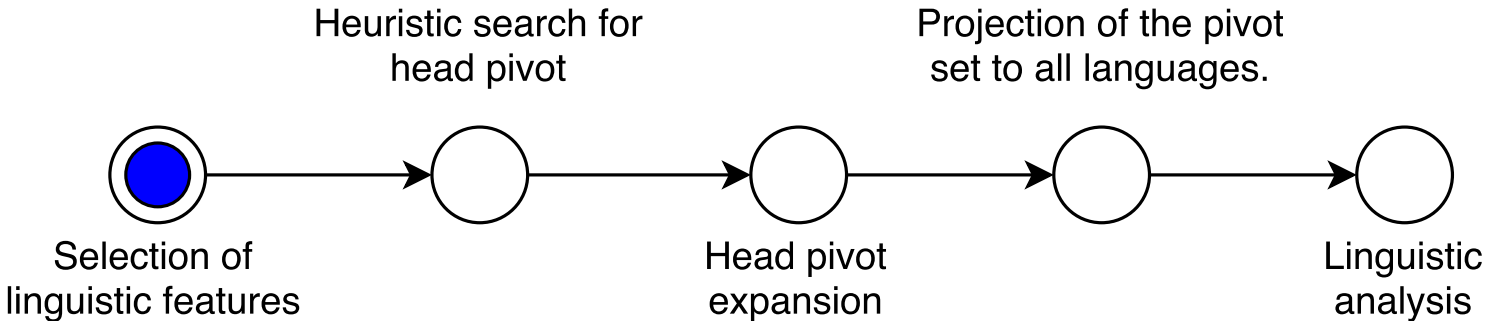
SuperPivot pipeline step 1



(1) Selection of linguistic features

We select a linguistic feature f of interest.

SuperPivot pipeline step 1



(1) Selection of linguistic features

We select a linguistic feature f of interest.

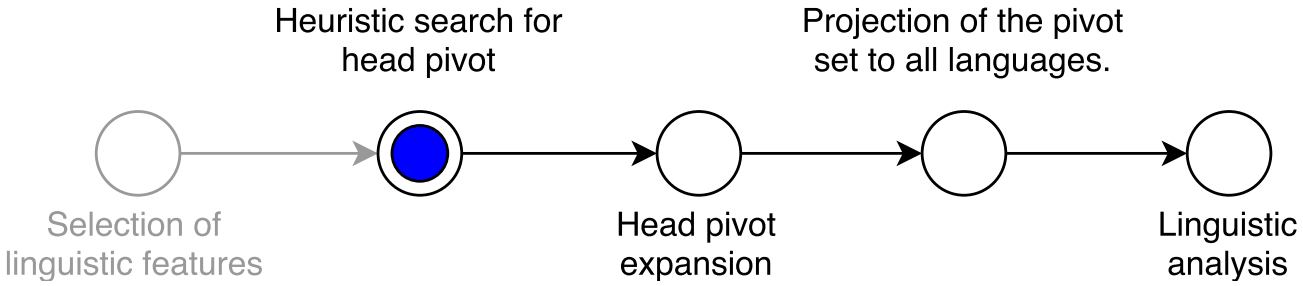
Example

We select past tense as feature f .

SuperPivot pipeline step 2 - Searching for a head pivot

(2) Heuristic search for head pivot

Find language l and its a head pivot p_h highly correlated with the linguistic feature



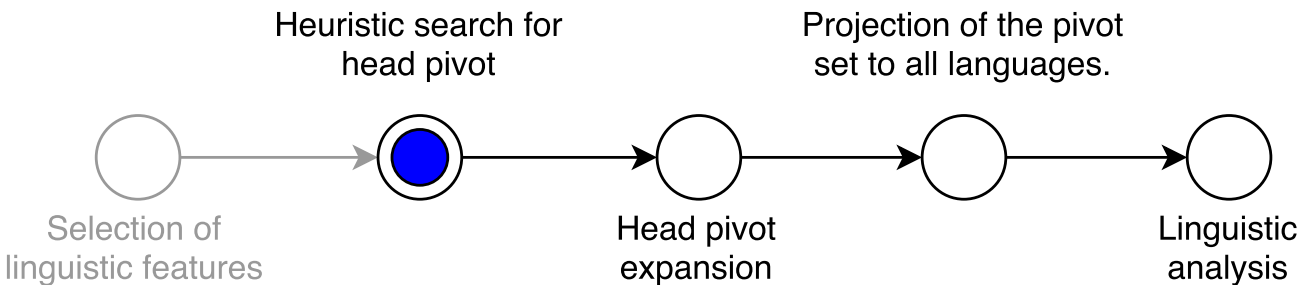
SuperPivot pipeline step 2 - Searching for a head pivot

(2) Heuristic search for head pivot

Find language l and its a head pivot p_h highly correlated with the linguistic feature

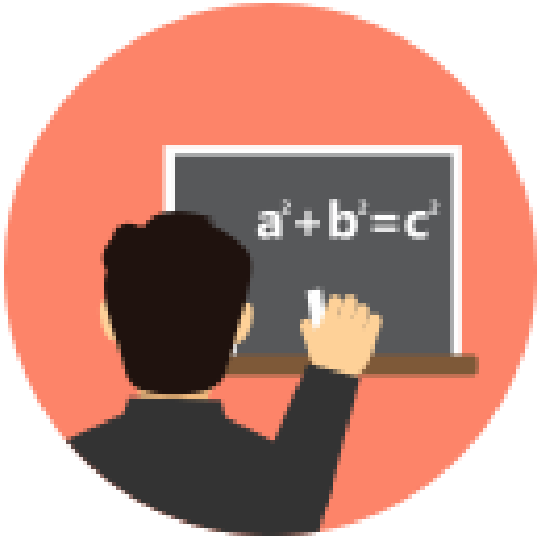
Example

- "ti" in Seychelles Creole (CRS) is an ideal head pivot for past-tense
- It is a surface marker, easily identifiable and not ambiguous



Practical: χ^2 test for feature selection

Practical introduction on using χ^2 test on the alignment edges.

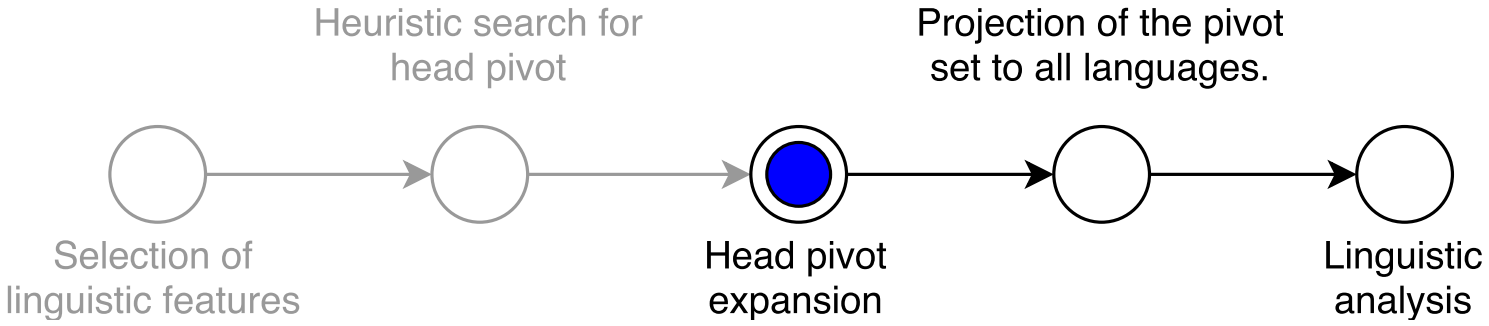


SuperPivot pipeline step 3 - Head pivot expansion

(3) Projection of head pivot to larger pivot set

Expansion of the pivot set of highly correlated surface markers

- More pivots makes the method more robust
- Moving to a larger pivot set increases recall
- Creation of a fine-grained map



Head pivot set extracted at step 3

Top tense pivots extracted using SuperPivot from 1163 languages.

Description	code	language	pivot
head pivots	**CRS**	**Seychelles C.**	*ti*
Top pivot set	GUX	Gourmanchéma	*den*
	MAW	Mampruli	*daa*
	GFK	Patpatar	*ga*
	YAL	Yalunka	*yi*
	TOH	Gitonga	*di*
	DGI	Northern Dagara	*tɩ*
	BUM	Bulu (Cameroon)	*nga*
	TCS	Torres Strait C.	*bin*
	NDZ	Ndogo	*giì*

Head pivot set extracted at step 3

Top tense pivots extracted using SuperPivot from 1163 languages.

Description	code	language	pivot
head pivots	**CRS**	**Seychelles C.**	*ti*
Top pivot set	GUX	Gourmanchéma	*den*
	MAW	Mampruli	*daa*
	GFK	Patpatar	*ga*
	YAL	Yalunka	*yi*
	TOH	Gitonga	*di*
	DGI	Northern Dagara	*tɩ*
	BUM	Bulu (Cameroon)	*nga*
	TCS	Torres Strait C.	*bin*
	NDZ	Ndogo	*giɩ̀*

Limitations

- We made a big assumption, "all markers are tokenizable"
  '-ed' in English past tense

Head pivot set extracted at step 3

Top tense pivots extracted using SuperPivot from 1163 languages.

Description	code	language	pivot
head pivots	**CRS**	**Seychelles C.**	*ti*
Top pivot set	GUX	Gourmanchéma	*den*
	MAW	Mampruli	*daa*
	GFK	Patpatar	*ga*
	YAL	Yalunka	*yi*
	TOH	Gitonga	*di*
	DGI	Northern Dagara	*tɩ*
	BUM	Bulu (Cameroon)	*nga*
	TCS	Torres Strait C.	*bin*
	NDZ	Ndogo	*giì*

Limitations

- We made a big assumption, "all markers are tokenizable"
  '-ed' in English past tense
- Tokenization imperfection in many languages

SuperPivot pipeline step 4 - Projection to all languages

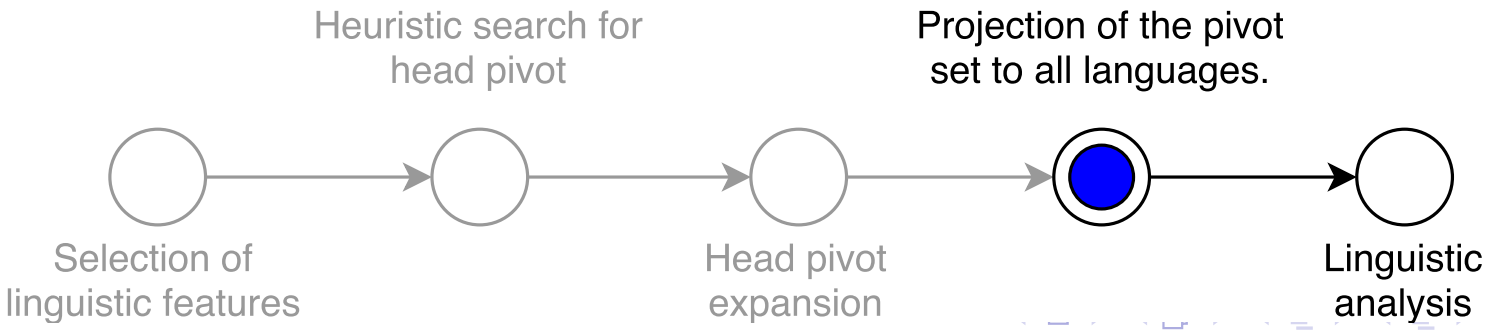
(4) Go to the level of sub-word, i.e. character n-gram

- Search for character n-gram patterns

English: Jesus knew what they pondered

Seychelles Creole: Me Zezi ti konn tou sa ki ti pe pase dan zot lespri

No annotation!



SuperPivot pipeline step 4 - Projection to all languages

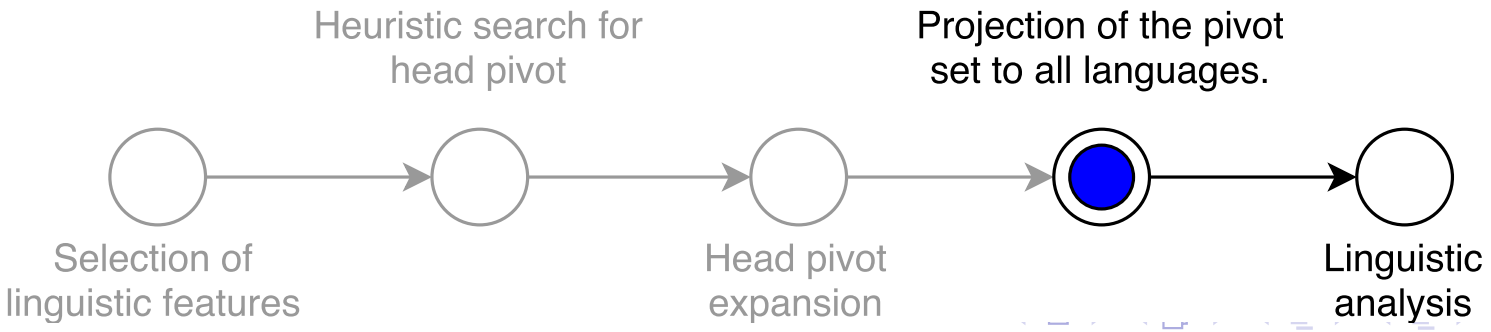
(4) Go to the level of sub-word, i.e. character n-gram

- Search for character n-gram patterns

English: Jesus knew what they pondered

Seychelles Creole: Me Zezi ti konn tou sa ki ti pe pase dan zot lespri

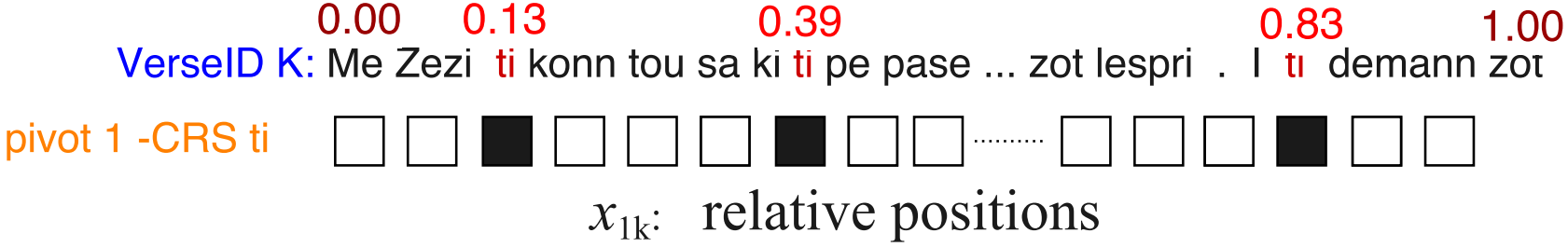
No annotation!



Localize the occurrence of the past tense marker

"I arrived yesterday, I'm staying today, and I will leave tomorrow."

Localize the occurrence of the past tense marker



Localize the occurrence of the past tense marker

VerseID K: Me Zezi ti konn tou sa ki ti pe pase ... zot lespri . I ti demann zot

pivot 1 -CRS ti

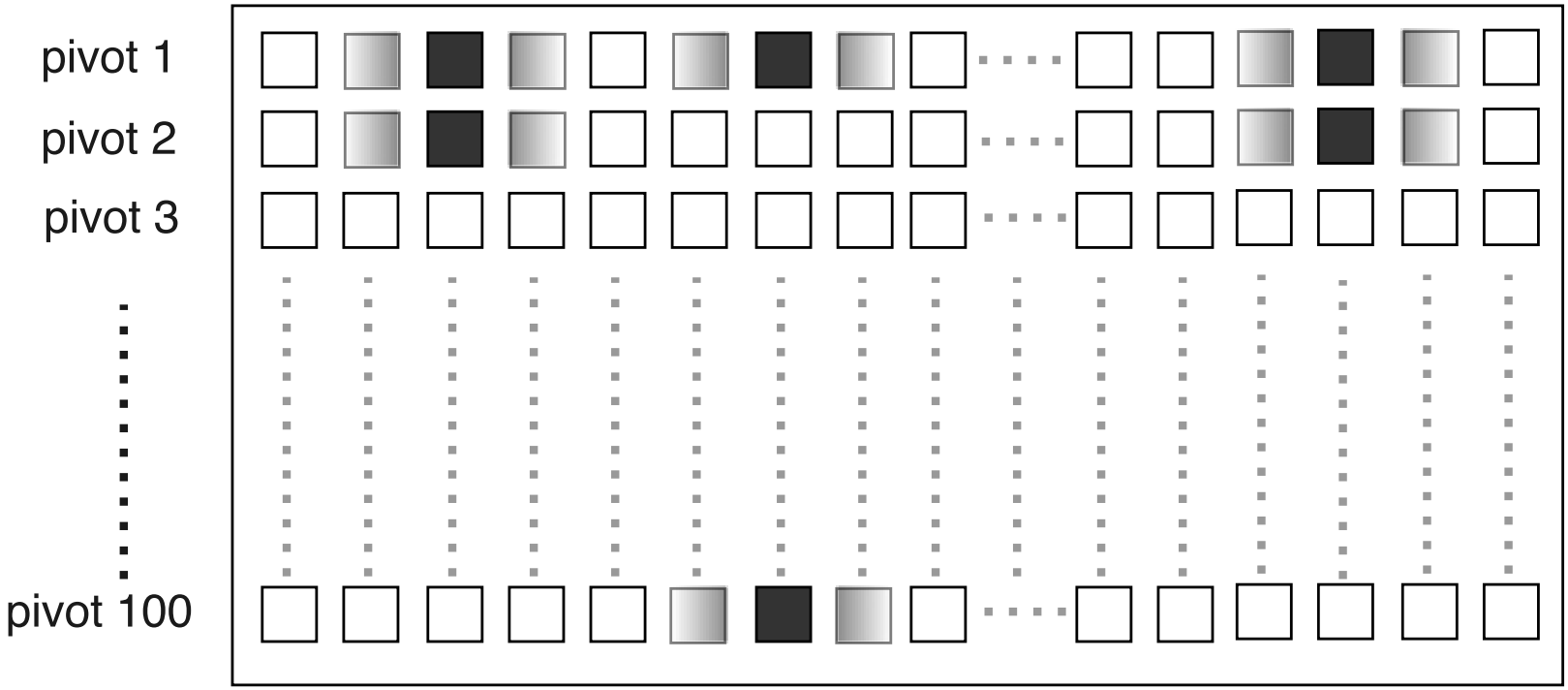
x_{1k}: relative positions

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$$

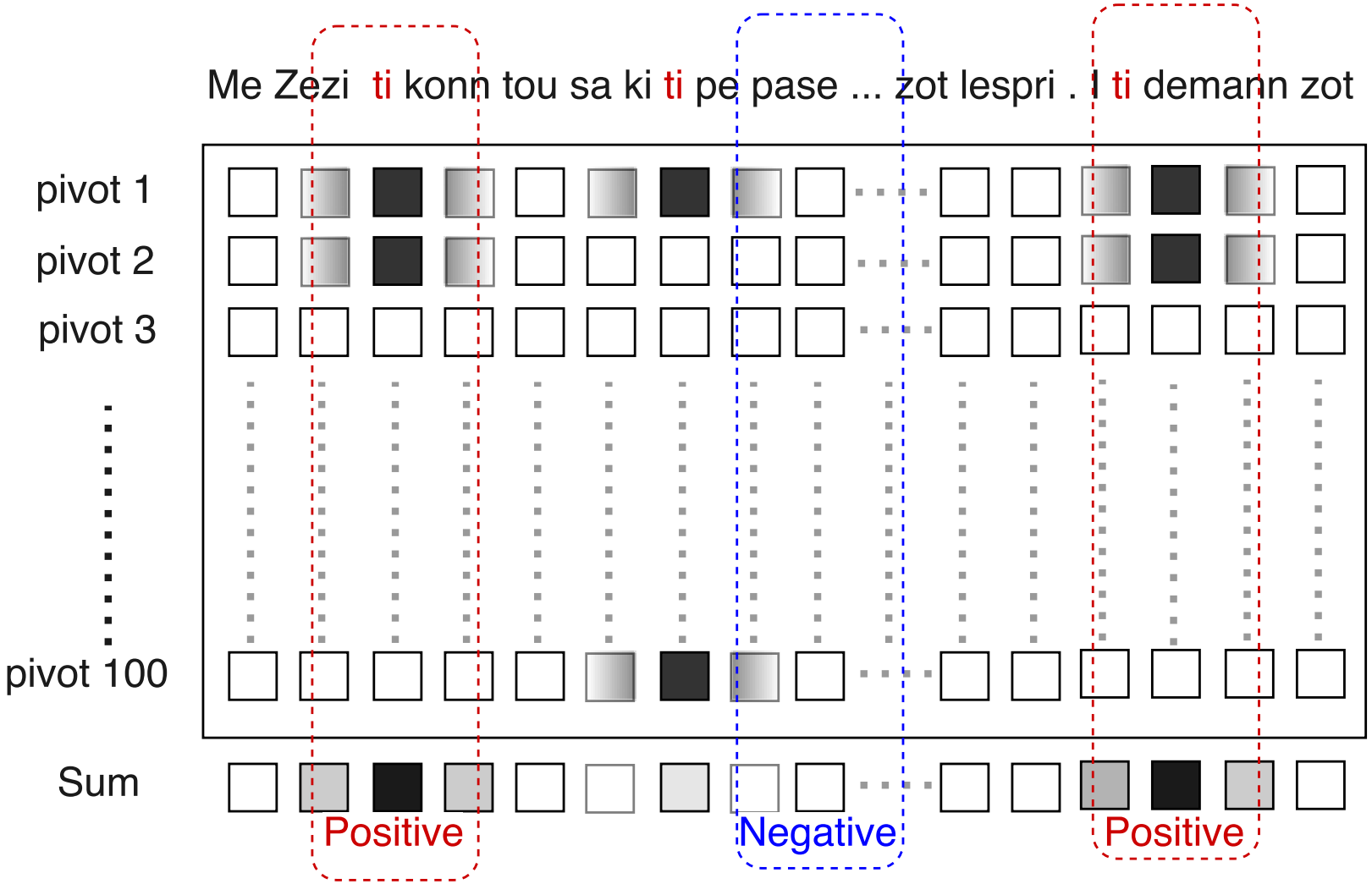
$$\widetilde{x_{1k}} = G(x_{1k}) * x_{1k}$$

Localize the occurrence of the past tense marker

Me Zezi ti konn tou sa ki ti pe pase ... zot lespri . I ti demann zot

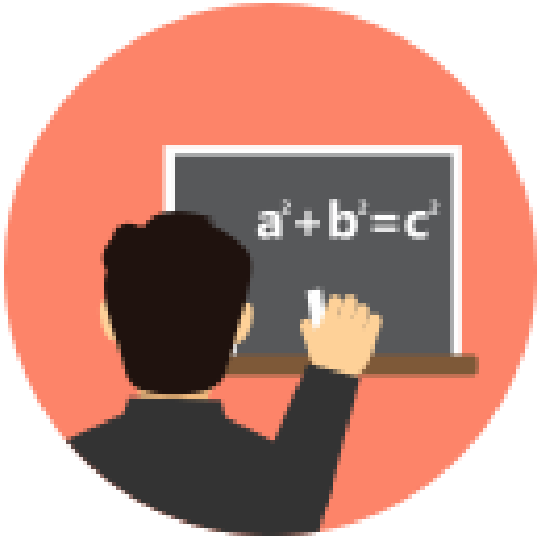


Localize the occurrence of the past tense marker



Practical: subword-level χ^2 for marker finding

Practical on using χ^2 to find subword markers.



Evaluation of the extracted sub-word markers in step 4

Evaluation Framework

We rank n-gram features and calculate the Mean Reciprocal Rank (MRR) for 10 languages has been calculated.

Example

Language	Top 10 n-gram				Frequent Tokens
	Rank	N	N-gram	Score	
English	1	2	<ed>	118	<predicted><stared><seized><appeared>
⋮	⋮	⋮	⋮	⋮	⋮

Mean Reciprocal Rank (MRR) Example

Language	Top 10 n-gram						Reciprocal rank
	Rank	N	N-gram	Score	Dir.	Frequent Tokens	
English	1	2	<ed>	299.9	+	asked, called, answered, wanted, replied	1
	2	2	<ok>	100.0	+	took, spoke, look, looked, looking	
	3	2	< .>	89.5	−	.	
	4	2	<d >	69.4	+	and, God, said, had, Lord	
	5	2	<as>	66.3	+	a, was, as, has, asked	
	6	2	<we>	61.3	+	were, we, went, We, power	
	7	2	<wa>	38.1	+	a, was, way, away, want	
	8	2	< y>	33.7	−	you, your, You, you're, Your	
	9	2	<yo>	33.5	−	you, your, You, everyone, anyone	
	10	2	<te>	31.8	+	Peter, After, tell, temple, guarantee	
English	1	3	<ed >	231.8	+	asked, called, answered, wanted, replied	1
	2	3	<aul>	87.1	+	a, Paul, A, Saul, Paul's	
	3	3	<pau>	85.8	+	a, Paul, A, Paul's, p	
	4	3	<hed>	76.2	+	he, He, touched, finished, washed	
	5	3	<saw>	75.2	+	a, saw, A, sawdust, sawed	
	6	3	<ked>	71.7	+	asked, looked, worked, talked, picked	
	7	3	<was>	71.1	+	a, was, as, A, As	
	8	3	<too>	70.7	+	to, took, too, stood, To	
	9	3	<ook>	67.7	+	took, look, looked, looking, Look	
	10	3	<pet>	64.8	+	Peter, trumpet, Peter's, p, trumpets	
English	1	4	<took>	129.5	+	to, took, too, To, O	1
	2	4	<aul >	95.6	+	a, Paul, A, Saul	
	3	4	<pete>	84.7	+	Peter, Peter's, p, trumpeterswill, competent	
	4	4	<, he>	74.4	+	,, he, He	
	5	4	< was>	73.9	+	a, was, as, A, As	
	6	4	< too>	73.7	+	to, took, too, To, tooth	
	7	4	<ed h>	72.1	+		
	8	4	<paul>	71.8	+	a, Paul, A, Paul's, p	
	9	4	<was >	71.1	+	a, was, as, A, As	
	10	4	<ched>	69.2	+	he, He, touched, watched, reached	

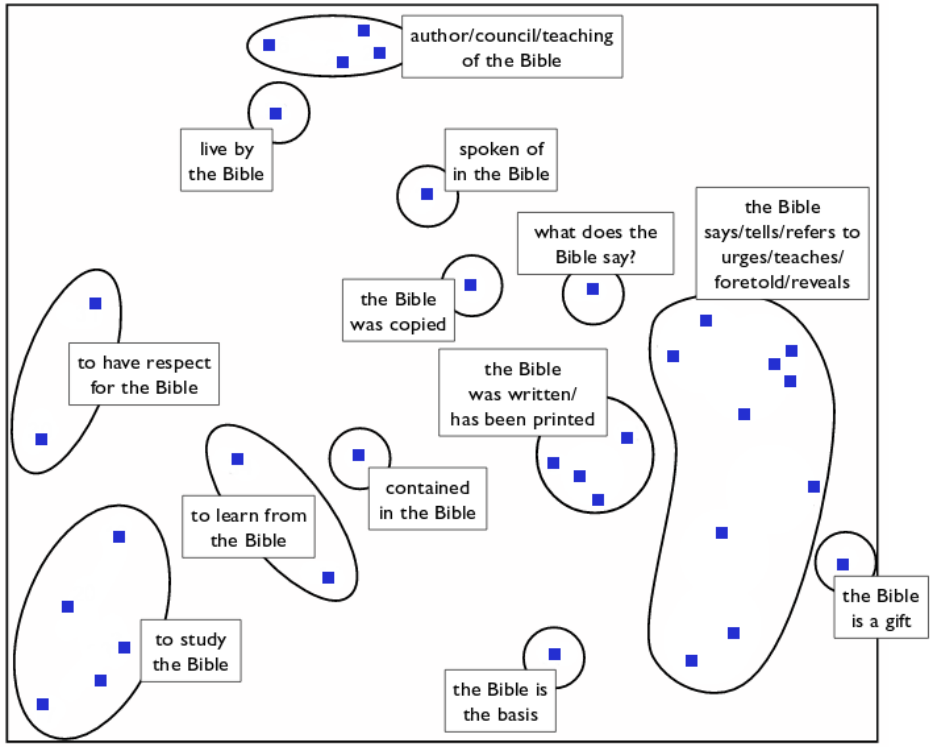
Evaluation of the extracted sub-word markers in step 4

MRR results for step 4:

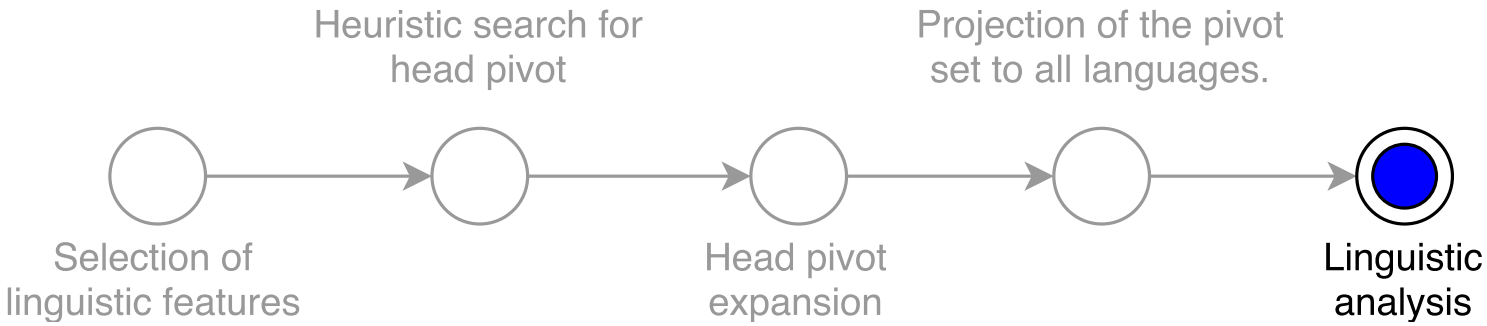
language	past	present	future	all
Arabic	1.00	0.39	0.77	0.72
Chinese	0.00	0.00	0.87	0.29
English	1.00	1.00	1.00	1.00
French	1.00	1.00	1.00	1.00
German	1.00	1.00	1.00	1.00
Italian	1.00	1.00	1.00	1.00
Persian	0.77	1.00	1.00	0.92
Polish	1.00	1.00	0.58	0.86
Russian	0.90	0.50	0.62	0.67
Spanish	1.00	1.00	1.00	1.00
all	0.88	0.79	0.88	0.85

Pipeline - Step 5

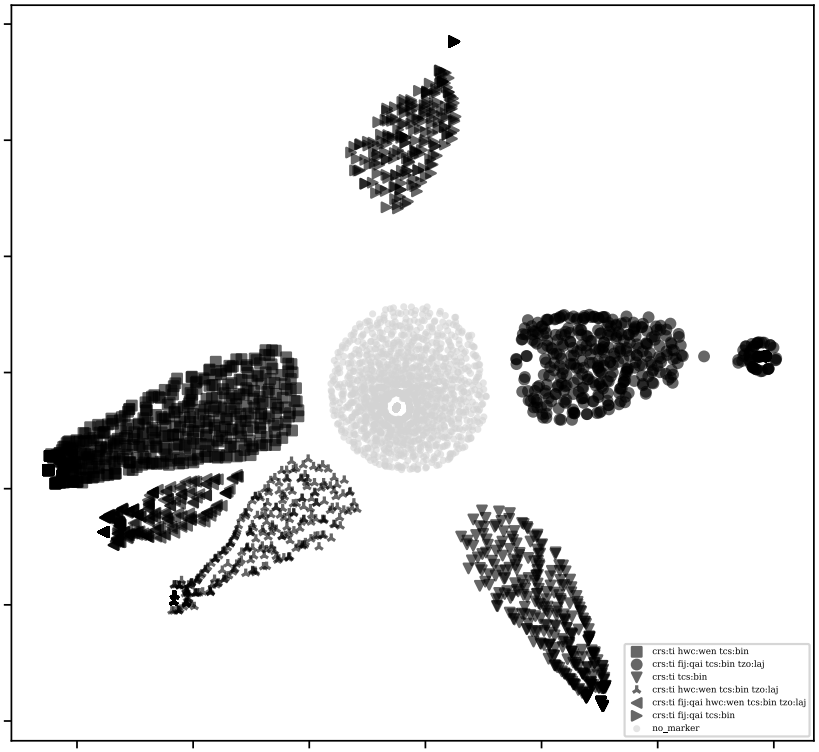
Pivot set can be used to create Cysouw's map for tense:



- A sentence containing the word 'bible'

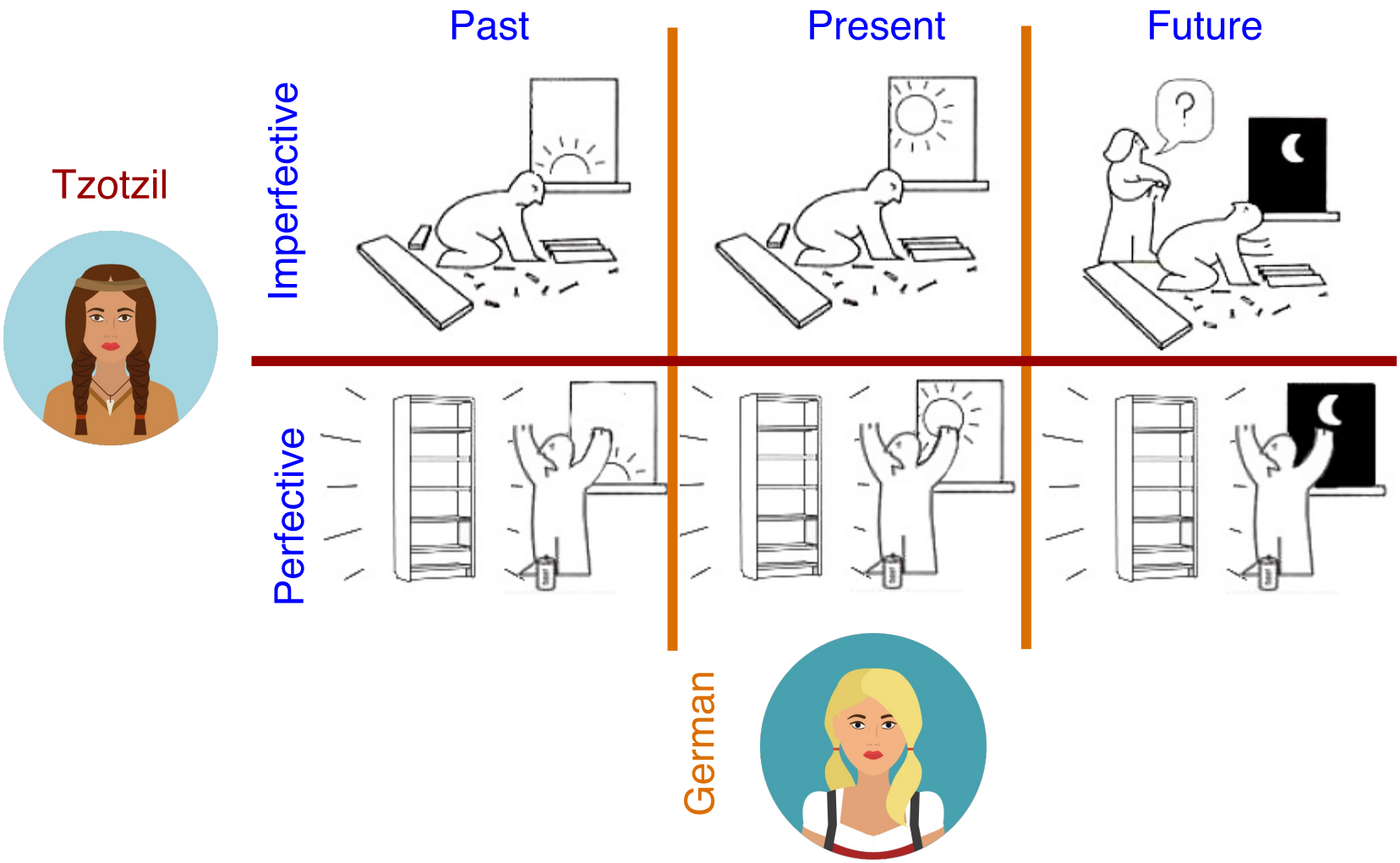


Step 5 - Creation of a tense map

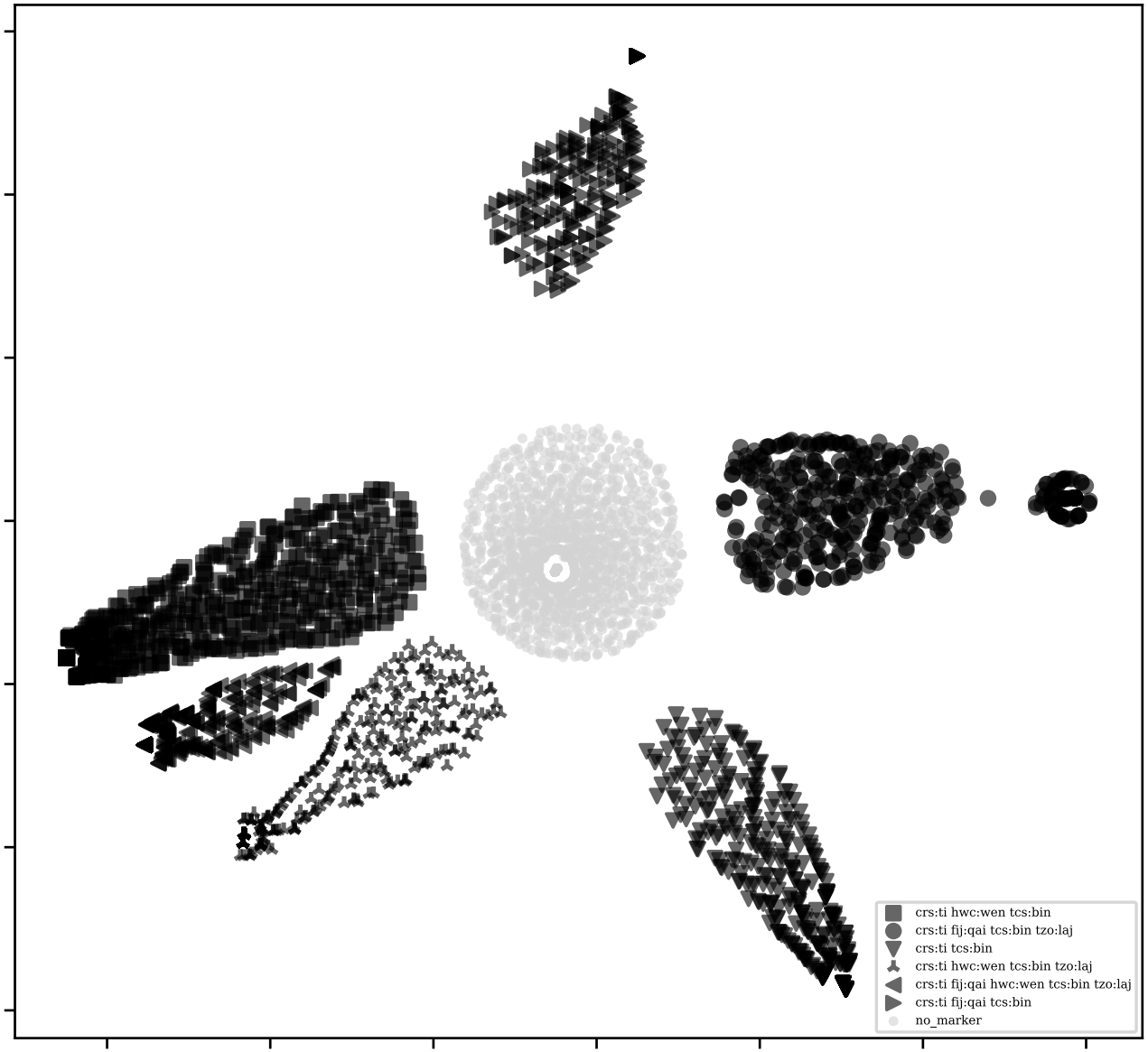


- Empirical & universal map of past tense based on parallel text.
- Each point = one verse.
- t-SNE visualization of verses based on markings (over 5/100 pivots)
- Verses with similar types of past tense are close.

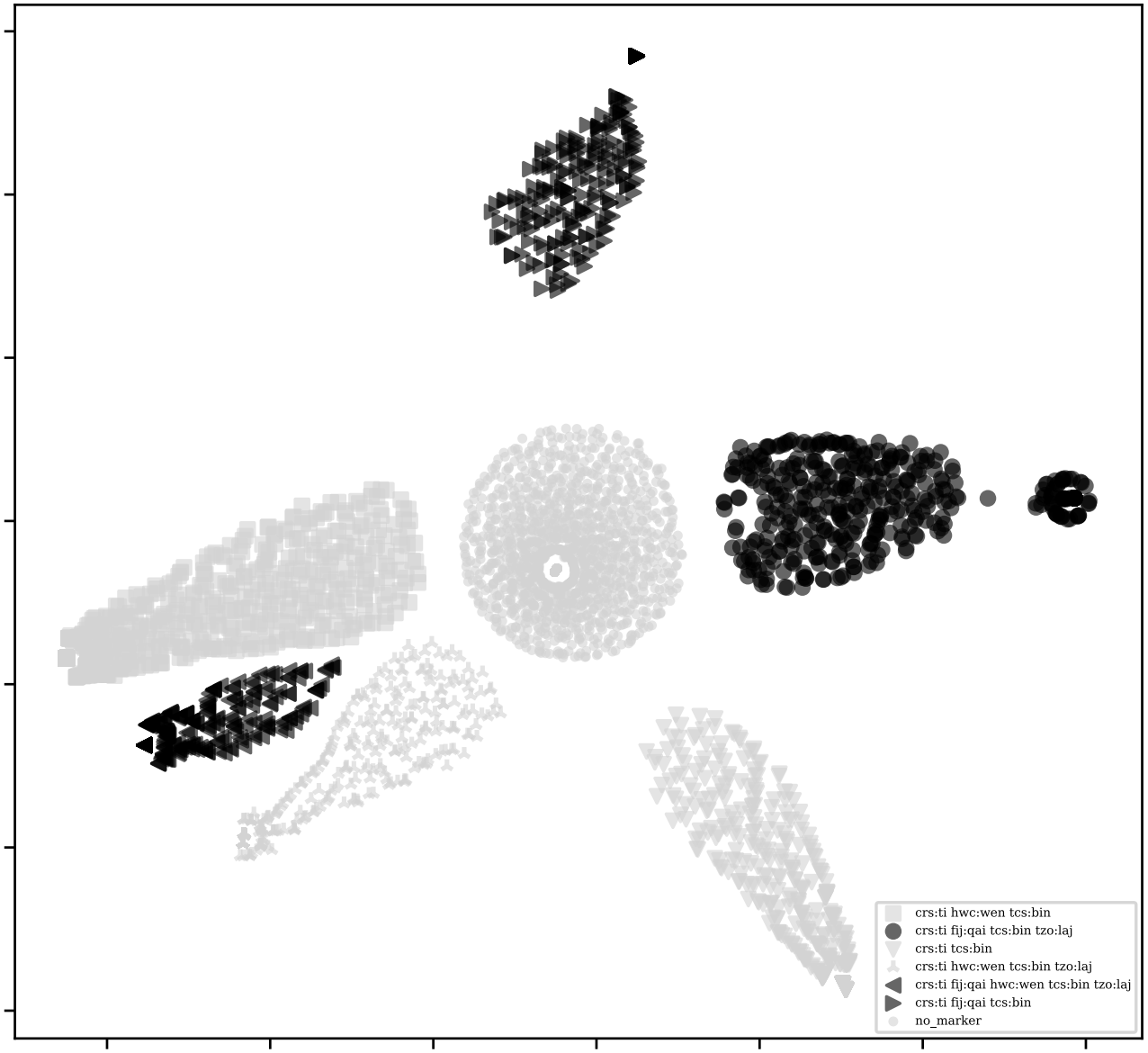
Tense/aspect example: Assembling Ikea shelves



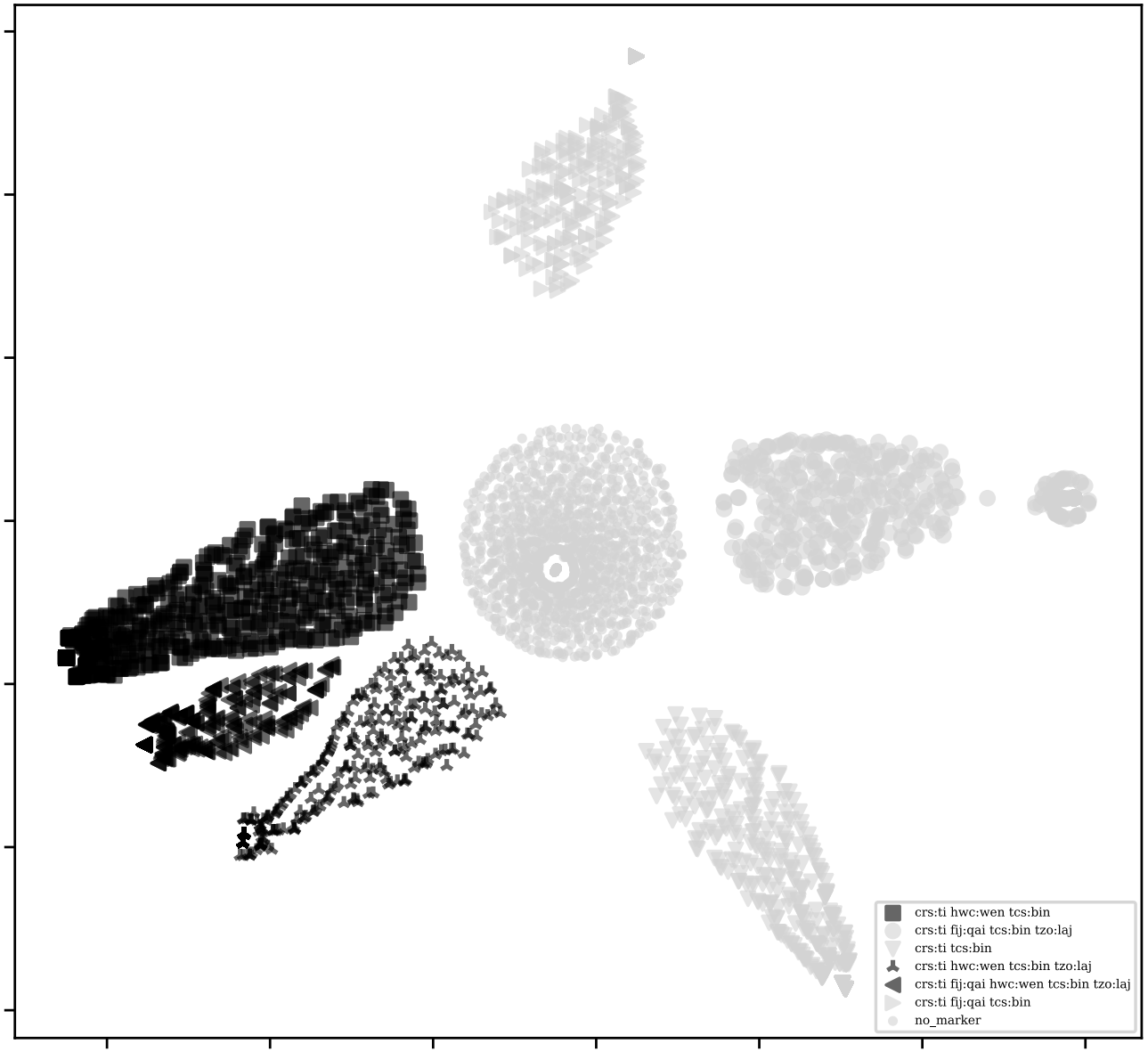
Past tense map: Seychelles Creole 'ti'



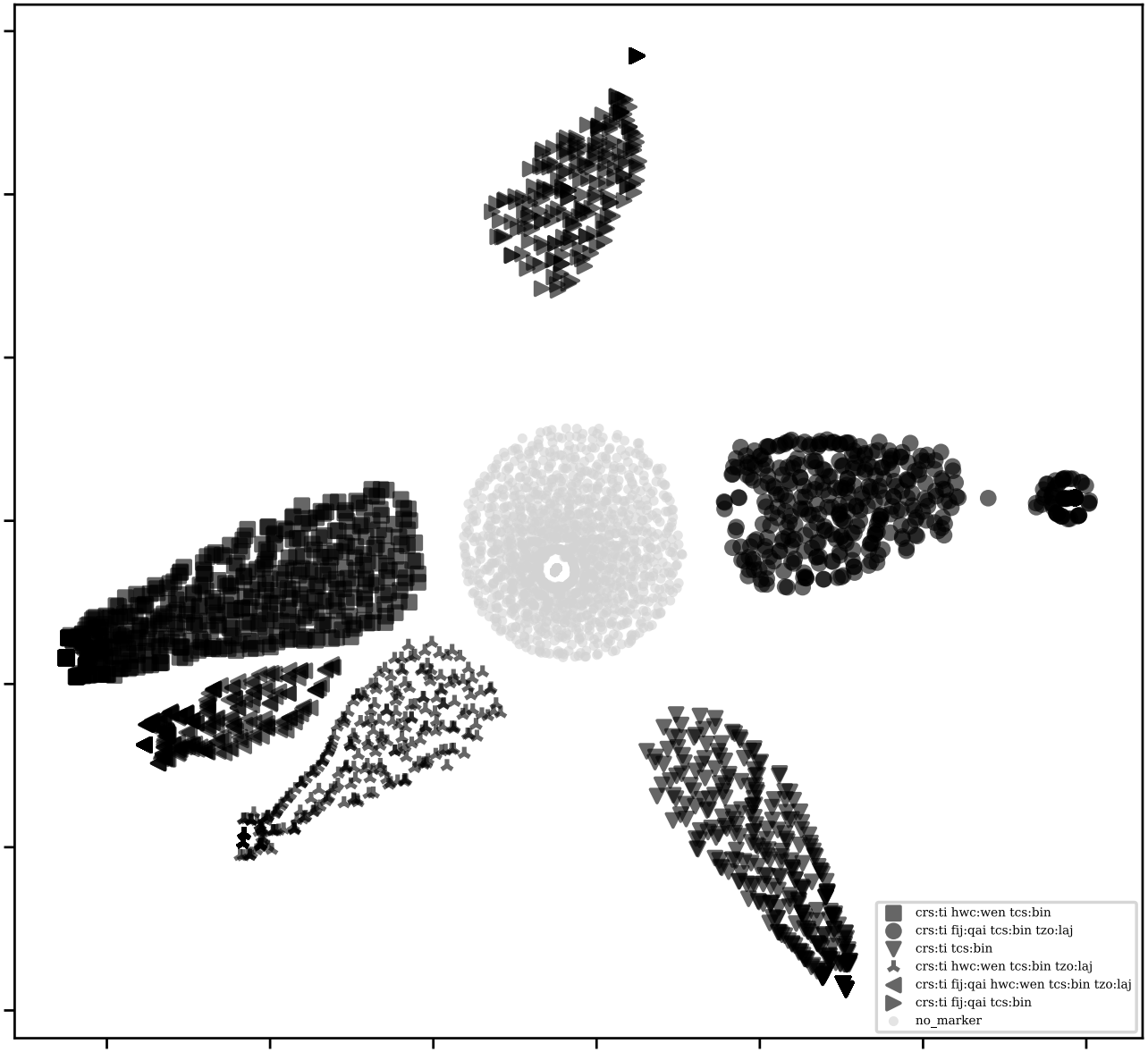
Past tense map: Tzotzil ‘laj’



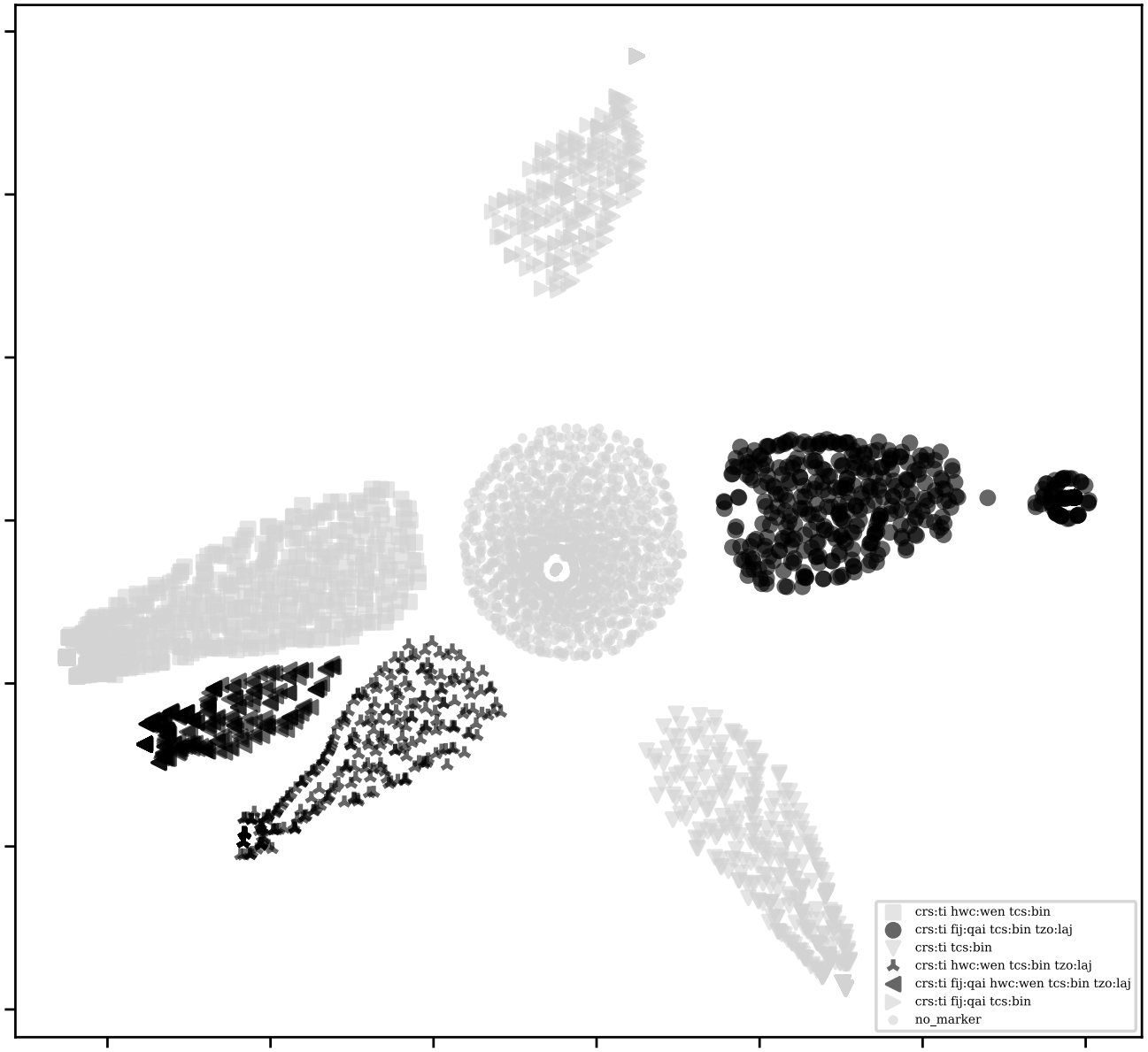
Past tense map: Fijian ‘qai’



Past tense map: Hawaiian Creole ‘wen’



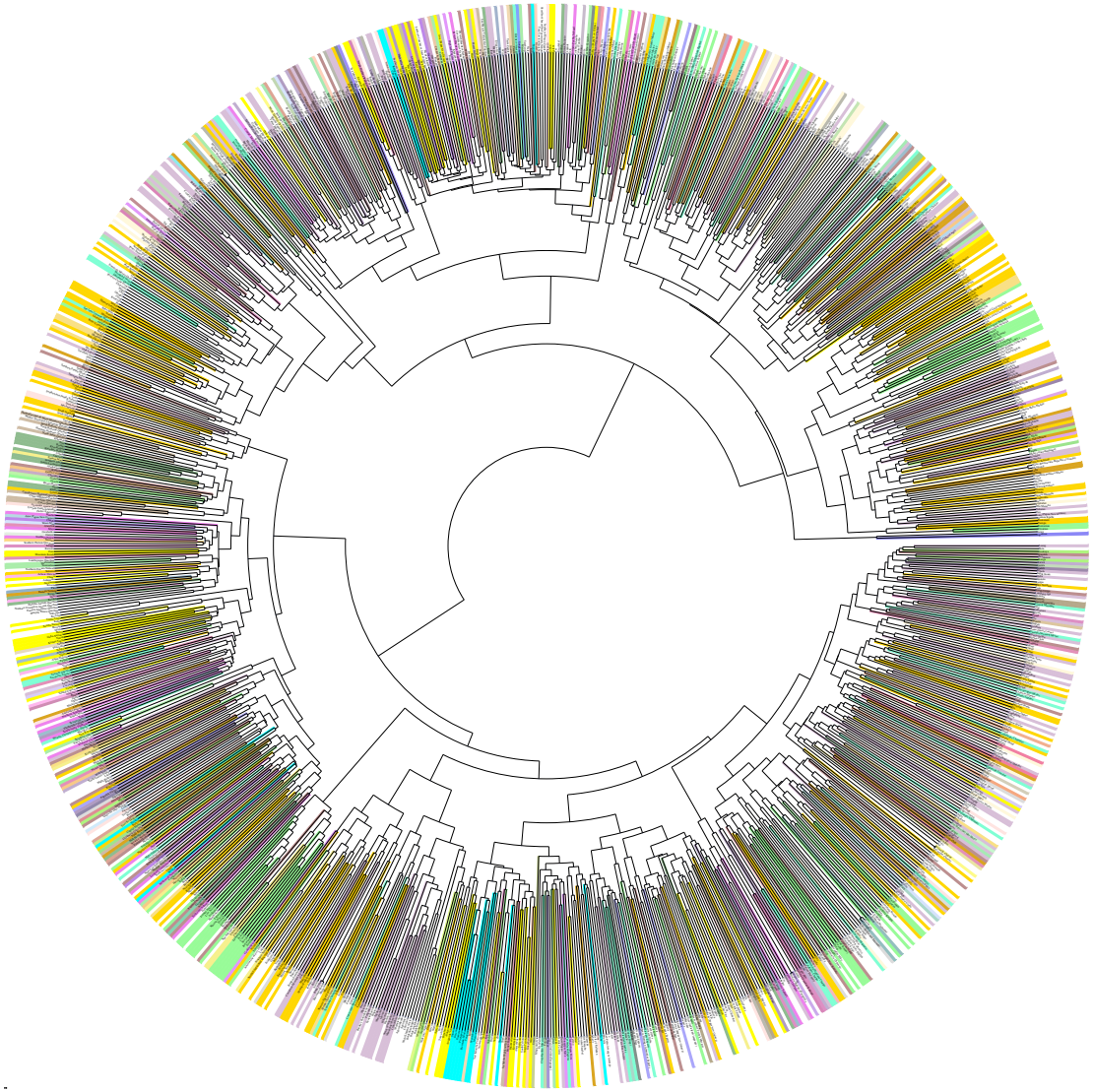
Past tense map: Torres Strait Creole ‘bin’



Outline

1. Problem definition
2. Linguistic Inspiration
3. Methods and Results
   - Data set
   - SuperPivot pipeline
     - Head pivot set
     - Subword-level marker finding
     - Linguistic Analysis
   - Language clustering based on tense marking
4. Contributions

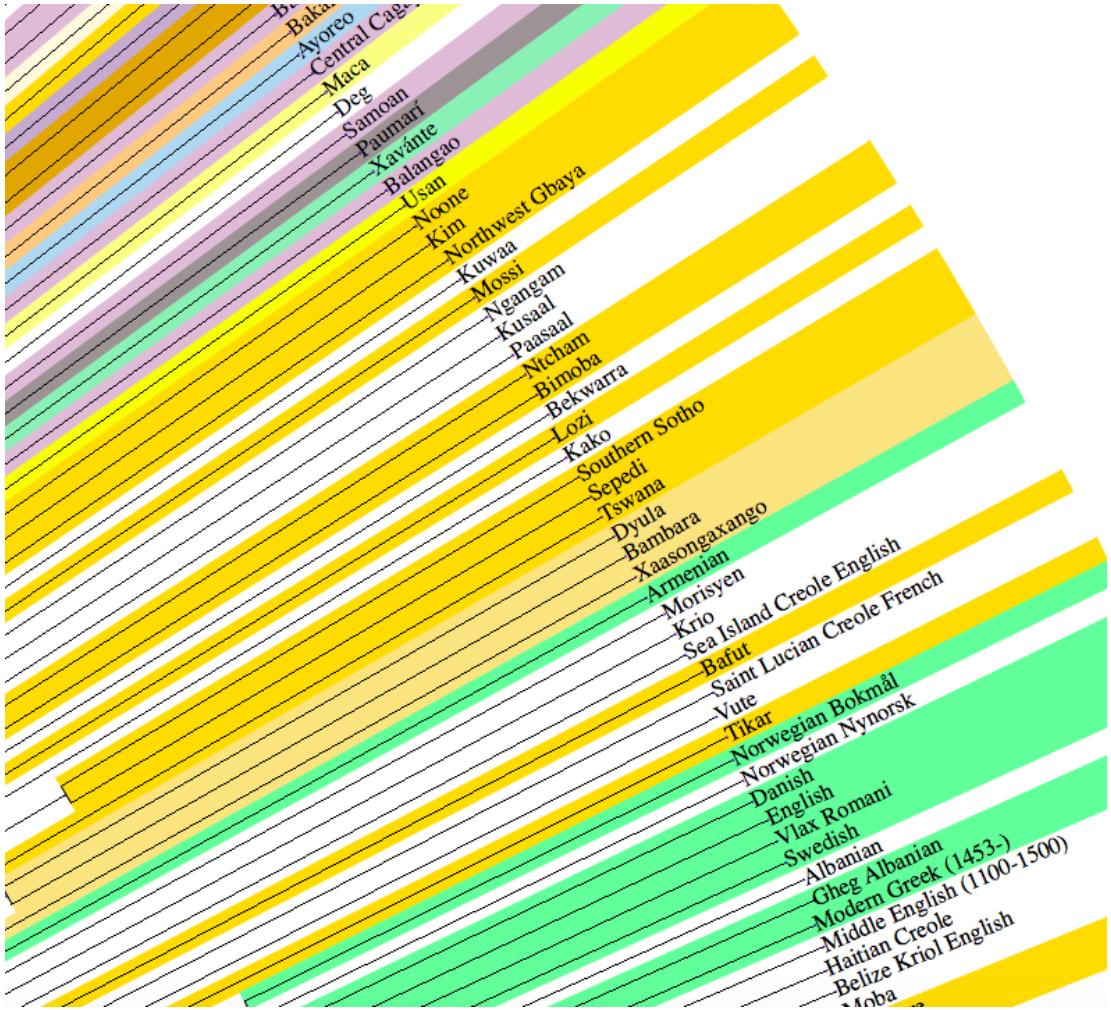
Language clustering based on tense marking



Remarks

- > 1000 languages
- Past present future
- Colored by family

Language clustering based on tense marking



Remarks

- > 1000 languages
- Past present future
- Colored by family

Summary of Contributions

- SuperPivot: computational framework for analysis of >1000 languages

Summary of Contributions

- SuperPivot: computational framework for analysis of >1000 languages
- Basic method: annotation projection

Summary of Contributions

- SuperPivot: computational framework for analysis of >1000 languages
- Basic method: annotation projection
- Instead of costly labels we project words

Summary of Contributions

- SuperPivot: computational framework for analysis of >1000 languages
- Basic method: annotation projection
- Instead of costly labels we project words
- Automated way of computing Cysouw's typological maps

Summary of Contributions

- SuperPivot: computational framework for analysis of >1000 languages
- Basic method: annotation projection
- Instead of costly labels we project words
- Automated way of computing Cysouw's typological maps
- Paper on arXiv:
  https://arxiv.org/abs/1704.08914

Summary of Contributions

- SuperPivot: computational framework for analysis of >1000 languages
- Basic method: annotation projection
- Instead of costly labels we project words
- Automated way of computing Cysouw's typological maps
- Paper on arXiv:
  https://arxiv.org/abs/1704.08914
- Practical Intro:
  https://github.com/ehsanasgari/CL2018SuperPivot

Summary of Contributions

- SuperPivot: computational framework for analysis of >1000 languages
- Basic method: annotation projection
- Instead of costly labels we project words
- Automated way of computing Cysouw's typological maps
- Paper on arXiv:
  https://arxiv.org/abs/1704.08914
- Practical Intro:
  https://github.com/ehsanasgari/CL2018SuperPivot