Multilingual Supervision Improves Semantic Disambiguation of Adpositions

Wesley Scivetti **Lauren Levine** **Nathan Schneider**
Georgetown University
{wss37, lel76, nathan.schneider}@georgetown.edu

Abstract

Adpositions display a remarkable amount of ambiguity and flexibility in their meanings, and are used in different ways across languages. We conduct a systematic corpus-based cross-linguistic investigation into the lexical semantics of adpositions, utilizing SNACS (Schneider et al., 2018), an annotation framework with data available in several languages. Our investigation encompasses 5 of these languages: Chinese, English, Gujarati, Hindi, and Japanese. We find substantial distributional differences in adposition semantics, even in comparable corpora. We further train classifiers to disambiguate adpositions in each of our languages. Despite the cross-linguistic differences in adpositional usage, sharing annotated data across languages boosts overall disambiguation performance, leading to the highest published scores on this task for all 5 languages.

1 Introduction

While often disregarded as mere "functional" elements of languages, adpositions in fact contain a substantial amount of inherent semantic information, and are often extended to new meanings and constructions. It is well known that speakers utilize different adpositional paradigms across languages to describe similar scenes and situations (Bowerman and Pederson, 1992; Levinson et al., 2003; Feist, 2008, *inter alia*). However, analyzing the meanings of adpositions at scale is made more difficult due to the fact that, due to considerable ambiguity, meanings of adpositions cannot be straightforwardly analyzed from unannotated corpora. Thus, there is a need for an annotation framework and meaning representation for adpositional semantics. Additionally, in order to investigate cross-linguistic differences in adpositions, this meaning representation must allow for robust application to adpositions across languages.

The SNACS framework (Schneider et al., 2018) is one such meaning representation which we argue is uniquely suited for this task. SNACS (**S**emantic **N**etwork of **A**dposition and **C**ase **S**upersenses) is a semantic framework for broad-coverage corpus annotation of high-level senses, referred to as "supersenses", for case markers and adpositions. Taking the English preposition **by** as an example, the supersense categories disambiguate whether a usage is spatial (*standing **by** the truck*), temporal (*finished **by** 8:00*), or causal (*eaten **by** moths*) (see §2 for an overview of the framework and other approaches to adpositional polysemy). SNACS is designed to be multilingual, and boasts manually annotated corpora for 9 languages thus far (§3). In addition to these datasets, the past work on English, Hindi, and Gujarati has shown promise on the task of automatic SNACS supersense classification, which could crucially aid in the future creation of larger datasets in these languages.

While the effort to create these new datasets has been substantial, there is as of yet no empirical comparison of the differences in distribution of supersenses across several languages. Here we offer such an analysis for five languages: Chinese, English, Gujarati, Hindi, and Japanese (described in §4). Additionally, we provide strong, state-of-the-art baselines for the task of monolingual supersense classification in each of the languages we investigate (§5). Moreover, we then show that the further performance gains can be achieved by sharing training data across the languages that have been annotated in the SNACS framework (§6). The success of this multilingual methodology indicates the consistency of the SNACS categorization framework across languages, despite the substantial surface differences in how these languages utilize adpositions and case.

2 Related Work

Here we give an overview of the SNACS framework (§2.1), other approaches to adpositional se-

Proceedings of the 31st International Conference on Computational Linguistics, pages 3655–3669
January 19–24, 2025.

mantics (§2.2), and previous work on disambiguation (§2.3).

2.1 The SNACS Framework

SNACS is a semantic framework for the annotation of high level senses, or supersenses, for case markings and adpositions (Schneider et al., 2018). Unlike other more general meaning representations (e.g. FrameNet (Baker et al., 1998), VerbNet (Kipper et al., 2006), PropBank (Kingsbury and Palmer, 2002), and the Prague Czech-English Dependency Treebank (Čmejrek et al., 2005)), the SNACS hierarchy is more specialized in its focus and narrow in its scope, and only attempts to describe semantic categories that are introduced by adpositions and case marking. The framework can be described as semi-coarse grained because it captures finer semantic granularity than general argument structure-based theta-roles (e.g. VerbNet), while not containing the level of detail present in Frame specific semantic roles (e.g. FrameNet) (Schneider et al., 2016).

The SNACS hierarchy consists of 55 different supersenses which are organized into 3 main branches: The CIRCUMSTANCE branch, the PARTICIPANT branch, and the CONFIGURATION branch, with each branch covering a different semantic domain.[1] The SNACS hierarchy is intended to be general enough that the same semantic categories can be applied consistently across languages.[2] The full SNACS hierarchy is shown in Figure 1 (taken from Schneider et al. 2022).

One important feature of SNACS annotation is its ability to capture extended or coerced meaning *construal* (Hwang et al., 2017), when adpositions are used to convey slightly different meanings in context than the adposition conveys as a matter of one of its core lexical meanings. Typically, this happens when a preposition's *prototypical* meaning is extended in context to a related but distinct or augmented meaning. In the SNACS hierarchy, construal is captured by giving each preposition two supersenses: a prototypical, lexically based supersense (called **Coding Function** or just **Function**) and a context-specific supersense (called the **Scene Role**). Example (1) shows a common English construal taken from the STREUSLE corpus

Figure 1: The English SNACS Hierarchy from Schneider et al. (2022), containing all SNACS supersenses (besides FOCUS, TOPICAL, and QUOTE, which are not used in English).

(Schneider et al., 2016), where the LOCUS (which describes a static location in space) is construed as the GOAL (which describes the end point of a movement event) in context.

(1) And then I put this paper **in**_GOAL↝LOCUS a drawer and lock it with a key .

In this example, the **function** is LOCUS because *in* prototypically introduces a static LOCUS, but in this context, the **scene role** of GOAL is more appropriate because the sentence is describing a dynamic event in which the *paper* ends up *in a drawer* as the endpoint of its movement.

2.2 Other Representations of Adpositional Semantics

Beyond the SNACS framework, many broad coverage meaning representations are potentially useful for the study of adpositions. As stated previously, FrameNet (Baker et al., 1998), VerbNet (Kipper et al., 2006), PropBank (Kingsbury and Palmer, 2002), and the Prague Czech-English Dependency Treebank (Čmejrek et al., 2005) all include information about the semantics of prepositions to various degrees, and with differing degrees of specificity. In most cases, the focus is on prepositional modifiers of verb phrases, though PCEDT and PropBank both provide the ability to capture modifiers of nominals as well. In fact, there is reason to believe that the functors described in PCEDT align relatively well with many SNACS supersenses, although there are some clear differences (Scivetti

[1] See Schneider et al. (2022) for a comprehensive description of the SNACS framework.

[2] The FOCUS, TOPICAL, and QUOTE supersenses are the only exceptions, and were later additions to specifically for Korean, Hindi and Japanese.

and Schneider, 2023).[3] Perhaps most similar in aim to the SNACS hierarchy is the semantic classification of prepositions outlined by The Preposition Project (Litkowski and Hargraves, 2005). This framework accomplishes similar goals as SNACS, but is generally finer-grained and has only been applied to English. In general, future work is needed to analyze the representations of adpositional semantics in all of the above frameworks across multiple languages.

2.3 Supersense Disambiguation

In English, Liu et al. (2021) present the state of the art for the task of SNACS supersense classification (also referred to as supersense disambiguation). Non-English results have only been reported alongside the introduction of the corpora for Hindi (Arora et al., 2022) and Gujarati (Mehta and Srikumar, 2023). Despite SNACS data being annotated in 9 different languages (see §3), we note that only English, Hindi, and Gujarati have any scores reported for supersense classification. Additionally, these scores all come from monolingual classifiers, so it is clear that there is a major gap in considering supersense classification from a multilingual and cross-lingual perspective.[4]

There is reason to believe that multilingual methods will be useful for boosting performance at SNACS supersense classification. Gonen and Goldberg (2016) utilize unlabeled multilingual data and a semi-supervised approach involving translation of adpositions between languages, finding that such semi-supervised signaling leads to performance gains for preposition sense disambiguation in a precurser framework to SNACS (Schneider et al., 2015). While their test sets are limited to English, their approach nevertheless demonstrates that cross-lingual signals could be generally informative for this task.

In this paper, we first present the results of SOTA monolingual classifiers for the languages of English, Hindi, Gujarati, Chinese, and Japanese in order to establish a baseline of performance for SNACS supersense classification. We then conduct multilingual experiments with data sharing between languages. The code and data used in these experiments is available on GitHub.[5]

3 Data

As noted in the previous section, the SNACS hierarchy is intended to facilitate the description of adposition and case semantics across languages, beyond just English. Indeed, thus far there have been attempts to annotate adpositions and case markers according to the SNACS guidelines in 9 different languages.[6] Non-English work on SNACS has concentrated on annotating translations of the novella *The Little Prince* (≈18k tokens) by Saint Exupéry, meaning that the available SNACS data across languages is largely parallel. English has two additional resources: STREUSLE (Schneider et al., 2016, 2018) is a >55k token corpus of the online reviews section of the English Web Treebank, and PASTRIE is a ≈22.5k token corpus of Reddit data (Kranzlein et al., 2020).

In this investigation, we focus on **English**, **Hindi**, **Gujarati**, **Chinese**, and **Japanese**. Although there are SNACS annotations of *The Little Prince* in Korean and German (full novella), and Finnish and Latin (select chapters), these resources are either annotated in an older version of the SNACS framework or are otherwise incompatible, and will thus be left out of the current analysis. Future work should aim to update these data resources to the most recent SNACSs version. Pointers to all of the available datasets for SNACS can be found on GitHub.[7]

4 Experiment 1: Distributional Analysis of Supersense Annotations across Languages

In this section, we investigate the distribution of supersense annotations across languages in parallel corpora. Specifically, we compare the distribution of preposition supersense annotations in the *The Little Prince* data of each language to that of every other language. As we have previously mentioned, for each of the 5 languages under investigation in this paper, there exists annotated SNACS data of the novella *The Little Prince*. The parallel nature

[3] PCEDT also has the benefit of having extensive parallel data between English and Czech, which could facilitate similar experiments to what we perform using SNACS.

[4] Arora (2023) also reports scores for English and Hindi, but has not yet been described in published work.

[5] https://github.com/WesScivetti/snacs/

[6] English (Schneider et al., 2016, 2018; Kranzlein et al., 2020), German (Prange and Schneider, 2021), Mandarin Chinese (Peng et al., 2020), Korean (Hwang et al., 2020), Finnish and Latin (Chen and Hulden, 2022), Hindi (Arora et al., 2022), Gujarati (Mehta and Srikumar, 2023), and Japanese (Aoyama et al., 2024)

[7] https://github.com/carmls/datasets

Language	Source	Sentences	Tokens	SNACS Tokens	SNACS Types	Unique Scene Roles	Unique Functions	Unique Construals
EN	STREUSLE	3814	55,588	5498	183	46	38	180
EN	Little Prince	1561	21,364	2051	87	47	38	124
GU	Little Prince	1483	18,516	4032	934	48	40	150
HI	Little Prince	1580	16,892	2923	116	47	38	138
JA	Little Prince	609	9,957	1801	27	49	40	134
ZH	Little Prince	1562	19,799	937	74	25	24	39

Table 1: Descriptive statistics for the datasets we use. SNACS Types refers to the number of unique lexemes which are tagged with a SNACS adpositional/case supersense. The Gujarati type counts are higher due to case markers not being segmented from their head nouns.

of this data gives us the opportunity to analyze the variation in the usage of adposition supersenses across different languages. Furthermore, this investigation allows us to estimate the potential for data sharing between various languages as a means of augmenting training data for the task of supersense classification, which will be discussed in §6.

Considering the parallel nature of the data, we may expect to see relatively similar distributions. However, upon analyzing the frequency distributions of the supersense tags, we found striking differences between different languages. Figure 2 shows the proportions of function supersense tags present in *The Little Prince* for each language. For each language, there is a long tail of rare supersenses (relative frequency <6%) which we group into a `Other` category. This covers ≈30–60% of the data in each language. In the remaining data, we see that while English has large proportions of LOCUS, GESTALT, and GOAL, Hindi and Gujarati are dominated by THEME, AGENT, and FOCUS, Japanese has large proportions of THEME and TOPICAL, and Chinese is dominated by LOCUS. The diversity of these distributions highlights the language-specific nuance of adposition senses, and also raises the question of how useful data sharing between various languages will be.

There is good reason to believe that the observed distributional differences across languages are not merely due to annotation idiosyncrasies or different applications of the SNACS framework. Past studies that introduced non-English SNACS datasets have highlighted a number of linguistic environments which give rise to such differences. For example, Aoyama et al. (2024) contrasted English and Japanese, where the Japanese と (to) is used to express manner:

(2) yukkuri-to/Manner ayashi-ta
slow-quo placate-past
"placated calmly"

While Japanese typically conveys manner using a postposition, English will more commonly convey manner using an adverb ("calmly" is more conventional than the prepositional "with calm"), thus contributing to divergence of supersense distributions. We refer to papers on the individual languages where extensive cross-lingual comparison is available: Aoyama et al. (2024) for Japanese, Arora et al. (2022) for Hindi, and Peng et al. (2020) for Chinese.

In order to more closely examine which languages are likely to useful to each other as supplemental training data for supersense classification, we quantify the distance between the supersense distributions of different languages using Jensen-Shannon (JS) distance, defined as the square root of the JS divergence, a symmetric measure of the similarity of two probability distributions:[8]

$$d_{JS}(P,Q) = \sqrt{JS(P \,||\, Q)} = \sqrt{\tfrac{1}{2}KL(P \,||\, M) + \tfrac{1}{2}KL(Q \,||\, M)}$$

where P and Q are two probability distributions, M is the mixture of the two distributions defined as $\frac{1}{2}(P+Q)$, and $KL(P \,||\, Q)$ is the Kullback–Leibler divergence (Kullback and Leibler, 1951) between P and Q. The Jensen-Shannon distance is bounded, $0 \leq d_{JS} \leq 1$, where 0 indicates that the distributions are identical, and 1 indicates that they are completely different. By applying Jensen-Shannon Distance as a metric to our data, we make the assumption that the frequency counts of the supersense annotations can be used to approximate the underlying distribution of prepositional supersenses in a language.

Figure 3 shows a heatmap of the distance scores of supersense distributions between all of the possible language pairs languages under investigation

[8] https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.jensenshannon.html

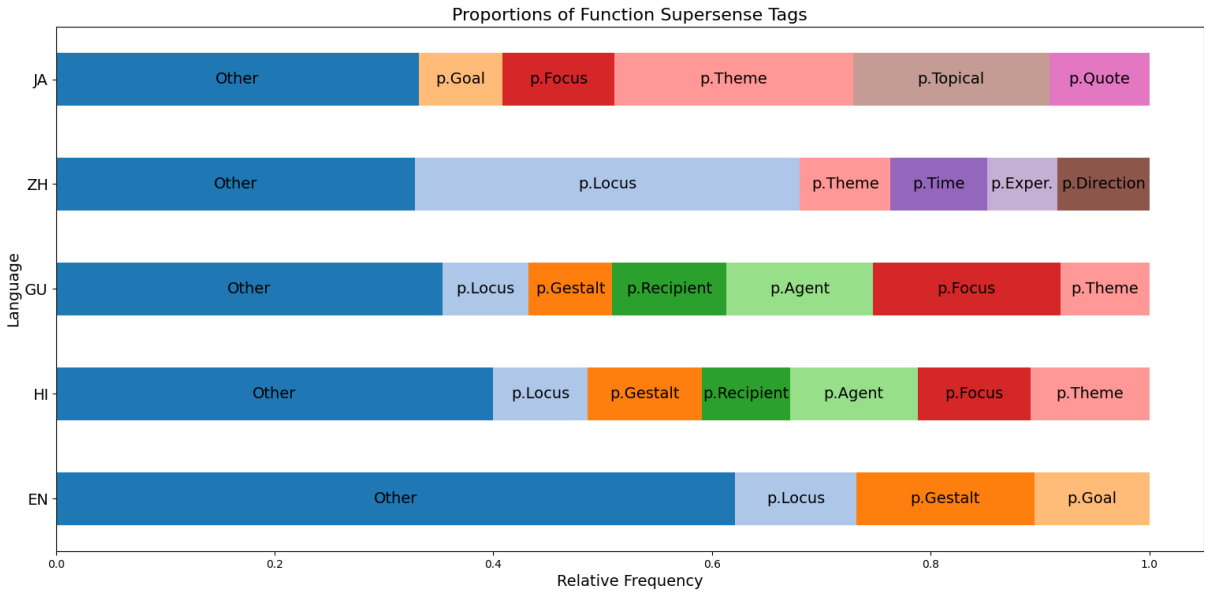

Figure 2: Proportions of function supersense tags in the original *The Little Prince* data for each language (tags with relative frequency <6% are grouped into the "Other" category).

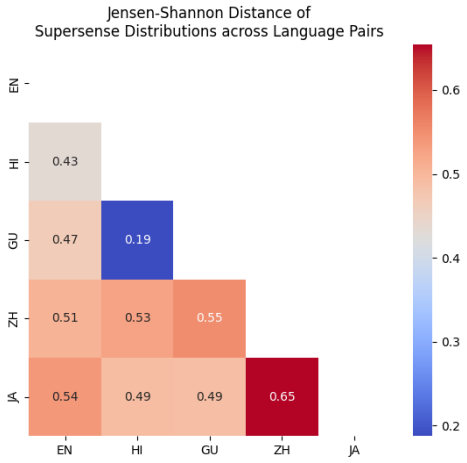

Figure 3: Jensen-Shannon Distance of supersense distributions across language pairs in the parallel corpora of *The Little Prince*.

in this study. Looking at this heatmap, we see that Hindi and Gujarati have considerably more similar supersense distributions (Jensen-Shannon Distance of 0.19) than any other language pair. As Hindi and Gujarati are from the same language family, this seems reasonable, and it suggests that they are likely to be useful data supplements for one another when training multilingual models for supersense classification. In alignment with this analysis, we choose to include data sharing between Hindi and Gujarati in our multilingual experiments in the next section.

5 Experiment 2: Monolingual Supersense Classification

In the previous section, we perform a cross-lingual analysis of the distribution of supersenses. Now, we turn to an investigation of automatic supersense classification. Before we begin our multilingual experiments, we now establish monolingual baselines for supersense classification for each of the languages under investigation.

5.1 Task Description

The goal of the supersense classification task is to assign the correct supersense label to every adposition or case-marked span in a text. While the task is a token classification task in most cases, due to the presence of some multi-word adpositions (e.g. *due to*), the task is instead formulated as a Span Classification task using a version of BIO tags, similar to Named Entity Recognition (NER) tasks.[9] It is important to note that in this task setup, target adpositions must be correctly identified in addition to being correctly classified. More lenient settings allow systems to utilize gold adposition spans before classification, leading to slightly higher scores (see Liu et al. (2021) who report scores with and without gold spans). In this work, we consider supersense classification straightforwardly as a span classification task, without providing gold or predicted span information to the models. And while

[9] See (Schneider et al., 2014) for a more complete description of the BIO tag system used in SNACS datasets.

Language	Paper	Dataset	Model	Scene Role F1	Function F1
English	Liu et al. (2021)	STREUSLE	bert-base	71.9	81.0
Hindi	Arora et al. (2022)	Little Prince	indic-transformers BERT	71.4	81.8
Gujarati	Mehta and Srikumar (2023)	Little Prince	MuRIL-large	66.2	73.2

Table 2: Previously reported monolingual scores for supersense classification. The indic-transformers BERT encoder comes from the indic-transformers library (Jain et al., 2020).

the scores for scene role and function are reported separately, we note that the models predict them as a joint tag.

For English, we use the largest corpus, STREUSLE, for all classification experiments, following the train/development/test split established in prior work (roughly an 80:10:10 split). In the other four languages, we use annotated data from *The Little Prince*. For Chinese, Hindi, and Gujarati, all 27 chapters have been annotated; we set aside chapters 1, 10, and 20 for development data and chapters 7, 17, and 27 for test data, and the remaining chapters are used for training the monolingual baseline classifiers (also roughly an 80:10:10 split). As only the first 10 chapters are available for Japanese, we set aside chapter 1 for development, chapter 7 for test, and used the remaining chapters for training. For the additional English corpora, our experiments will utilize the STREUSLE corpus, but not the PASTRIE corpus.[10] For English, we report scores on the STREUSLE test set, and for non-English languages, we report scores on the held out test portion of *The Little Prince* for that language.

As stated previously, past results on SNACS supersense classification only exist for English, Gujarati, and Hindi. The relevant published results for SNACS supersense classification are shown in Table 2.[11]

5.2 Classifier Architecture & Results

Most recent results on supersense classification (Liu et al., 2021; Arora et al., 2022; Mehta and Srikumar, 2023) feed contextualized pretrained embeddings into a biLSTM layer, then into a Conditional Random Field (CRF) layer which outputs the supersense tag probabilities. However, recent preliminary work by Arora (2023) indicates that simply fine-tuning a pretrained model (e.g. BERT (Devlin et al., 2019) with a linear token classification head) may be a competitive alternative to the LSTM+CRF architecture. We find that this is the case, and leads to competitive results with existing scores for English, Hindi, and Gujarati. We also perform a Bayesian hyperparameter search, which further boosts our results over those in past work.[12]

We tune hyperparameters on the development set to find the best model in each setting, and then report test set metrics for the best model settings in Table 3.[13] Where applicable, we included the delta from previously published scores in parentheses. As can be seen in Table 3, our best models achieves state-of-the-art performance on English STREUSLE and achieves scores which substantially excede those reported in previous work on Hindi and Gujarati, with the caveat that test splits are different from the past work on these two languages, and thus results are not directly comparable. Our classifiers also provide respectable baselines in Chinese and Japanese, where no past results have been published.

We also report the best monolingual performance when using xlm-roberta-large (XLM-R, Conneau et al., 2020) as an encoder instead of a monolingual encoder as the base of our classifier. While we expect that monolingual encoders may outperform multilingual ones in the monolingual setting, we also report the results using XLM-R as the encoder in order to be directly comparable to our later multilingual experiments, which all use XLM-R.

Overall, we see that the monolingual encoder and multilingual encoder classifiers give similar results within each respective language, and one is not clearly outperforming the other. We also see that for all languages with previously reported results, we improve both the scene role and function F1 scores substantially, with gains as high as 9.5%

[10] We do not use PASTRIE in this work due to the potential of licensing issues surrounding Reddit data.

[11] Arora (2023) also reports scores for English and Hindi, but has not yet been described in published work.

[12] Details on the hyperparameter search as well as best performing hyperparameters are reported in Appendix A.

[13] There is no official BERT-large for Chinese, so we use BERT-base. The Japanese monolingual model can be found at: https://huggingface.co/nlp-waseda/roberta-large-japanese. For Hindi and Gujarati, we do not fine-tune a monolingual model, and instead rely on MuRIL-large (Khanuja et al., 2021) as it had the highest performance out of available models in these languages.

Lang	Model	Scene Role F1	Function F1
EN	RoBERTa-large	80.7 (+8.8)	88.3 (+7.3)
	XLM-R-large	81.4 (+9.5)	88.1 (+7.1)
GU	MuRIL-large	74.4 (+8.2)	82.0 (+8.8)
	XLM-R-large	76.1 (+9.9)	84.7 (+11.5)
HI	MuRIL-large	76.1 (+4.7)	85.8 (+4.0)
	XLM-R-large	76.1 (+4.7)	84.6 (+2.8)
JA	RoBERTa-large-japanese	64.7	81.9
	XLM-R-large	62.0	81.3
ZH	BERT-base-chinese	88.6	90.8
	XLM-R-large	88.9	91.9

Table 3: F1 scores for the best-performing monolingual classifier for each language. Changes from the previously published baselines in Table 2 are shown in parentheses. English results are for the STREUSLE corpus. Note that our *Little Prince* train/dev/test splits are different from previous Hindi and Gujarati results, so deltas relative to previous reported results should be taken as ballpark figures.

for the English scene role F1 score. We hypothesize that this is due to several factors, primarily the larger encoder size in English and Hindi, as well as our extensive hyperparameter search. We also fine-tune the embeddings of our pretrained models, while past work has generally kept those embeddings static as inputs to the classification layer. We note that for all models, the performance on the function F1 is higher than for scene role F1. This is to be expected because the function of an adposition is more general and consistent, and the scene role of an adposition is more contextually variable.

Comparing the languages, Japanese had the lowest scores, which to be expected as it has the smallest amount of available training data. English has some of the highest scores, which is also expected because it has the most available training data. Surprisingly, performance for Chinese is the strongest, despite having the same amount of training data available as both Hindi and Gujarati. This strong performance on Chinese may be due to language specific factors in the distribution of supersense tags (see §4). Additionally, as shown in Table 1, Chinese has a substantially lower number of supersense tags that are used, which may be contributing to high classifier performance. This high performance also aligns with the especially high inter-annotator agreement reported in Peng et al. (2020) for the original Chinese SNACS release.

6 Experiment 3: Classification Using Multilingual Data Sharing

6.1 Task Description

Having established monolingual baselines in each of our languages, we now investigate whether data sharing between languages can lead to further performance gains. For our multilingual investigation, we conduct experiments with (1) data sharing between unrelated languages, and (2) data sharing between related languages.

In machine learning approaches to natural language processing, we often see that more data, even if it is noisy data, can boost the performance of a model (Halevy et al., 2009). This is especially true in low-resource settings, such as the supersense classification task, where each of our individual datasets is relatively small. To this end, we perform two experiments pooling data across unrelated languages. As there is substantially more annotated SNACS data in English than any other language, for the first experiment we supplement the training data of each non-English language with the English training data, evaluating on the development and test data for the non-English language. For the second experiment, we pool all of the available data together to train a single multilingual classifier which we then evaluate on the development and test data of all of the available languages.

We also experiment by sharing data specifically between the related languages Hindi and Gujarati. As the two languages are genetically similar, it is more likely that there will be a mutual benefit in sharing data in comparison to two unrelated languages, as suggested by work on transfer learning (Pan and Yang, 2010). This is supported by our distributional analysis in §4. For all of the multilingual experiments, we leverage XLM-R as our encoder, as it demonstrated strong performance even in the monolingual setting.

6.2 Results

We report the F1 scores for these data sharing experiments in Table 4. Overall, we see that data sharing across languages is useful in almost every setting. Combining all 5 language data together in the multilingual setting leads to the best performance in each of our languages (including English).

However, some of our results go against expectations. Based on §4, we expected Hindi and Gujarati to be the most useful languages to each other due to the relative similarity in their supersense distri-

butions. However, the opposite seems to be true, as both languages benefit more from the addition of English than they do from each other. This is in part likely due to the larger dataset size of English. Additionally, the Gujarati *The Little Prince* tokenization differs somewhat from the conventions in Hindi *The Little Prince*. Notably, more adpositions/case are tokenized separately in Hindi, while in Gujarati, they are left attached to the noun and then a supersense is predicted for the entire noun + case string. While adpositions and case markings do often have corresponding equivalents across the two languages, this difference in tokenization may be limiting model generalization. In all data augmentation settings, it is worth noting that adding additional language data increases the difficulty of the task by adding more attested supersenses, thus expanding the number of labels for the classification task. Despite this increase in task difficulty, the larger amount of data seems to nevertheless be helpful, especially when adding English STREUSLE as the supplemental data source.

It is also interesting that the languages that benefit the least from the additional English data are Hindi and Chinese, our two highest resource non-English languages in our datasets. It's possible that these languages are better represented in XLM-R's pretraining data, and so it already has more robust representations of Chinese and Hindi SNACS targets. Thus, adding in less relevant English data isn't as helpful. This could be seen as an example of the "curse of multilinguality" (Conneau et al., 2020), where already resource-rich languages are not benefited in the multilingual setting, while lower-resource languages (Gujarati and Japanese in our case) are better served by multilingual data augmentation. This may also work to explain the lack of reciprocal gains in data sharing between Hindi and Gujarati, though, as stated previously, tokenization differences could be playing a large part in this as well. We now turn to a deeper analysis of how multilingual data sharing effects performance on individual supersenses across languages.

6.3 Analysis

As we have seen, multilingual data sharing consistently boosts performance for each of our languages. This is interesting particularly in light of our findings in §4, which show that supersense distributions are substantially varied across languages. While we hypothesized that the more similar two languages are in their supersense distributions, the more helpful data sharing between those will be, this does not seem to be the case at least for Hindi and Gujarati. As a further example of this, English is roughly as useful for boosting performance in Japanese as it is in Hindi, even though Hindi displays a higher supersense-based Jensen-Shannon Distance with English than Japanese does.

Evaluation Language	Train Data	Scene Role F1	Function F1
EN	All langs	**82.7** (+1.3)	**90.4** (+2.3)
HI	All langs	**79.2** (+3.1)	85.3 (+0.7)
	+ EN	77.9 (+1.8)	**86.9** (+2.3)
	+ GU	76.6 (+0.5)	83.0 (−1.6)
GU	All langs	**78.3** (+2.2)	84.6 (−0.1)
	+ EN	78.0 (+1.9)	**85.3** (+0.6)
	+ HI	77.5 (+1.4)	84.2 (−0.5)
ZH	All langs	**90.1** (+1.2)	**92.3** (+0.4)
	+ EN	89.2 (+0.3)	91.5 (−0.4)
JA	All langs	**67.1** (+5.1)	**84.3** (+3.0)
	+ EN	63.9 (+1.9)	83.4 (+2.1)

Table 4: F1 scores for multilingual classifier experiments. All reported models use XLM-R as the encoder. Changes from the monolingual XLM-R encoder models in Table 3 are shown in parentheses.

Why might data sharing for unrelated languages be so useful for supersense classification? We hypothesize that it is related to the low-resource nature of this problem and the relatively sparse tagset. With the large number of supersense labels (55 in total), and the large number of construals between of scene and function roles, there are potentially hundreds of possible labels for the classifiers to discriminate between. Because of this, there is a long tail distribution of rare supersenses and construals, which are inherently difficult to predict and might occur only one or two times in the training set. In light of this fact, it's possible that unrelated languages with radically different distributions are actually helpful to one another: what's common in one language may be uncommon in another, and each language benefits from the additional training examples for a given supersense.

To test this hypothesis, we perform an analysis of the supersenses which are most benefited by data sharing. For each language, we bin the supersense labels into 4 groups, based on their frequency. We then calculate a supersense-specific F1 score, and compare the difference in this F1 in the monolingual and multilingual settings. We the average the gains for each frequency bin, and report the results in Table 5.

While there is some variation across languages, the strongest performance gains across languages

Freq Bin	EN	GU	HI	JA	ZH	AVG
Top 25%	0.8%	−4.2%	0.0%	3.3%	0.2%	0.0%
25–50%	−2.7%	0.7%	**10.6%**	10.9%	−2.2%	3.4%
50–75%	**9.3%**	9.7%	−6.7%	6.9%	**4.1%**	**4.7%**
Bot. 25%	−9.6%	**12.3%**	9.3%	**11.1%**	−6.0%	3.3%

Table 5: Average gains using data sharing across all languages, divided by frequency of the adposition supersenses. We see that the most frequent supersenses are not helped much by data augmentation, while the less frequent supersenses are helped more on average.

tend to be for the relatively rare supersense labels. In fact, for the most frequent supersenses, gains tend to be quite low or nonexistent. This shows that the overall performance gains from data sharing are primarily coming from increased performance on the long tail of rare supersenses in each language. Furthermore, this demonstrates that even if languages have radically different distributions in terms of frequency of supersenses, the supersense labels themselves are stable enough to be useful in predicting long tail examples, demonstrating the cross-lingual robustness of the SNACS framework.

7 Conclusion

In this paper, we have shown that supersenses have substantially different distributions on parallel corpus data across the 5 languages we investigate. We show that these distributions align with typological and genetic similarity in the case of Hindi and Gujarati, which have the most similar distributions and are the most closely related languages in our sample. We also show that Japanese is the most divergent from the other 4 languages.

In the monolingual classification setting, we find that it is possible to make substantial improvements over past classifier performance through a combination of hyperparameter tuning and slight tweaks to model architecture. Our current work provides strong SOTA baselines for English, Hindi, Gujarati, Mandarin Chinese, and Japanese.

In the multilingual setting, data sharing yields strong performance gains across the board, especially when data from all languages is combined together. We generally see modest improvement by adding English to other languages, especially for scene role. Regarding Gujarati and Hindi, it is interesting that data sharing did not seem to help as much as English, though it did lead to slight performance gains especially in Gujarati. While tokenization differences may have played a role in the lack of gains, it is also important to note that the amount of data sharing shared between these languages was much less than for English: we only shared the training split from *The Little Prince*, while for English the models received all of English STREUSLE as additional training data, which is roughly 3.5 times the amount of supplemental data. It seems that in this case, the size of the additional data outweighed the benefits of sharing between genetically similar languages. This is reinforced by our analysis of supersense-by-supersense gains, which show that the long tail of rare supersenses are most positively impacted by the additional data from data sharing. The difference is especially stark for Japanese (the language with the lowest amount of data). Data sharing is least helpful overall for Chinese and English, though we still observe slight performance gains.

From this work, we conclude that substantial gains in SNACS supersense classification accuracy are well within the realm of possibility. Furthermore, while languages seem to use different supersenses at radically different frequencies even when describing similar situations, we nevertheless find that the supersense categories themselves are robust enough to be useful as training data across languages. This is a positive sign for the SNACS framework, which is intended to be general enough to be cross-linguistically applicable. We hope that the promising results here inspire future work on additional languages using the SNACS framework, as well as demonstrate the general usefulness of multilingual data sharing for low-resource linguistic tasks.

Limitations

We hope to expand this analysis in the future to include all 9 languages which have SNACS data, though we stress that these 9 are all European and Asian languages, and cannot be regarded as a typological sample of world languages in general. Another limitation of the current methodology is that we do not experiment with more recent prompting-based models. Future work will investigate prompt-based experiments with more recent models, as well as fine-tuning open-source prompt-based models such as FlanT5 (Chung et al., 2022), in the hopes of boosting performance even further. Finally, we do not analyze our results in depth in terms of the individual adposition lexeme accuracies. In future work, we plan to analyze these results in more detail at a word type level, in order to

gain more insight into how data sharing effects performance on different lexemes across languages.

Acknowledgements

We thank Maitrey Mehta for helpful comments regarding the tokenization of Gujarati SNACS and its differences with Hindi. This research was supported in part by NSF award IIS-2144881.

References

Tatsuya Aoyama, Chihiro Taguchi, and Nathan Schneider. 2024. J-SNACS: Adposition and Case Supersenses for Japanese Joshi. In *The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*.

Aryaman Arora. 2023. snacs: Models for parsing SNACS datasets. https://github.com/aryamanarora/snacs.

Aryaman Arora, Nitin Venkateswaran, and Nathan Schneider. 2022. MASALA: Modelling and analysing the semantics of adpositions in linguistic annotation of Hindi. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5696–5704, Marseille, France. European Language Resources Association.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Melissa Bowerman and Eric Pederson. 1992. Crosslinguistic perspectives on topological spatial relationships. In *87th Annual Meeting of the American Anthropological Association, San Francisco, CA*.

Daniel Chen and Mans Hulden. 2022. My case, for an adposition: Lexical polysemy of adpositions and case markers in Finnish and Latin. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2610–2616, Marseille, France. European Language Resources Association.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Martin Čmejrek, Jan Cuřín, Jan Hajič, and Jiří Havelka. 2005. Prague Czech-English dependency treebank: resource for structure-based MT. In *Proceedings of the 10th EAMT Conference: Practical applications of machine translation*, Budapest, Hungary. European Association for Machine Translation.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Michele I. Feist. 2008. Space between languages. *Cognitive Science*, 32(7):1177–1199.

Hila Gonen and Yoav Goldberg. 2016. Semi supervised preposition-sense disambiguation using multilingual data. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2718–2729, Osaka, Japan. The COLING 2016 Organizing Committee.

Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE intelligent systems*, 24(2):8–12.

Jena D. Hwang, Archna Bhatia, Na-Rae Han, Tim O'Gorman, Vivek Srikumar, and Nathan Schneider. 2017. Double Trouble: The Problem of Construal in Semantic Annotation of Adpositions. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, page 178–188, Vancouver, Canada. Association for Computational Linguistics.

Jena D. Hwang, Hanwool Choe, Na-Rae Han, and Nathan Schneider. 2020. K-SNACS: Annotating Korean adposition semantics. In *Proceedings of the Second International Workshop on Designing Meaning Representations*, pages 53–66, Barcelona Spain (online). Association for Computational Linguistics.

Kushal Jain, Adwait Deshpande, Kumar Shridhar, Felix Laumann, and Ayushman Dash. 2020. Indic-Transformers: An Analysis of Transformer Language Models for Indian Languages. In *Proceedings of the NeurIPS 2020 Workshop: ML Retrospectives, Surveys & Meta-Analyses (ML-RSA)*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

Paul Kingsbury and Martha Palmer. 2002. From TreeBank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Michael Kranzlein, Emma Manning, Siyao Peng, Shira Wein, Aryaman Arora, and Nathan Schneider. 2020. PASTRIE: A corpus of prepositions annotated with supersense tags in Reddit international English. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 105–116, Barcelona, Spain. Association for Computational Linguistics.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Stephen Levinson, Sérgio Meira, The Language Group, and Cognition. 2003. "natural concepts" in the spatial topological domain-adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language*, 79(3):485–516.

Ken Litkowski and Orin Hargraves. 2005. The preposition project. In *Proc. of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, page 171–179, Colchester, Essex, UK.

Nelson F. Liu, Daniel Hershcovich, Michael Kranzlein, and Nathan Schneider. 2021. Lexical semantic recognition. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, page 49–56, Online. Association for Computational Linguistics.

Maitrey Mehta and Vivek Srikumar. 2023. Verifying annotation agreement without multiple experts: A case study with Gujarati SNACS. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10941–10958, Toronto, Canada. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Siyao Peng, Yang Liu, Yilun Zhu, Austin Blodgett, Yushi Zhao, and Nathan Schneider. 2020. A corpus of adpositional supersenses for Mandarin Chinese. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5986–5994, Marseille, France. European Language Resources Association.

Jakob Prange and Nathan Schneider. 2021. Draw *mir* a sheep: A supersense-based analysis of German case and adposition semantics. *KI-Künstliche Intelligenz*, 35(3):291–306.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014. Discriminative lexical semantic segmentation with gaps: Running the mwe gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Archna Bhatia, Na-Rae Han, Tim O'Gorman, Sarah R. Moeller, Omri Abend, Adi Shalev, Austin Blodgett, and Jakob Prange. 2022. Adposition and Case Supersenses v2.6: Guidelines for English. arXiv:1704.02134 [cs.CL].

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Meredith Green, Abhijit Suresh, Kathryn Conger, Tim O'Gorman, and Martha Palmer. 2016. A Corpus of Preposition Supersenses. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 99–109, Berlin, Germany. Association for Computational Linguistics.

Nathan Schneider, Jena D. Hwang, Vivek Srikumar, Jakob Prange, Austin Blodgett, Sarah R. Moeller, Aviram Stern, Adi Bitan, and Omri Abend. 2018. Comprehensive Supersense Disambiguation of English Prepositions and Possessives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 185–196, Melbourne, Australia. Association for Computational Linguistics.

Nathan Schneider, Vivek Srikumar, Jena D. Hwang, and Martha Palmer. 2015. A hierarchy with, of, and for preposition supersenses. In *Proceedings of the 9th Linguistic Annotation Workshop*, page 112–123, Denver, Colorado, USA. Association for Computational Linguistics.

Wesley Scivetti and Nathan Schneider. 2023. Meaning representation of English prepositional phrase roles: SNACS supersenses vs. tectogrammatical functors. In *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 68–73, Nancy, France. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing.

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Hyperparameter Search Details

For each of our models, we performed a Bayesian hyperparameter sweep using Weights and Biases (Biewald, 2020). The hyperparameters that we tuned were: learning rate, batch size, warmup steps, weight decay, and learning rate scheduler type. We tuned learning rate as a continuous variable ranging from `1e-7` to `1e-3`. Batch size was set to be either 4, 8, 16, or 24. The learning rate scheduler was either linear or cosine, and the warmup steps were a continuous variable ranging from 0 to 500. Weight decay was set to 0.0, 0.01, or 0.1. All model hyperparameters besides these were kept at their defaults from the huggingface trainer (Wolf et al., 2020). For each of our models, we allowed the hyperparameter sweep to run for a maximum of 75 runs, though some sweeps were stopped early if we did not observe any improvements after several runs. In Table 6, we report the hyperparameters for the best performing classifiers, across all languages, models, and data sharing settings. After the sweep was complete, we selected the best run from the sweep based on the F1 score on the development set. We then tested this best performing run on our held-out test set.

B Results by Supersense

In this section, we report metrics for individual supersenses for our best performing models across all settings and languages.

Language	Model	Data Sharing	Batch Size	LR	LR Scheduler	Warmup Steps	Weight Decay
Chinese	bert-base-chinese	monolingual	24	5.5e-5	cosine	49	.1
Chinese	xlm-roberta-large	monolingual	16	6.4-e5	cosine	457	0
Chinese	xlm-roberta-large	+English	8	5.7e-5	cosine	225	0
English	roberta-large	monolingual	16	4.9e-5	cosine	390	.1
English	xlm-roberta-large	monolingual	24	4.7e-5	cosine	28	.1
Gujarati	MuRIL-large	monolingual	4	3.3e-5	linear	296	.1
Gujarati	xlm-roberta-large	monolingual	16	5.0e-5	linear	244	.1
Gujarati	xlm-roberta-large	+English	16	3.8e-5	linear	65	.01
Gujarati	xlm-roberta-large	+Hindi	24	1.9e-5	cosine	82	.1
Hindi	MuRIL-large	monolingual	24	4.0e-5	linear	18	.1
Hindi	xlm-roberta-large	monolingual	24	7.3e-5	linear	442	.1
Hindi	xlm-roberta-large	+English	24	1.6e-5	linear	312	.1
Hindi	xlm-roberta-large	+Gujarati	24	1.9e-5	cosine	82	.1
Japanese	roberta-large-japanese	monolingual	16	1.0e-4	cosine	325	.1
Japanese	xlm-roberta-large	monolingual	8	4.4e-5	linear	113	.01
Japanese	xlm-roberta-large	+English	8	1.9e-5	cosine	469	.1
All	xlm-roberta-large	+All	24	2.5e-5	linear	419	0

Table 6: Hyperparameters for best models for each of our languages and data settings.

Construal	Frequency	Mono	XLM-R	XLM-R+Eng	XLM-R+All
p.Theme-p.Theme	36	0.90	0.88	0.87	0.90
p.Focus-p.Focus	25	0.67	0.73	0.80	0.70
p.Content-p.Quote	15	0.86	0.93	0.81	0.93
p.Manner-p.Manner	9	0.88	0.63	0.71	0.75
p.Experiencer-p.Topical	8	0.44	0.40	0.64	0.71
p.Topical-p.Topical	8	0.55	0.71	0.67	0.5
p.Locus-p.Locus	7	0.71	0.78	0.75	0.71
p.Theme-p.Focus	7	0.36	0.53	0.67	0.55
p.Agent-p.Topical	6	0.57	0	0	0.56
p.Beneficiary-p.Goal	6	0.67	0.8	0.83	0.29
p.Stimulus-p.Theme	6	0.80	0.55	0.71	0.71
p.Theme-p.Topical	6	0.50	0.44	0.55	0.50
p.Agent-p.Agent	5	0.80	0.80	0.73	0.73
p.Explanation-p.Circumstance	5	0	0	0.33	0
p.Gestalt-p.Topical	5	0.60	0.50	0.60	0.33
p.Goal-p.Goal	5	0.67	0.44	0.36	0.53
p.QuantityValue-p.QuantityValue	5	0.73	0.89	0.89	0.89

Table 7: F1 scores for Japanese supersense construals with frequency at least 5 in the test set. "Mono" stands for the monolingual encoder, in this case roberta-large-japanese. Scores are reported for the monolingual model, the model with additional english training data, and the model with all languages combined.

Construal	Frequency	Mono	XLM-R	XLM-R+Eng	XLM-R+All
p.Locus-p.Locus	62	0.95	0.95	0.98	0.97
p.Theme-p.Theme	13	0.96	1.00	0.89	0.96
p.ComparisonRef-p.ComparisonRef	12	0.96	0.80	0.86	0.96
p.Experiencer-p.Experiencer	8	1.00	1.00	1.00	0.94
p.Time-p.Time	7	0.71	0.92	0.86	0.86

Table 8: F1 scores for Chinese supersense construals with frequency at least 5 in the test set. "Mono" stands for the monolingual encoder, in this case bert-base-chinese. Scores are reported for the monolingual model, the model with additional english training data, and the model with all languages combined.

Construal	Frequency	Mono	XLM-R	XLM-R+Eng	XLM-R+Guj	XLM-R+All
p.Focus-p.Focus	34	0.91	0.97	0.97	0.99	0.99
p.Locus-p.Locus	26	0.79	0.82	0.85	0.87	0.82
p.Theme-p.Theme	23	0.79	0.74	0.76	0.68	0.72
p.ComparisonRef-p.ComparisonRef	16	0.85	0.90	0.88	0.81	0.71
p.Experiencer-p.Recipient	16	0.84	0.86	0.87	0.91	0.90
p.` d-p.` d	15	1.00	0.97	1.00	1.00	1.00
p.Whole-p.Whole	14	0.85	0.93	0.96	0.89	0.93
p.Originator-p.Agent	13	0.81	0.75	0.72	0.72	0.80
p.Gestalt-p.Gestalt	11	0.59	0.77	0.69	0.46	0.61
p.Stimulus-p.Theme	11	0.84	0.72	0.82	0.88	0.85
p.Topic-p.Theme	11	0.70	0.67	0.62	0.59	0.47
p.Topic-p.Topic	9	0.70	0.61	0.80	0.52	0.67
p.Agent-p.Agent	8	0.75	0.67	0.71	0.75	0.82
p.Manner-p.Manner	8	0.74	0.86	0.88	0.82	0.93
p.Possessor-p.Possessor	8	0.94	0.94	0.94	0.89	0.89
p.Extent-p.Extent	7	0.86	0.86	0.77	0.77	0.77
p.Goal-p.Locus	7	0.63	0.62	0.57	0.78	0.80
p.Source-p.Source	6	0.92	0.73	0.71	0.75	0.83
p.Characteristic-p.Characteristic	5	0.80	0.67	0.80	0.67	0.67
p.Experiencer-p.Agent	5	0.91	0.77	0.67	0.77	0.83
p.Time-p.Time	5	0.77	0.89	0.80	0.83	0.89

Table 9: F1 scores for Hindi supersense construals with frequency at least 5 in the test set. "Mono" stands for the monolingual encoder, in this case muRIL-large. Scores are reported for the monolingual model, the model with additional english training data, additional Gujarati data, and the model with all languages combined.

Construal	Frequency	Mono	XLM-R	XLM-R+Eng	XLM-R+Hi	XLM-R+All
p.Focus-p.Focus	90	0.90	0.91	0.91	0.92	0.93
p.Theme-p.Theme	34	0.74	0.74	0.67	0.73	0.63
p.Originator-p.Agent	31	0.95	0.92	0.84	0.94	0.90
p.Experiencer-p.Recipient	29	0.93	0.92	0.92	0.89	0.89
p.` -p.`	29	0.89	0.84	0.85	0.85	0.91
p.Locus-p.Locus	23	0.80	0.82	0.79	0.78	0.78
p.Whole-p.Whole	16	0.70	0.88	0.78	0.89	0.88
p.Gestalt-p.Gestalt	13	0.59	0.67	0.69	0.50	0.69
p.Goal-p.Locus	13	0.67	0.77	0.77	0.62	0.62
p.Recipient-p.Recipient	13	0.83	0.93	0.83	0.89	0.89
p.Possessor-p.Possessor	11	0.70	0.78	0.76	0.91	0.82
p.Stimulus-p.Theme	10	0.50	0.77	0.67	0.75	0.77
p.Topic-p.Topic	10	0.62	0.75	0.64	0.70	0.54
p.Characteristic-p.Identity	9	0.67	0.59	0.74	0.71	0.56
p.ComparisonRef-p.ComparisonRef	9	0.63	0.70	0.95	0.55	0.67
p.Manner-p.Manner	9	0.73	0.74	0.67	0.63	0.80
p.Agent-p.Agent	8	0.63	0.53	0.63	0.67	0.60
p.Characteristic-p.Characteristic	7	0.40	0.40	0.67	0.47	0.29
p.Circumstance-p.Circumstance	7	0.36	0.62	0.63	0.67	0.57
p.Frequency-p.Frequency	7	1.00	1.00	1.00	1.00	1.00
p.Time-p.Time	7	0.80	0.71	0.77	0.67	0.77
p.Originator-p.Gestalt	6	0.73	0.91	0.73	0.67	0.80
p.Source-p.Source	6	0.92	0.92	1.00	1.00	0.92
p.Explanation-p.Explanation	5	0.80	0.80	0.80	0.67	0.89
p.PartPortion-p.Characteristic	5	0.00	0.00	0.00	0.00	0.00
p.Purpose-p.Purpose	5	0.83	0.67	0.83	0.83	0.80

Table 10: F1 Scores for Gujarati supersense construals with frequency of at least 5. "Mono" refers to the monolingual model, in this case muRIL-large. Scores are reported for the monolingual model, the model with additional English training data, additional Hindi data, and the model with all languages combined.

Construal	Frequency	Mono	XLM-R	XLM-R+All
p.Locus-p.Locus	73	0.88	0.92	0.95
p.Possessor-p.Possessor	29	0.97	0.93	0.92
p.Purpose-p.Purpose	26	0.93	0.89	0.91
p.Time-p.Time	19	0.97	0.97	0.95
p.SocialRel-p.Gestalt	18	0.91	0.91	0.89
p.Topic-p.Topic	18	0.75	0.63	0.81
p.ComparisonRef-p.ComparisonRef	17	0.85	0.76	0.81
p.Goal-p.Goal	16	0.91	0.93	0.91
p.Gestalt-p.Gestalt	12	0.78	0.69	0.70
p.QuantityItem-p.QuantityItem	11	0.91	0.91	0.91
p.Recipient-p.Goal	11	0.91	0.87	0.87
p.Characteristic-p.Locus	9	0.50	0.67	0.57
p.Duration-p.Duration	9	0.60	0.90	0.82
p.Explanation-p.Explanation	9	0.94	0.95	1.00
p.Stimulus-p.Topic	9	0.90	0.90	0.95
p.Direction-p.Direction	8	0.57	0.63	0.53
p.OrgMember-p.Possessor	8	0.93	0.80	0.80
p.QuantityItem-p.Whole	8	0.94	0.94	0.94
p.Agent-p.Gestalt	7	0.64	0.57	0.67
p.Experiencer-p.Gestalt	7	0.71	0.71	0.71
p.Org-p.Locus	7	0.92	0.93	0.92
p.Theme-p.Theme	7	0.67	0.67	0.75
p.Characteristic-p.Characteristic	6	0.67	0.80	0.80
p.Manner-p.Manner	6	0.71	0.67	0.73
p.Circumstance-p.Circumstance	5	0.53	0.57	0.50
p.Originator-p.Source	5	0.83	0.77	0.83
p.SocialRel-p.Ancillary	5	0.75	0.89	0.73
p.StartTime-p.StartTime	5	1.00	1.00	1.00
p.Whole-p.Whole	5	0.77	0.77	0.77

Table 11: F1 Scores for English supersense construals with frequency of at least 5. "Mono" refers to the monolingual model, in this case roberta-large. Scores are reported for the monolingual models, and the model with all languages combined.