Past, Present, Future: A Computational Investigation of the Typology of Tense in 1000 Languages

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

[1]Center for Information and Language Processing, LMU Munich, Germany
[2]Applied Science and Technology, University of California, Berkeley, CA, USA,
inquiries@cislmu.org

Abstract

We present SuperPivot, an analysis method for low-resource languages that occur in a superparallel corpus, i.e., in a corpus that contains an order of magnitude more languages than parallel corpora currently in use. We show that SuperPivot performs well for the crosslingual analysis of the linguistic phenomenon of tense. We produce analysis results for more than 1000 languages, conducting – to the best of our knowledge – the largest crosslingual computational study performed to date. We extend existing methodology for leveraging parallel corpora for typological analysis by overcoming a limiting assumption of earlier work: We only require that a linguistic feature is overtly marked in a *few* of thousands of languages as opposed to requiring that it be marked in *all* languages under investigation.

1 Introduction

Significant linguistic resources such as machine-readable lexicons and part-of-speech (POS) taggers are available for at most a few hundred languages. This means that the majority of the languages of the world are low-resource. Low-resource languages like Fulani are spoken by tens of millions of people and are politically and economically important; e.g., to manage a sudden refugee crisis, NLP tools would be of great benefit. Even "small" languages are important for the preservation of the common heritage of humankind that includes natural remedies and linguistic and cultural diversity that can potentially enrich everybody. Thus, developing analysis methods for low-resource languages is one of the most important challenges of NLP today.

We address this challenge by proposing a new method for analyzing what we call *superparallel corpora*, corpora that are by an order of magnitude more parallel than corpora that have been available in NLP to date. The corpus we work with in this paper is the Parallel Bible Corpus (PBC) that consists of translations of the New Testament in 1169 languages. Given that no NLP analysis tools are available for most of these 1169 languages, how can we extract the rich information that is potentially hidden in such superparallel corpora?

The method we propose is based on two hypotheses. **H1 Existence of overt encoding.** For any important linguistic distinction f that is frequently encoded across languages in the world, there are a few languages that encode f overtly on the surface. **H2 Overt-to-overt and overt-to-non-overt projection.** For a language l that encodes f, a projection of f from the "overt languages" to l in the superparallel corpus will identify the encoding that l uses for f, both in cases in which the encoding that l uses is overt and in cases in which the encoding that l uses is non-overt. Based on these two hypotheses, our method proceeds in 5 steps.

1. Selection of a linguistic feature. We select a linguistic feature f of interest. Running example: We select past tense as feature f.

2. Heuristic search for head pivot. Through a heuristic search, we find a language l^h that contains a *head pivot* p^h that is highly correlated with the linguistic feature of interest.

Running example: "ti" in Seychelles Creole (CRS). CRS "ti" meets our requirements for a head pivot well as will be verified empirically in §3. First, "ti" is a surface marker: it is easily identifable through whitespace tokenization and it is not ambiguous, e.g., it does not have a second meaning apart from being a grammatical marker. Second, "ti" is a good marker for past tense in

terms of both "precision" and "recall". CRS has mandatory past tense marking (as opposed to languages in which tense marking is facultative) and "ti" is highly correlated with the general notion of past tense.

This does not mean that every clause that a linguist would regard as past tense is marked with "ti" in CRS. For example, some tense-aspect configurations that are similar to English present perfect are marked with "in" in CRS, not with "ti" (e.g., ENG "has commanded" is translated as "in ordonn").

Our goal is not to find a head language and a head pivot that is a perfect marker of f. Such a head pivot probably does not exist; or, more precisely, linguistic features are not completely rigorously defined. In a sense, one of the significant contributions of this work is that we provide more rigorous definitions of past tense across languages; e.g., "ti" in CRS is one such rigorous definition of past tense and it automatically extends (through projection) to 1000 languages in the superparallel corpus.

3. Projection of head pivot to larger pivot set. Based on an alignment of the head language to the other languages in the superparallel corpus, we project the head pivot to all other languages and search for highly correlated surface markers, i.e., we search for additional pivots in other languages. This projection to more pivots achieves three goals. First, it makes the method more *robust*. Relying on a single pivot would result in many errors due to the inherent noisiness of linguistic data and because several components we use (e.g., alignment of the languages in the superparallel corpus) are imperfect. Second, as we discussed above, the head pivot does not necessarily have high "recall"; our example was that CRS "ti" is not applied to certain clauses that would be translated using present perfect in English. Thus, moving to a larger pivot set *increases recall*. Third, as we will see below, the pivot set can be leveraged to create a *fine-grained map of the linguistic feature*. Consider clauses referring to eventualities in the past that English speakers would render in past progressive, present perfect and simple past tense. Our hope is that the pivot set will cover these distinctions, i.e., one of the pivots marks past progressive, but not present prefect and simple past, another pivot marks present perfect, but not the other two and so on. It is beyond the scope of this paper to verify that we can produce such an analysis for all linguistic features, but a promising example of this type of map, including distinctions like progressive and perfective aspect, is given in §4.

Running example: We compute the correlation of "ti" with words in other languages and select the 100 highest correlated words as pivots. Examples of pivots we find this way are Torres Strait Creole "bin" (from English "been") and Tzotzil "laj". "laj" is a perfective marker, e.g., "Laj meltzaj -uk" 'LAJ be-made subj' means "It's done being built" (Aissen, 1987).

4. Projection of pivot set to all languages. Now that we have a large pivot set, we project the pivots to all other languages to search for linguistic devices that express the linguistic feature f. Up to this point, we have made the assumption that it is easy to segment text in all languages into pieces of a size that is not too small (individual characters of the Latin alphabet would be too small) and not too large (entire sentences as tokens would be too large). Segmentation on standard delimiters is a good approximation for the majority of languages – but not for all: it undersegments some (e.g., the polysynthetic language Inuit) and oversegments others (e.g., languages that use punctuation marks as regular characters).

For this reason, we do not employ tokenization in this step. Rather we search for character n-grams ($2 \leq n \leq 6$) to find linguistic devices that express f. This implementation of the search procedure is a limitation – there are many linguistic devices that cannot be found using it, e.g., templates in templatic morphology. We leave addressing this for future work (§6).

Running example: We find "-ed" for English and "-te" for German as surface features that are highly correlated with the 100 past tense pivots.

5. Linguistic analysis. The result of the previous steps is a superparallel corpus that is richly annotated with information about linguistic feature f. This structure can be exploited for *the analysis of a single language* l^i that may be the focus of a linguistic investigation. Starting with the character n-grams that were found in the step "projection of pivot set to all languages", we can explore their use and function, e.g, for the mined n-gram "-ed" in English (assuming English is the language l^i and it is unfamiliar to us). Many of the other 1000 languages provide annotations of linguistic

feature f for l^i: both the languages that are part of the pivot set (e.g., Tzotzil "laj") and the mined n-grams in other languages that we may have some knowledge of (e.g., "-te" in German).

We can also use the structure we have generated for *typological analysis across languages* following the work of Michael Cysouw. He has pioneered a new methodology for typology ((Cysouw, 2014), §5). We do not contribute any innovations to typology in this paper, but our method is a significant advancement computationally over Cysouw's work because we overcome many of his limiting assumptions. Most importantly, our method scales to thousands of languages as we demonstrate below whereas Cysouw worked on a few dozen.

Running example: We sketch the type of analysis that our new method makes possible in §4.

The above steps "1. heuristic search for head pivot" and "2. projection of head pivot to larger pivot set" are based on H1: we assume the **existence of overt coding** in a subset of languages.

The above steps "2. projection of head pivot to larger pivot set" and "3. projection of pivot set to all languages" are based on H2: we assume that **overt-to-overt and overt-to-non-overt projection** is possible.

In the rest of the paper, we will refer to the method that consists of steps 1 to 5 as *SuperPivot*: "linguistic analysis of SUPERparallel corpora using surface PIVOTs".

We make three contributions. (i) Our basic hypotheses are H1 and H2. (H1) For an important linguistic feature, there exist a few languages that mark it overtly and easily recognizably. (H2) It is possible to project overt markers to overt and non-overt markers in other languages. Based on these two hypotheses we design SuperPivot, a new method for analyzing highly parallel corpora, and show that it performs well for the crosslingual analysis of the linguistic phenomenon of tense. (ii) Given a superparallel corpus, SuperPivot can be used for the analysis of *any low-resource language* represented in that corpus. In the supplementary material, we present results of our analysis for three tenses (past, present, future) for 1163[1] languages. An evaluation of accuracy is presented in Table 3.2. (iii) We extend Michael Cysouw's pioneering work on typological analysis using parallel corpora by overcoming several limiting factors. The most important is that Cysouw's method is only applicable if markers of the relevant linguistic feature are recognizable on the surface in *all* languages. In contrast, we only assume that markers of the relevant linguistic feature are recognizable on the surface in *a small number of* languages.

2 SuperPivot: Description of method

1. Selection of a linguistic feature. The linguistic feature of interest f is selected by the person who performs a SuperPivot analysis, i.e., by a linguist, NLP researcher or data scientist. Henceforth, we will refer to this person as the linguist.

In this paper, $f \in F = \{\text{past}, \text{present}, \text{future}\}$.

2. Heuristic search for head pivot. There are several ways for finding the head language and the head pivot. Perhaps the linguist knows a language that has a good head pivot. Or she is a trained typologist and can find the head pivot by consulting the typological literature.

In this paper, we use our knowledge of English and an alignment from English to all other languages to find head pivots. (See below for details on alignment.) We define a "query" in English and search for words that are highly correlated to the query in other languages. For future tense, the query is simply the word "will", so we search for words in other languages that are highly correlated with "will". For present tense, the query is the union of "is", "are" and "am". So we search for words in other languages that are highly correlated with the "merger" of these three words. For past tense, we POS tag the English part of PBC and merge all words tagged as past tense into one past tense word.[2] We then search for words in other languages that are highly correlated with this artificial past tense word.

As an additional constraint, we do not select the most highly correlated word as the head pivot, but the most highly correlated word in a Creole language. Our rationale is that Creole languages are more regular than other languages because they are young and have not accumulated "historical baggage" that may make computational analysis more difficult.

Table 1 lists the three head pivots for F.

3. Projection of head pivot to larger pivot set. We first use fast_align (Dyer et al., 2013) to align

[1] We exclude six of the 1169 languages because they do not share enough verses with the rest.

[2] Past tense is defined as tags BED, BED*, BEDZ, BEDZ*, DOD*, VBD, DOD. We use NLTK (Bird, 2006).

the head language to all other languages in the corpus. This alignment is on the word level.

We compute a score for each word in each language based on the number of times it is aligned to the head pivot, the number of times it is aligned to another word and the total frequencies of head pivot and word. We use χ^2 (Casella and Berger, 2008) as the score throughout this paper. Finally, we select the k words as pivots that have the highest association score with the head pivot.

We impose the constraint that we only select one pivot per language. So as we go down the list, we skip pivots from languages for which we already have found a pivot. We set $k = 100$ in this paper. Table 1 gives the top 10 pivots.

4. Projection of pivot set to all languages. As discussed above, the process so far has been based on tokenization. To be able to find markers that cannot be easily detected on the surface (like "-ed" in English), we identify non-tokenization-based character n-gram features in step 4.

The immediate challenge is that without tokens, we have no alignment between the languages anymore. We could simply assume that the occurrence of a pivot has scope over the entire verse. But this is clearly inadequate, e.g., for the sentence "I arrived yesterday, I'm staying today, and I will leave tomorrow", it is incorrect to say that it is marked as past tense (or future tense) in its entirety. Fortunately, the verses in the New Testament mostly have a simple structure that limits the variation in where a particular piece of content occurs in the verse. We therefore make the assumption that a particular relative position in language l_1 (e.g., the character at relative position 0.62) is aligned with the same relative position in l_2 (i.e., the character at relative position 0.62). This is likely to work for a simple example like "I arrived yesterday, I'm staying today, and I will leave tomorrow" across languages.

In our analysis of errors, we found many cases where this assumption breaks down. A well-known problematic phenomenon for our method is the difference between, say, VSO and SOV languages: the first class puts the verb at the beginning, the second at the end. However, keep in mind that we accumulate evidence over $k = 100$ pivots and then compute aggregate statistics over the entire corpus. As our evaluation below shows, the "linear alignment" assumption does not seem to do much harm given the general robustness of our method.

One design element that increases robustness is that we find the two positions in each verse that are most highly (resp. least highly) correlated with the linguistic feature f. Specifically, we compute the relative position x of each pivot that occurs in the verse and apply a Gaussian filter ($\sigma = 6$ where the unit of length is the character), i.e., we set $p(x) \approx 0.066$ (0.066 is the density of a Gaussian with $\sigma = 6$ at $x = 0$) and center a bell curve around x. The total score for a position x is then the sum of the filter values at x summed over all occurring pivots. Finally, we select the positions x_{min} and x_{max} with lowest and highest values for each verse.

χ^2 is then computed based on the number of times a character n-gram occurs in a window of size w around x_{max} (positive count) and in a window of size w around x_{min} (negative count). Verses in which no pivot occurs are used for the negative count in their entirety. The top-ranked character n-grams are then output for analysis by the linguist. We set $w = 20$.

5. Linguistic analysis. We now have created a structure that contains rich information about the linguistic feature: for each verse we have relative positions of pivots that can be projected across languages. We also have maximum positions within a verse that allow us to pinpoint the most likely place in the vicinity of which linguistic feature f is marked in all languages. This structure can be used for the analysis of individual low-resource languages as well as for typological analysis. We will give an example of such an analysis in §4.

6. Hierarchical clusterings of markers and languages. As an additional evaluation, we worked on hierarchical clusterings of past, present and future pivots. As detailed in §2.4, we represent each verse by a vector of length 100 showing which pivot markers are used to express this verse. The other way of looking at these data is that for each marker we have an occurrence distribution over verses and we may exploit these data to demonstrate the distance between markers. For the purpose of comparing two markers, we propose calculation of the Jensen-Shannon divergence between the normalized occurrence distribution over verses:

$$D_{m_{p_i}, m_{p_j}} = JSD(\hat{m_{p_i}}, \hat{m_{p_j}}),$$

where $\hat{m_{p_i}}$ and $\hat{m_{p_i}}$, are the normalized occurrence distributions over verses. We compare the obtained distance between markers with genetic

distance of their corresponding languages using WALS information (Dryer et al., 2005). For visualization purposes, we perform Unweighted Pair Group Method with Arithmetic Mean (UPGMA) hierarchical clustering on the pairwise distance matrix of the marker for each tense separately (Johnson, 1967).

In addition to clustering of pivot markers for each tense separately, we performed the same comparison for all top markers of 1107 languages[3] and take the average distances of languages in past, present, and future marking. This allows us to compare the average tense behavior of languages.

$$D_{l_i,l_j} = \frac{1}{3}(JSD_{past} + JSD_{present} + JSD_{future}),$$

3 Data, experiments and results

3.1 Data

We use a New Testament subset of the Parallel Bible Corpus (PBS) (Mayer and Cysouw, 2014) that consists of 1556 translations of the the Bible in 1169 unique languages. We consider two languages to be different if they have different ISO 639-3 codes.

The translations are aligned on the verse level. However, many translations do not have complete coverage, so that most verses are not present in at least one translation. One reason for this is that sometimes several consecutive verses are merged, so that one verse contains material that is in reality not part of it and the merged verses may then be missing from the translation. Thus, there is a trade-off between number of parallel translations and number of verses they have in common. Although some preprocessing was done by the authors of the resource, many translations are not preprocessed. For example, Japanese is not tokenized. We also observed some incorrectness and sparseness in the metadata. One example is that one Fijian translation (see §4) is tagged fij_hindi, but it is Fijian, not Fiji Hindi.

We use the 7958 verses with the best coverage across languages.

3.2 Experiments

1. Selection of a linguistic feature. We conduct three experiments for the linguistic features past tense, present tense and future tense.

[3]We exclude languages that have fewer than 7000 verses in common with the pivot language to ensure quality of marker.

2. Heuristic search for head pivot. We use the queries described in §2 for finding the following three head pivots. (i) Past tense head pivot: "ti" in Seychellois Creole (CRS) (McWhorter, 2005). (ii) Present tense head pivot: "ta" in Papiamentu (PAP) (Andersen, 1990). (iii) Future tense head pivot: "bai" in Tok Pisin (TPI) (Traugott, 1978; Sankoff, 1990).

3. Projection of head pivot to larger pivot set. Using the method described in §2, we project each head pivot to a set of $k = 100$ pivots. Table 1 gives the top 10 pivots for each tense.

4. Projection of pivot set to all languages. Using the method described in §2, we compute highly correlated character n-gram features, $2 \leq n \leq 6$, for all 1163 languages.

See §4 for the last step of SuperPivot: **5. Linguistic analysis.**

3.3 Evaluation

We rank n-gram features and retain the top 10, for each linguistic feature, for each language and for each n-gram size. We process 1556 translations. Thus, in total, we extract $1556 \times 5 \times 10$ n-grams.

Table 3.2 shows Mean Reciprocal Rank (MRR) for 10 languages. The rank for a particular ranking of n-grams is the first n-gram that is highly correlated with the relevant tense; e.g., character subsequences of the name "Paulus" are evaluated as incorrect, the subsequence "-ed" in English as correct for past. MRR is averaged over all n-gram sizes, $2 \leq n \leq 6$. Chinese has consistent tense marking only for future, so results are poor. Russian and Polish perform poorly because their central grammatical category is aspect, not tense. The poor performance on Arabic is due to the limits of character n-gram features for a "templatic" language.

During this evaluation, we noticed a surprising amount of variation within translations of one language; e.g., top-ranked n-grams for some German translations include names like "Paulus". We suspect that for literal translations, linear alignment (§2) yields good n-grams. But many translations are free, e.g., they change the sequence of clauses. This deteriorates mined n-grams. See §6.

Hierarchical clusterings of markers. Hierarchical clusterings of past, present and future pivots using JSD between the normalized occurrence distribution over verses are shown in Figure 1, Figure 2, and Figure 3 for past, present, and future

	past			present			future		
	code	language	pivot	code	language	pivot	code	language	pivot
head pivots	CRS	Seychelles C.	*ti*	PAP	Papiamentu	*ta*	TPI	Tok Pisin	*bai*
	GUX	Gourmanchéma	*den*	NOB	Norwegian Bokmål	*er*	LID	Nyindrou	*kameh*
	MAW	Mampruli	*daa*	HIF	Fiji Hindi	*hei*	GUL	Sea Island C.	*gwine*
	GFK	Patpatar	*ga*	AFR	Afrikaans	*is*	TGP	Tangoa	*pa*
	YAL	Yalunka	*yi*	DAN	Danish	*er*	BUK	Bugawac	*oc*
	TOH	Gitonga	*di*	SWE	Swedish	*är*	BIS	Bislama	*bambae*
	DGI	Northern Dagara	*tı*	EPO	Esperanto	*estas*	PIS	Pijin	*bae*
	BUM	Bulu (Cameroon)	*nga*	ELL	Greek	*είναι*	APE	Bukiyip	*eke*
	TCS	Torres Strait C.	*bin*	HIN	Hindi	*haai*	HWC	Hawaiian C.	*goin*
	NDZ	Ndogo	*gìì*	NAQ	Khoekhoe	*ra*	NHR	Nharo	*gha*

Table 1: Top ten past, present, and future tense pivots extracted from 1163 languages. C. = Creole

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

Table 2: MRR results for step 4. See text for details.

tenses respectively. In addition to markers clusterings, the average tense behavior clustering of 1107 languages is depicted in Figure 4. In these figures languages are colored based on their language families using WALS (Dryer et al., 2005), languages without family information on WALS are uncolored. We observed that most of pivot past and future markers belong to Niger Congo family and present markers are mostly within Indo-European family. It can be seen that in many cases languages with the same family behave accordingly in tense marking. For instance, in past tense marking Oto-Manguean languages use almost the same marker of ni with small writing variations (Figure 1). Although Tezoatlán Mixtec did not have a record on WALS, since its marker is the same as other Oto-Manguean languages and works almost identical to ni in Oto-Manguean languages, we may guess this language is also Oto-Manguean, which turned out to be true when we performed further searches.[4] There were many of such cases for which we could guess the family of language based on their tense marking similarities in Figure 1, Figure 2 and Figure 3.

We use normalized JSD $(0 \leq JSD \leq 1)$ for comparison of each pair of languages/markers; this allows us to investigate whether a simple threshold of 0.5 accurately predicts whether two languages are genetically related or not. The results are summarized in Table 3.3. Although the average tense marking divergence has a low recall, it expresses a high precision of 0.36, where the random chance is $\frac{1}{103} \approx 0.01$. Thus, it means that if divergence of tense marking is low the languages are very likely to be genetically related. This conclusion is supported by Figure 4 where many small clusters of nodes have the same color. This suggests that our method may help in completion of WALS.

4 A map of past tense

To illustrate the potential of our method we select five out of the 100 past tense pivots that give rise to large clusters of distinct combinations. Starting with CRS, we find other pivots that "split" the set of verses that contain the CRS past tense pivot "ti" into two parts that have about the same size. This gives us two sets. We now look for a pivot that splits one of these two sets about evenly and so on. After iterating four times, we arrive at five pivots: CRS "ti", Fijian (FIJ) "qai", Hawaiian Creole (HWC) "wen", Torres Strait Creole (TCS) "bin" and Tzotzil (TZO) "laj".

Figure 5 shows a t-SNE (Maaten and Hinton, 2008) visualization of the large clusters of combinations that are found for these five languages, including one cluster of verses that do not contain any of the five pivots.

This figure is a map of past tense for all 1163 languages, not just for CRS, FIJ, HWC, TCS and TZO: once the interpretation of a particular cluster has been established based on CRS, FIJ, HWC, TCS and TZO, we can investigate this cluster in the 1164 other languages by looking at the verses

[4] https://www.ethnologue.com/language/mxb

	avg tense ($\frac{696}{1107}$ lang. - 103 fam.)	past ($\frac{55}{100}$ lang. - 15 fam.)	present ($\frac{70}{100}$ lang. - 17 fam.)	future ($\frac{44}{100}$ lang. - 16 fam.)
accuracy	0.93	0.55	0.81	0.58
precision	0.36	0.18	0.75	0.16
recall	0.01	0.59	0.37	0.61
TNR	0.99	0.55	0.96	0.58

Table 3: Language family similarity prediction results based on coordinated marking of verses. Only languages with records on WALS are included in this evaluation. TNR: true negative rate.

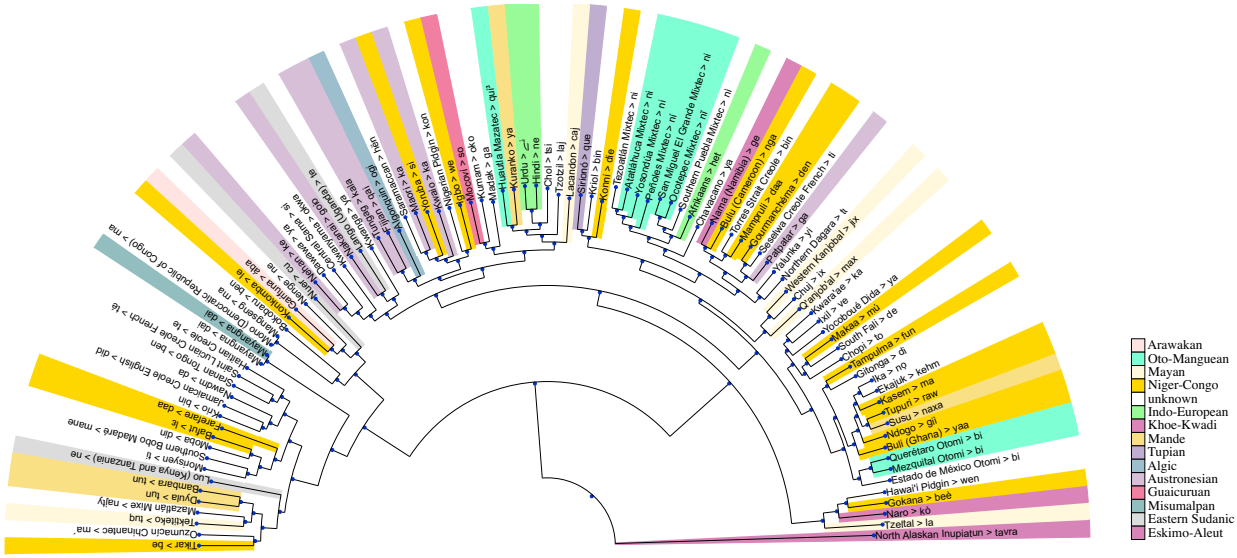

Figure 1: Clustering of 100 pivot past tense markers. Each node is colored based on its family information. Languages with no record on WALS remained white. This clustering is based on JSD of markers in marking 5960 verses in bible. We observed that most of pivot past and future markers belong to Niger Congo family and present markers are mostly within Indo-European family. It can be seen that in many cases languages with the same family behave accordingly in tense marking. For instance, in past tense marking Oto-Manguean languages use almost the same marker of ni with small writing variations (Figure 1). Although Tezoatlán Mixtec did not have a record on WALS, since its marker is the same as other Oto-Manguean languages and works almost identical to ni in Oto-Manguean languages, we may guess this language is also Oto-Manguean, which turned out to be true when we performed further searches.

that are members of this cluster. This methodology supports the empirical investigation of questions like "how is progressive past tense expressed in language X"? We just need to look up the cluster(s) that correspond to progressive past tense, look up the verses that are members and retrieve the text of these verses in language X.

To give the reader a flavor of the distinctions that are reflected in these clusters, we now list phenomena that are characteristic of verses that contain only one of the five pivots; these phenomena identify properties of one language that the other four do not have.

CRS "ti". CRS has a set of markers that can be systematically combined, in particular, a progressive marker "pe" that can be combined with the past tense marker "ti". As a result, past progressive sentences in CRS are generally marked with "ti". Example: "43004031 Meanwhile, the disciples were urging Jesus, 'Rabbi, eat something.'" "crs_bible 43004031 Pandan sa letan, bann disip ti pe sipliy Zezi, 'Met! Manz en pe.'"

The other four languages do not consistently use the pivot for marking the past progressive; e.g., HWC uses "was begging" in 43004031 (instead of "wen") and TCS uses "kip tok strongwan" 'keep talking strongly' in 43004031 (instead of "bin").

FIJ "qai". This pivot means "and then". It is highly correlated with past tense in the New Testament because most sequential descriptions of events are descriptions of past events. But there are also some non-past sequences. Example:

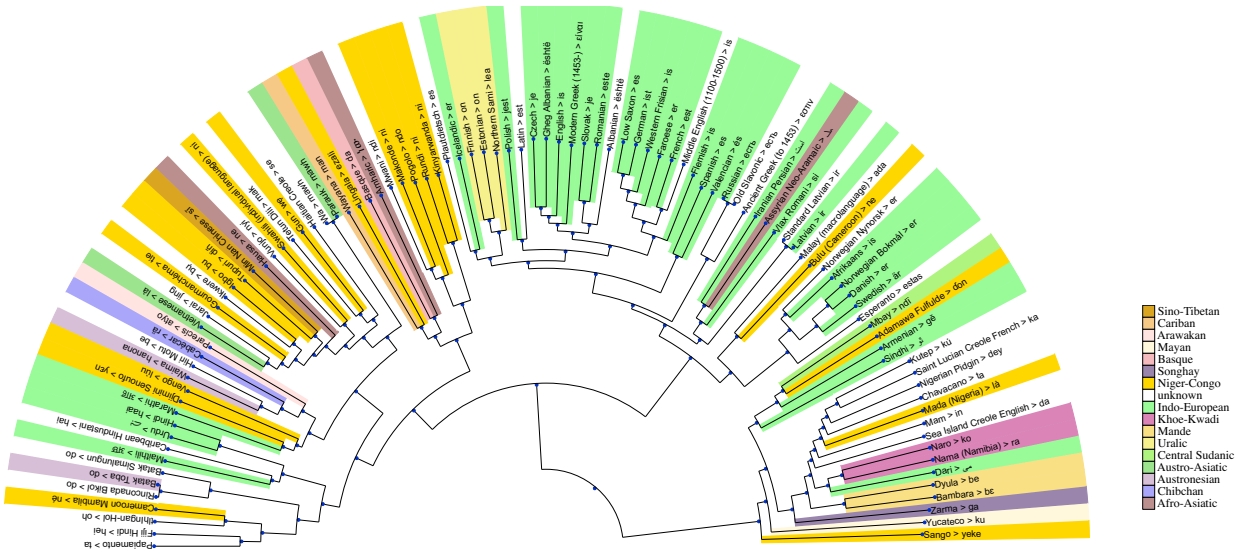

Figure 2: Clustering of 100 pivot present tense markers. Each node is colored based on its family information. Languages with no record on WALS remained white. This clustering is based on JSD of markers in marking 6590 verses in bible. It can be seen that in many cases languages with the same family behave accordingly in tense marking.

"eng_newliving 44009016 And I will show him how much he must suffer for my name's sake." "fij_hindi 44009016 Au na qai vakatakila vua na levu ni ka e na sota kaya e na vukuqu." This verse is future tense, but it continues a temporal sequence (it starts in the preceding verse) and therefore FIJ uses "qai". The pivots of the other four languages are not general markers of temporal sequentiality, so they are not used for the future.

HWC "wen". HWC is less explicit than the other four languages in some respects and more explicit in others. It is less explicit in that not all sentences in a sequence of past tense sentences need to be marked explicitly with "wen", resulting in some sentences that are indistinguishable from present tense. On the other hand, we found many cases of noun phrases in the other four languages that refer implicitly to the past, but are translated as a verb with explicit past tense marking in HWC. Examples: "hwc_2000 40026046 Da guy who wen set me up . . . " 'the guy who WEN set me up', "eng_newliving 40026046 . . . my betrayer . . . "; "hwc_2000 43008005 . . . Moses wen tell us in da Rules . . . " 'Moses WEN tell us in the rules', "eng_newliving 43008005 The law of Moses says . . . "; "hwc_2000 47006012 We wen give you guys our love . . . ", "eng_newliving 47006012 There is no lack of love on our part . . . ". In these cases, the other four languages (and English too) use a noun phrase with no tense marking that is translated as a tense-marked clause in HWC.

While preparing this analysis, we realized that HWC "wen" unfortunately does not meet one of the criteria we set out for pivots: it is not unambiguous. In addition to being a past tense marker (derived from standard English "went"), it can also be a conjunction, derived from "when". This ambiguity is the cause for some noise in the clusters marked for presence of HWC "wen" in the figure.

TCS "bin". Conditionals is one pattern we found in verses that are marked with TCS "bin", but are not marked for past tense in the other four languages. Example: "tcs_bible 46015046 Wanem i bin kam pas i da nomal bodi ane den da spiritbodi i bin kam apta." 'what came first is the normal body and then the spirit body came after', "eng_newliving 46015046 What comes first is the natural body, then the spiritual body comes later." Apparently, "bin" also has a modal aspect in TCS: generic statements that do not refer to specific events are rendered using "bin" in TCS whereas the other four languages (and also English) use the default unmarked tense, i.e., present tense.

TZO "laj". This pivot indicates perfective aspect. The other four past tense pivots are not perfective markers, so that there are verses that are marked with "laj", but not marked with the past tense pivots of the other four languages. Example: "tzo_huixtan 40010042 . . . ja'ch-ac'bat ben-

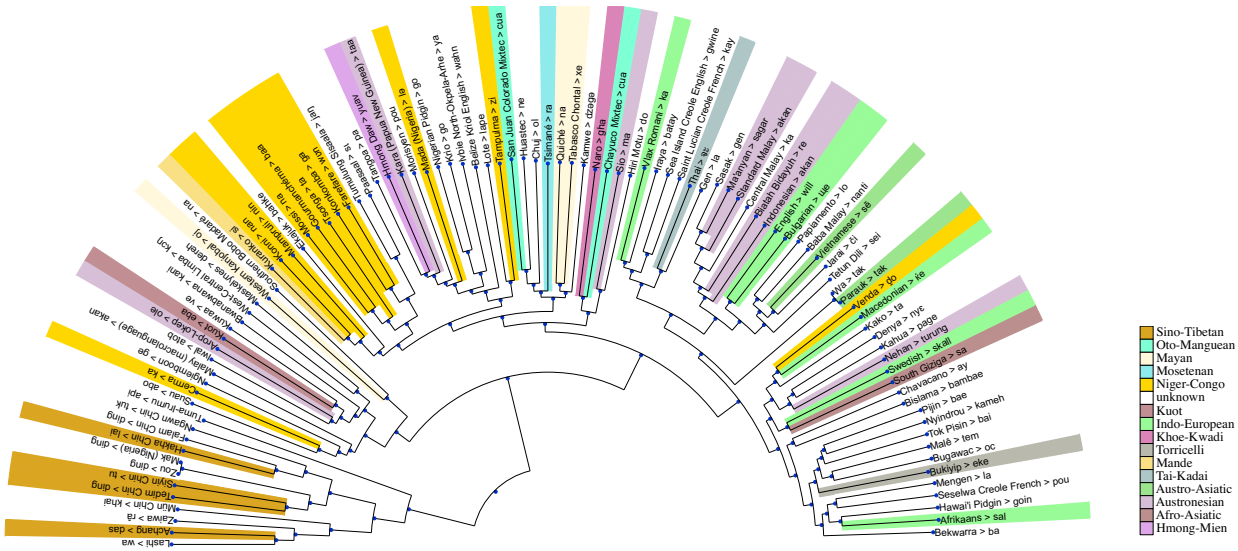

Figure 3: Clustering of 100 pivot future tense markers. Each node is colored based on its family information. Languages with no record on WALS remained white. This clustering is based on JSD of markers in marking 5733 verses in bible. It can be seen that in many cases languages with the same family behave accordingly in tense marking.

dición yu'un hech laj spas . . . " (literally "a blessing . . . LAJ make"), "eng_newliving 40010042 . . . you will surely be rewarded." Perfective aspect and past are correlated in the real world since most events that are viewed as simple wholes are in the past. But future events can also be viewed this way as the example shows.

Similar maps for present and future tenses are presented in the Figure 6 and Figure 7.

5 Related work

Our work is inspired by (Cysouw, 2014; Cysouw and Wälchli, 2007); see also (Dahl, 2007; Wälchli, 2010). Cysouw creates maps like Figure 5 by manually identifying occurrences of the proper noun "Bible" in a parallel corpus of Jehovah's Witnesses' texts. Areas of the map correspond to semantic roles, e.g., the Bible as actor (it tells you to do something) or as object (it was printed). This is a definition of semantic roles that is complementary to and different from prior typological research because it is empirically grounded in real language use across a large number of languages. It allows typologists to investigate traditional questions from a radically new perspective.

The field of **typology** is important for both theoretical (Greenberg, 1960; Whaley, 1996; Croft, 2002) and computational (Heiden et al., 2000; Santaholma, 2007; Bender, 2009, 2011) linguistics. Typology is concerned with all areas of linguistics: morphology (Song, 2014), syntax (Comrie, 1989; Croft, 2001; Croft and Poole, 2008; Song, 2014), semantic roles (Hartmann et al., 2014; Cysouw, 2014), semantics (Koptjevskaja-Tamm et al., 2007; Dahl, 2014; Wälchli and Cysouw, 2012; Sharma, 2009) etc. Typological information is important for many NLP tasks including discourse analysis (Myhill and Myhill, 1992), information retrieval (Pirkola, 2001), POS tagging (Bohnet and Nivre, 2012), parsing (Bohnet and Nivre, 2012; McDonald et al., 2013), machine translation (Hajič et al., 2000; Kunchukuttan and Bhattacharyya, 2016) and morphology (Bohnet et al., 2013).

Tense is a central phenomenon in linguistics and the languages of the world differ greatly in whether and how they express tense (Traugott, 1978; Bybee and Dahl, 1989; Dahl, 2000, 1985; Santos, 2004; Dahl, 2007; Santos, 2004; Dahl, 2014).

Low resource. Even resources with the widest coverage like World Atlas of Linguistic Structures (WALS) (Dryer et al., 2005) have little information for hundreds of languages. Many researchers have taken advantage of parallel information for extracting linguistic knowledge in low-resource settings (Resnik et al., 1997; Resnik, 2004; Mihalcea and Simard, 2005; Mayer and Cysouw, 2014;

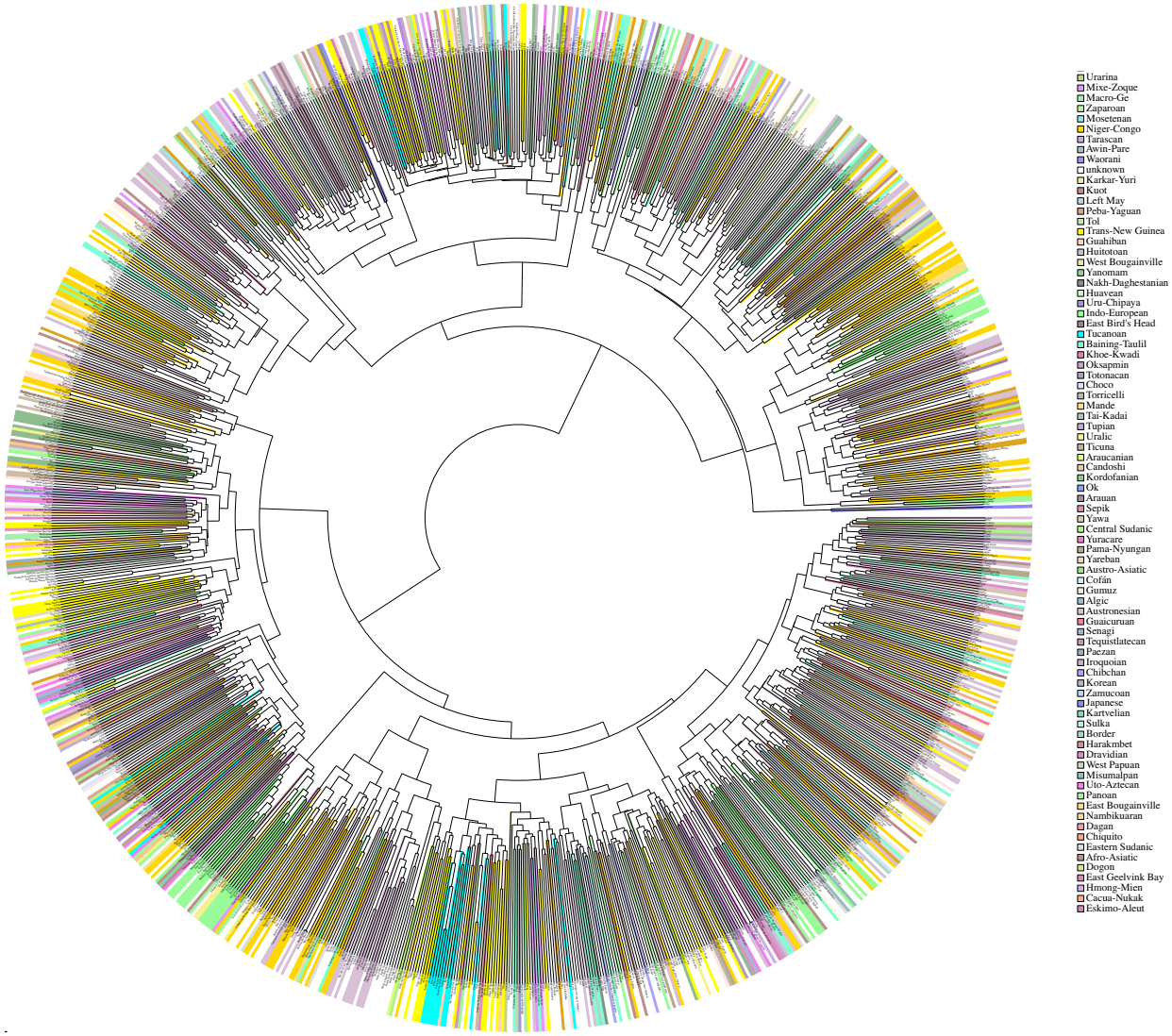

Figure 4: Clustering of 1107 languages based on the average Jensen-Shannon divergence in past, present, and future marking of their respective top markers. Each node is colored based on its family information. Languages with no record on WALS remained white. It can be seen that many small clusters of nodes have the same color, which together with our quantitative evaluation supports that if divergence of tense marking is low the languages are very likely to be genetically related.

Christodouloupoulos and Steedman, 2015; Lison and Tiedemann, 2016).

Parallel projection. Parallel projection across languages has been used for variety of NLP tasks. Machine translation aside, which is the most natural task on parallel corpora (Brown et al., 1993), parallel projection has been used for sense disambiguation (Ide, 2000), parsing (Hwa et al., 2005), paraphrasing (Bannard and Callison-Burch, 2005), part-of-speech tagging (Mukerjee et al., 2006), coreference resolution (de Souza and Orăsan, 2011), event marking (Nordrum, 2015), morphological segmentation (Chung et al., 2016), bilingual analysis of linguistic marking (McEnery and Xiao, 1999; Xiao and McEnery, 2002), as well as language classification (Asgari and Mofrad, 2016).

6 Conclusion

We presented SuperPivot, an analysis method for low-resource languages that occur in a superparallel corpus, i.e., in a corpus that contains an order of magnitude more languages than parallel corpora currently in use. We showed that SuperPivot performs well for the crosslingual analysis of the linguistic phenomenon of tense. We produced analysis results for more than 1000 languages, conducting – to the best of our knowledge

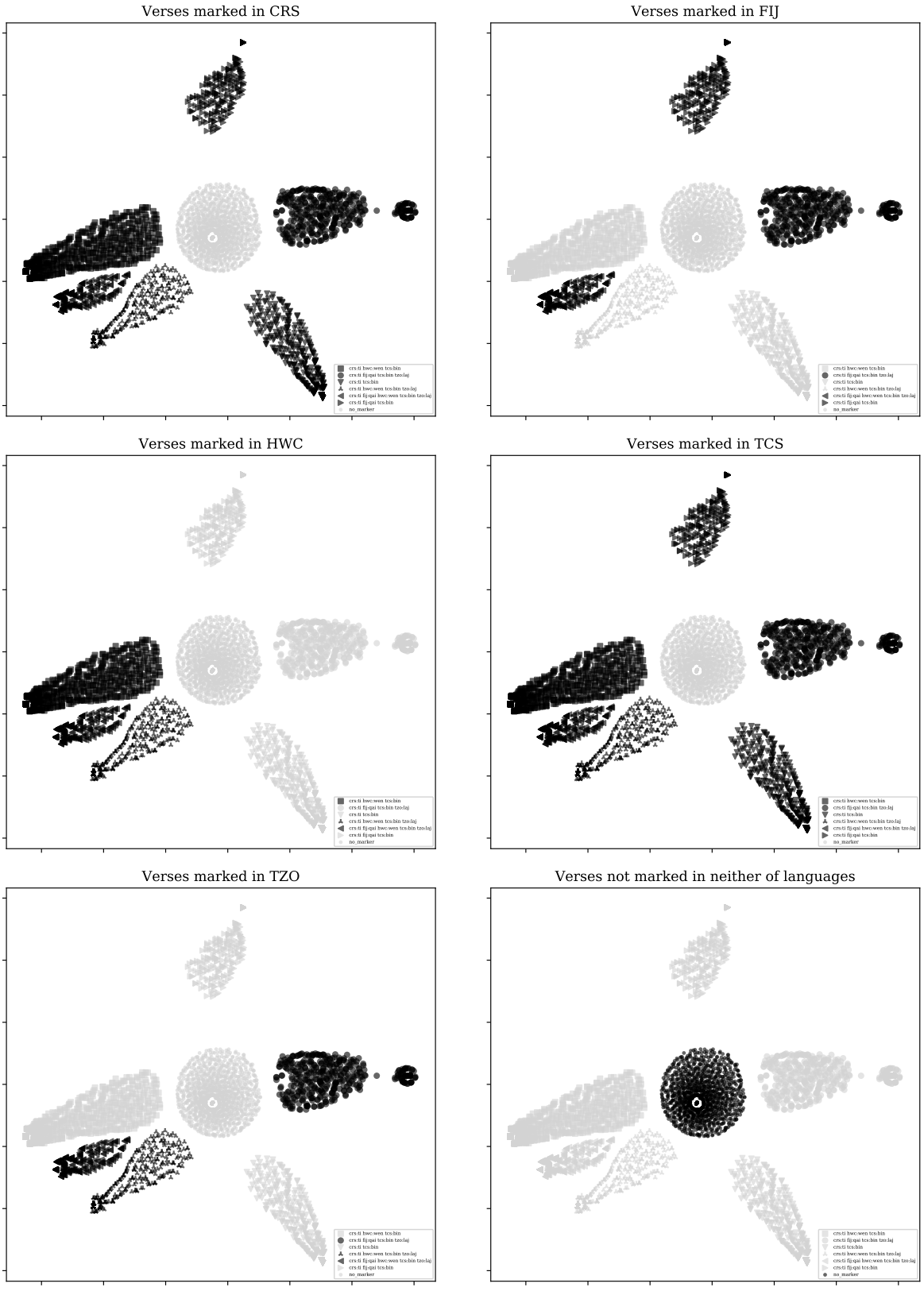

Figure 5: A map of past tense based on the largest clusters of verses with particular combinations of the past tense pivots from Seychellois Creole (CRS), Fijian (FIJ), Hawaiian Creole (HWC), Torres Strait Creole (TCS) and Tzotzil (TZO). For each of the five languages, we present a subfigure that highlights the subset of verse clusters that are marked by the pivot of that language. The sixth subfigure highlights verses not marked by any of the five pivots.

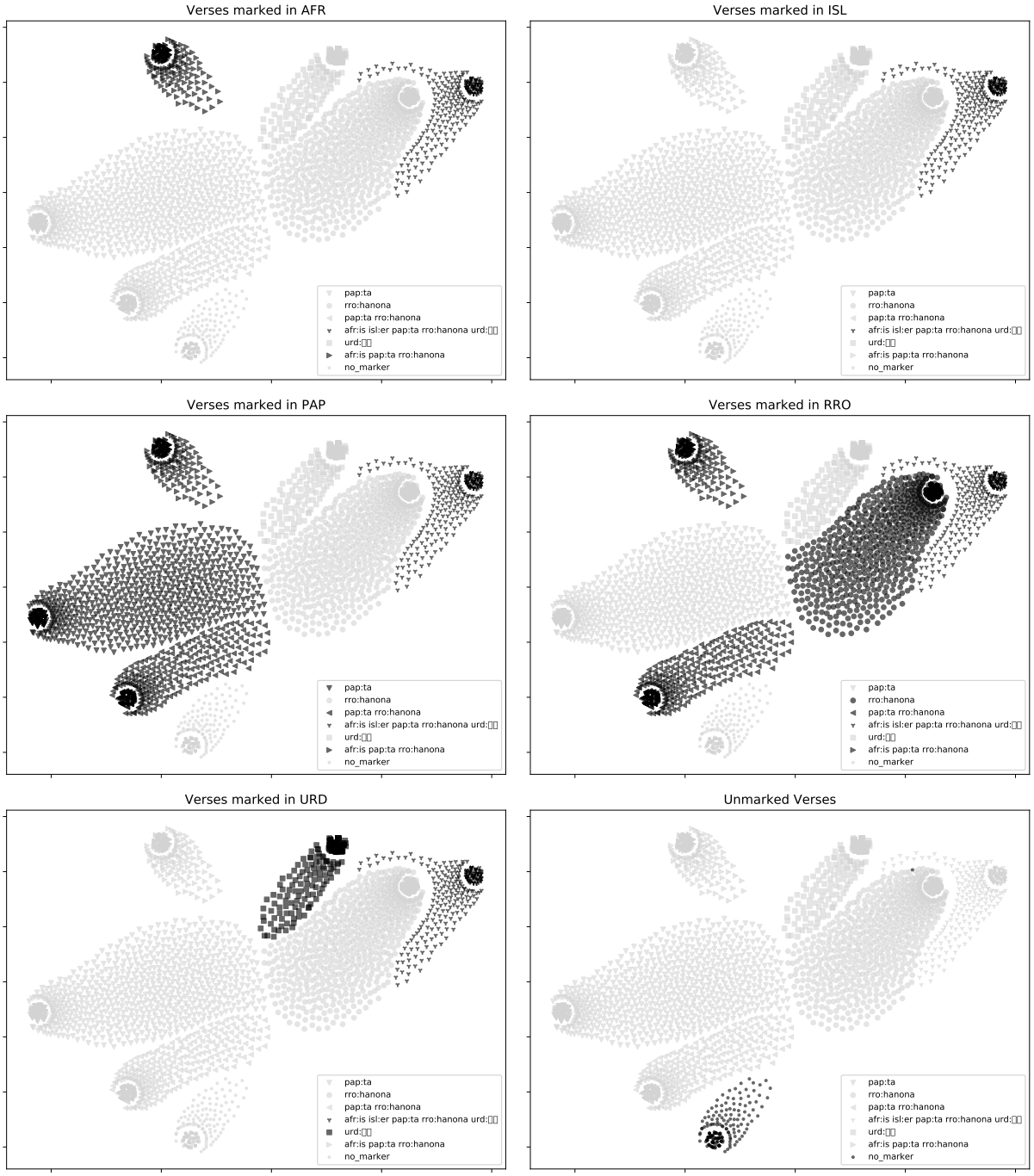

Figure 6: A map of present tense based on the largest clusters of verses with particular combinations of the past tense pivots from Papiamento (PAP), Waima (RRO), Afrikaans (ARF), Urdu (URD) and Icelandic (ISL). For each of the five languages, we present a subfigure that highlights the subset of verse clusters that are marked by the pivot of that language. The sixth subfigure highlights verses not marked by any of the five pivots.

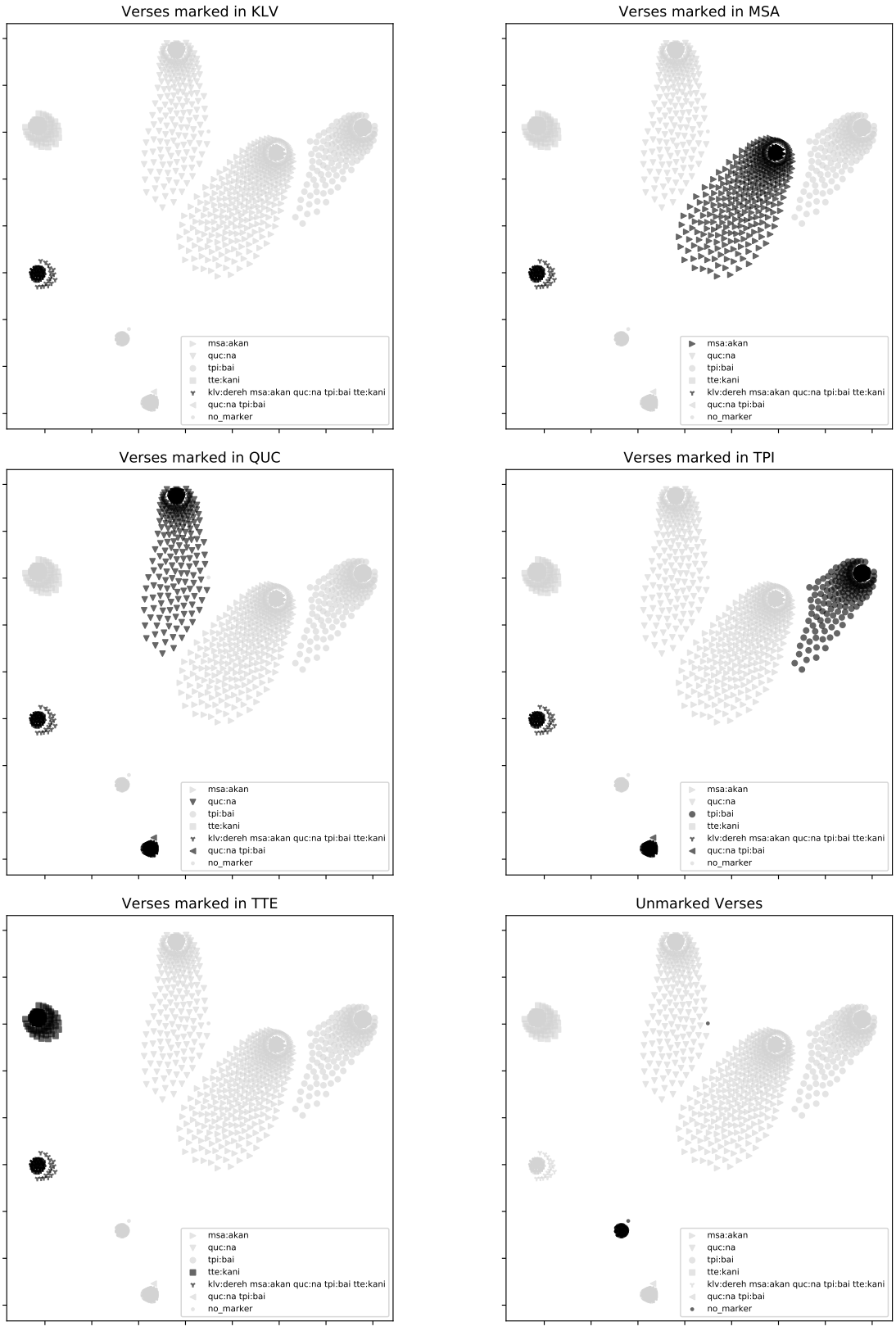

Figure 7: A map of future tense based on the largest clusters of verses with particular combinations of the past tense pivots from Bwanabwana (TTE), Tok Pisin (TPI), Quiché (QUC), Malay (MSA) and Maskelynes (KLV). For each of the five languages, we present a subfigure that highlights the subset of verse clusters that are marked by the pivot of that language. The sixth subfigure highlights verses not marked by any of the five pivots.

– the largest crosslingual computational study performed to date. We extended existing methodology for leveraging parallel corpora for typological analysis by overcoming a limiting assumption of earlier work. We only require that a linguistic feature is overtly marked in a *few* of thousands of languages as opposed to requiring that it be marked in *all* languages under investigation.

There are at least three **future directions** that seem promising to us: (i) creating a common map of tense along the lines of Figure 5, but unifying the three tenses; (ii) generalizing character n-grams to more general features, so that templates in templatic morphology, reduplication and other more complex manifestations of linguistic features can be captured; and (iii) segmenting verses into clauses and performing linear alignment not on the verse level (which caused many errors in our experiments), but on the clause level instead.

References

Judith L. Aissen. 1987. *Tzotzil Clause Structure*. Springer.

Roger W Andersen. 1990. Papiamentu tense-aspect, with special attention to discourse. *Pidgin and creole tense-mood-aspect systems* pages 59–96.

Ehsaneddin Asgari and Mohammad RK Mofrad. 2016. Comparing fifty natural languages and twelve genetic languages using word embedding language divergence (weld) as a quantitative measure of language distance. *arXiv preprint arXiv:1604.08561* .

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pages 597–604.

Emily M Bender. 2009. Linguistically naïve!= language independent: why nlp needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*. Association for Computational Linguistics, pages 26–32.

Emily M Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology* 6(3):1–26.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics, pages 69–72.

Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, pages 1455–1465.

Bernd Bohnet, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter, and Jan Hajič. 2013. Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Association for Computational Linguistics* 1:415–428.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics* 19(2):263–311.

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

George Casella and Roger L. Berger. 2008. *Statistical Inference*. Thomson.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation* 49(2):375–395.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. *arXiv preprint arXiv:1603.06147* .

Bernard Comrie. 1989. *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press.

William Croft. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press on Demand.

William Croft. 2002. *Typology and universals*. Cambridge University Press.

William Croft and Keith T Poole. 2008. Inferring universals from grammatical variation: Multidimensional scaling for typological analysis. *Theoretical linguistics* 34(1):1–37.

Michael Cysouw. 2014. Inducing semantic roles. *Perspectives on semantic roles* pages 23–68.

Michael Cysouw and Bernhard Wälchli. 2007. Parallel texts: using translational equivalents in linguistic typology. *STUF-Sprachtypologie und Universalienforschung* 60(2):95–99.

Östen Dahl. 1985. *Tense and aspect systems*. Basil Blackwell.

Östen Dahl. 2000. *Tense and Aspect in the Languages of Europe*. Walter de Gruyter.

Östen Dahl. 2007. From questionnaires to parallel corpora in typology. *STUF-Sprachtypologie und Universalienforschung* 60(2):172–181.

Östen Dahl. 2014. The perfect map: Investigating the cross-linguistic distribution of tame categories in a parallel corpus. *Aggregating Dialectology, Typology, and Register Contents Analysis. Linguistic Variation in Text and Speech. Linguae & litterae* 28:268–289.

José Guilherme Camargo de Souza and Constantin Orăsan. 2011. Can projected chains in parallel corpora help coreference resolution? In *Discourse Anaphora and Anaphor Resolution Colloquium*. Springer, pages 59–69.

Matthew S Dryer, David Gil, Bernard Comrie, Hagen Jung, Claudia Schmidt, et al. 2005. *The world atlas of language structures*. Oxford University Press.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*. pages 644–648.

Joseph H Greenberg. 1960. A quantitative approach to the morphological typology of language. *International journal of American linguistics* 26(3):178–194.

Jan Hajič, Jan Hric, and Vladislav Kuboň. 2000. Machine translation of very close languages. In *Proceedings of the sixth conference on Applied natural language processing*. Association for Computational Linguistics, pages 7–12.

Iren Hartmann, Martin Haspelmath, and Michael Cysouw. 2014. Identifying semantic role clusters and alignment types via microrole coexpression tendencies. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"* 38(3):463–484.

Serge Heiden, Sophie Prévost, Benoit Habert, Helka Folch, Serge Fleury, Gabriel Illouz, Pierre Lafon, and Julien Nioche. 2000. Typtex: Inductive typological text classification by multivariate statistical analysis for nlp systems tuning/evaluation. In *Maria Gavrilidou, George Carayannis, Stella Markantonatou, Stelios Piperidis, Gregory Stainhaouer (éds) Second International Conference on Language Resources and Evaluation*. pages p–141.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural language engineering* 11(03):311–325.

Nancy Ide. 2000. Cross-lingual sense determination: Can it work? *Computers and the Humanities* 34(1):223–234.

Stephen C Johnson. 1967. Hierarchical clustering schemes. *Psychometrika* 32(3):241–254.

Maria Koptjevskaja-Tamm, Martine Vanhove, and Peter Koch. 2007. Typological approaches to lexical semantics. *Linguistic typology* 11(1):159–185.

Anoop Kunchukuttan and Pushpak Bhattacharyya. 2016. Faster decoding for subword level phrase-based smt between related languages. *arXiv preprint arXiv:1611.00354* .

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* 9(Nov):2579–2605.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel bible corpus. *Oceania* 135(273):40.

Ryan T McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *ACL (2)*. pages 92–97.

Tony McEnery and Richard Xiao. 1999. Domains, text types, aspect marking and english-chinese translation. *Languages in Contrast* 2(2):211–229.

John H McWhorter. 2005. *Defining creole*. Oxford University Press.

Rada Mihalcea and Michel Simard. 2005. Parallel texts. *Natural Language Engineering* 11(03):239–246.

Amitabha Mukerjee, Ankit Soni, and Achla M Raina. 2006. Detecting complex predicates in hindi using pos projection across parallel corpora. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*. Association for Computational Linguistics, pages 28–35.

John Myhill and Myhill. 1992. *Typological discourse analysis: Quantitative approaches to the study of linguistic function*. Blackwell Oxford.

Lene Nordrum. 2015. Exploring spontaneous-event marking though parallel corpora: Translating english ergative intransitive constructions into norwegian and swedish. *Languages in Contrast* 15(2):230–250.

Ari Pirkola. 2001. Morphological typology of languages for ir. *Journal of Documentation* 57(3):330–348.

Philip Resnik. 2004. Exploiting hidden meanings: Using bilingual text for monolingual annotation. *Computational Linguistics and Intelligent Text Processing* pages 283–299.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1997. Creating a parallel corpus from the book of 2000 tongues. In *Proceedings of the Text Encoding Initiative 10th Anniversary User Conference (TEI-10)*. Citeseer.

Gillian Sankoff. 1990. The grammaticalization of tense and aspect in tok pisin and sranan. *Language Variation and Change* 2(03):295–312.

Marianne Elina Santaholma. 2007. Grammar sharing techniques for rule-based multilingual nlp systems. *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA)* .

Diana Santos. 2004. *Translation-based corpus studies: Contrasting English and Portuguese tense and aspect systems*. 50. Rodopi.

Devyani Sharma. 2009. Typological diversity in new englishes. *English World-Wide* 30(2):170–195.

Jae Jung Song. 2014. *Linguistic typology: Morphology and syntax*. Routledge.

Elizabeth Closs Traugott. 1978. On the expression of spatio-temporal relations in language. *Universals of human language* 3:369–400.

Bernhard Wälchli. 2010. The consonant template in synchrony and diachrony. *Baltic linguistics* 1.

Bernhard Wälchli and Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics* 50(3):671–710.

Lindsay J Whaley. 1996. *Introduction to typology: the unity and diversity of language*. Sage Publications.

RZ Xiao and AM McEnery. 2002. A corpus-based approach to tense and aspect in english-chinese translation. In *The 1st International Symposium on Contrastive and Translation Studies between Chinese and English*.