Effective QA-driven Annotation of Predicate-Argument Relations Across Languages

Jontahan Davidov[1] **Aviv Slobodkin**[1] **Shmuel Tomi Klein**[1]
Reut Tsarfaty[1] **Ido Dagan**[1] **Ayal Klein**[2]
[1]Bar-Ilan University
[2]Ariel University
yonatand58@gmail.com

Abstract

Explicit representations of predicate-argument relations form the basis of interpretable semantic analysis, supporting reasoning, generation, and evaluation. However, attaining such semantic structures requires costly annotation efforts and has remained largely confined to English. We leverage the Question-Answer driven Semantic Role Labeling (QA-SRL) framework — a natural-language formulation of predicate-argument relations — as the foundation for extending semantic annotation to new languages. To this end, we introduce a cross-linguistic projection approach that reuses an English QA-SRL parser within a constrained translation and word-alignment pipeline to automatically generate question-answer annotations aligned with target-language predicates. Applied to Hebrew, Russian, and French — spanning diverse language families — the method yields structurally rich training data and fine-tuned, language-specific parsers that outperform strong multilingual LLM baselines (GPT-4o, LLaMA-Maverick). By leveraging QA-SRL as a transferable natural-language interface for semantics, our approach enables efficient and broadly accessible predicate-argument parsing across languages.

1 Introduction

Despite the representational power of large language models (LLMs), explicit predicate-argument representations remain a cornerstone of natural language understanding. By decomposing sentences into elementary meaning units — *who did what to whom, how, when* and *where* — such structures enable fine-grained modeling of meaning and support tasks that depend on precise semantic alignment across texts, including faithfulness and attribution analysis in generation, controlled text production, and systematic evaluation of generated content (Bhattacharyya et al., 2022; Dryjański et al., 2022; Fan et al., 2023; Zhang et al., 2025).

Traditional semantic role labeling (SRL) frameworks such as FrameNet, PropBank and OntoNotes (Baker et al., 1998; Kingsbury and Palmer, 2002; Weischedel et al., 2013) provide such explicit representation of predicate-argument structure through predefined role inventories and frame lexicons — an approach also adopted by later meaning representations such as AMR, UCCA, and other frameworks (Banarescu et al., 2013; Abend and Rappoport, 2013; Oepen et al., 2015). However, these linguistically grounded schemas rely on expert annotation and language-specific resources, making large-scale annotation costly and, consequently, largely limited to English. As a result, cross-lingual semantic annotation has progressed slowly and unevenly, despite the conceptual centrality of predicate-argument representations in NLP.

Question-Answer driven SRL (QA-SRL) (He et al., 2015) introduced a natural-language alternative: representing predicate-argument relations through question-answer pairs rather than symbolic role labels. Each predicate is associated with simple questions (e.g., *Who* ***shot*** *someone?*) and their corresponding answer spans, capturing the underlying roles without schema-specific design (see Table 1 for a full example). This formulation yields an interpretable, annotation-friendly, and LLM-compatible representation of semantic structure. Subsequent work extended this approach to deverbal nominalizations (QANom; Klein et al., 2020) as well as to a broader QA-based semantic framework (Klein et al., 2022) encompassing additional types of predication (Pyatkin et al., 2020; Pesahov et al., 2023). In practice, QA-SRL has been shown to be effective across multiple downstream tasks (Brook Weiss et al., 2021; Sultan and Shahaf, 2022; Caciularu et al., 2023; Cattan et al., 2024; Zhang et al., 2025), underscoring its role

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2484–2502
March 24-29, 2026 ©2026 Association for Computational Linguistics

Both were shot in the confrontation with police and have been recovering in hospital since the massive attack.			
QA-SRL	1	When was someone shot?	in the confrontation ; the attack
	2	Who was shot?	Both
	3	Who shot someone?	police
	4	Where has someone been recovering?	in hospital
	5	How long was someone recovering from something?	since the attack
	6	Who was recovering from something?	Both
	7	What was someone recovering from?	shot
QANom	8	Who confronted with something?	Both
	9	What did someone confront with?	police

Table 1: An example sentence annotated with QA-SRL and QANom.

as a broadly applicable representation of predicate-argument structure.

However, existing QA-SRL resources have so far been developed exclusively for English. To realize its broader potential, we seek to leverage QA-SRL's natural-language format as a vehicle for scaling predicate-argument annotations to new languages with minimal language-specific prerequisites. Achieving this requires a systematic, cross-lingual extension that preserves QA-SRL's accessibility while maintaining the structured correspondence between predicates, questions, and answer spans that underpins its semantic fidelity.

In this work, we propose a multilingual QA-SRL projection approach that fulfills this goal. Our algorithm reuses a high-quality English QA-SRL parser within a refined projection pipeline that translates the English question-answer annotations into the target language. It combines constrained machine translation, word alignment, and QA-structure preservation mechanisms specifically tailored to QA-SRL's semi-structured format. The resulting projected annotations are used to supervise the fine-tuning of lightweight, language-specific QA-SRL parsers, making automatic predicate-argument analysis efficiently attainable and widely accessible across languages. To ground the task and illustrate the expected outputs across languages, Table 2 provides concrete examples of QA-SRL predictions for the same predicate in Hebrew, Russian, and French.

We validate this approach on three typologically diverse languages — Hebrew, Russian, and French — spanning Semitic, Slavic, and Romance families. For each, we construct projected verbal and nominal QA-SRL datasets, curate gold evaluation subsets, and train language-specific QA parsers that substantially outperform strong multilingual LLM baselines such as GPT-4o and LLaMA-Maverick.

Taken together, our results establish a scalable methodology for extending predicate-argument semantics to new languages. By operationalizing multilingual QA-SRL as a projection-based process, we demonstrate how explicit predicate-argument representations — long central to semantic analysis — can be scaled cross-linguistically with minimal cost, leveraging natural language itself as the medium of meaning transfer.

2 Background and Related Work

This section situates our work in two strands of prior research: efforts to build multilingual semantic resources, and the development of QA-based representations.

2.1 Multilingual Semantic Resources

Large-scale semantic resources remain concentrated in English. While syntactic treebanks such as Universal Dependencies provide broad coverage (Nivre et al., 2016; de Marneffe et al., 2021), comparable semantic annotations such as predicate-argument relations are scarce beyond English, limiting interpretable, content-grounded modeling in most languages.

Cross-lingual SRL has relied on translating English corpora and projecting PropBank-style roles via alignments, or on inducing training data from parallel corpora. A prominent example is Universal Proposition Bank (UPB), which constructs multilingual propbanks through a two-stage pipeline combining monolingual SRL with multilingual parallel data, and has recently been extended and improved in UP2.0 (Jindal et al., 2022). Beyond projection, another core obstacle is *formalism heterogeneity*: different languages adopt different SRL inventories (e.g., PropBank vs. AnCora vs. PDT-Vallex), making cross-lingual transfer de-

Language	Sentence	Question	Answer	English gloss
Hebrew	הוועדה **אישרה** את המדיניות החדשה	מי אישר משהו?	הוועדה	Who approved something? → The committee
		מה מישהו אישר?	את המדיניות החדשה	What did someone approve? → the new policy
Russian	Комитет **одобрил** новую политику	Кто что одобрил?	Комитет	Who approved something? → The committee
		что кто-то одобрил?	новую политику	What did someone approve? → the new policy
French	Le comité a **approuvé** la nouvelle politique	qui a approuvé quelque chose?	Le comité	Who approved something? → The committee
		qu'est-ce que quelqu'un a approuvé?	la nouvelle politique	What did someone approve? → the new policy

Table 2: Examples of model predictions in different target languages.

pend on schema mapping. Conia et al. (2021) address this by training a unified model over heterogeneous SRL resources that implicitly learns cross-inventory correspondences without relying on word alignment or translation. In parallel, lexical-semantic resources such as VerbAtlas define cross-frame semantic roles and prototypical argument structures, offering an alternative to purely predicate-specific PropBank role sets and supporting more uniform role semantics (Di Fabio et al., 2019).

Alongside these efforts, recent cross-lingual approaches include translated training data (Fei et al., 2020), X-SRL's multilingual projection (Daza and Frank, 2020), alignment-free modeling (Cai and Lapata, 2020), and divergence-aware corrections (Youm et al., 2024). Similar strategies underlie AMR transfer: contextual word alignments (Sheth et al., 2021) or translate-then-parse baselines (Uhrig et al., 2021) achieve strong results, but again require adaptation to formalism-specific schemata. Closer to our direction is CLaP (Parekh et al., 2024), which improves projection by conditioning label translation on context and label semantics rather than word-level alignments, yet still operates within a schema-based SRL framework. Recent surveys stress that reliance on English-centric linguistic formalisms, schema mapping, and costly alignment pipelines constrains scalability across languages (Hämmerl et al., 2024).

These limitations motivate exploring natural-language-based predicate-argument representations, which can mitigate reliance on rigid role-schema mappings and enable more scalable, cost-effective transfer across languages.

2.2 QA-based Semantic Role Labeling

QA-driven semantic role labeling offers an alternative to schema-based SRL by representing predicate–argument relations directly in natural language, rather than through fixed role inventories. As illustrated in Table 1, each predicate is queried with simple questions capturing its roles, and the answers mark the corresponding argument spans. This formulation eliminates the need for predefined role sets or cross-schema mapping, enables intuitive crowdsourced annotation, and naturally aligns with the capabilities of modern language models (He et al., 2015).

The QA-SRL framework was extended to include deverbal nominalizations (Klein et al., 2020), preserving the same question formats and capturing eventive structure in nominal domains. Verbal and nominal QA-SRL have since been modeled jointly through a text-to-text framework (Klein et al., 2022), with subsequent advances yielding a state-of-the-art English parser (Cattan et al., 2024, see Appendix A for details). This parser provides the English annotations that we project in our cross-lingual pipeline.

While this work focus on verbal and nominal QA-SRL, which encode the core propositional content of sentences, other extensions of the QA-based semantic paradigm cover adjectival predicates (Pesahov et al., 2023), discourse relations (Pyatkin et al., 2020) and additional noun-related semantics (Tseytlin et al., 2025). We leave the cross-linguistic extension of these for future exploration.

Recent studies have demonstrated the utility of QA-SRL across a range of downstream applications, including cross-text predicate-argument alignment (Brook Weiss et al., 2021), fine-grained summarization evaluation (Zhang et al., 2025), localization of factuality assessment (Cattan et al., 2024), event similarity in analogical reasoning (Sultan and Shahaf, 2022), and multi-document pre-training (Caciularu et al., 2023). These applications establish QA-SRL as a general-purpose, interpretable content representation at the fine-grained level of predicate-argument relations. Yet crucially, all prior work has been limited to English. Our contribution is to address this gap by scaling QA-SRL to new languages through a cross-lingual translation pipeline.

3 Multilingual QA-SRL Projection Algorithm

3.1 Overview

Goal & Motivation Our goal is to demonstrate a general methodology for attaining predicate-argument representations automatically and at low cost in new languages. Concretely, our methodology produces large-scale, high-quality QA-SRL training datasets in a target language $\mathcal{L}$ for fine-tuning a language-specific QA-SRL parser. Our proposed projection pipeline, introduced in the current section, takes a target-language corpus and automatically generates question-answer annotations that serve as supervision for training these parsers, making the trained models the ultimate output of the process.

In contrast to traditional SRL projection approaches that transfer schema-bound role labels, our method projects natural language question-answer pairs, avoiding the need for role mapping and leveraging core translation capabilities. On the other hand, we observe that a naive English back-translation approach for attaining QA-SRL QAs in $\mathcal{L}$ yields inadequate QA-SRL outputs (see Appendix B for a detailed discussion), which motivates our more refined projection approach.

Target Language Requirements Our projection pipeline is applicable in any target language $\mathcal{L}$ with the following broadly available resource requirements: (i) a reasonably sized corpus with POS annotations to identify verbal and candidate nominal predicates; (ii) machine translation (MT) from $\mathcal{L}$ to English; (iii) word alignment tools; and (iv) a pretrained language model capable of simple in-context operations — primarily constrained translation (§3.3). In practice, these prerequisites are already satisfied for most commonly used languages, as contemporary multilingual resources and models provide dependable coverage for well over a hundred languages, encompassing the vast majority of the world's population.[1]

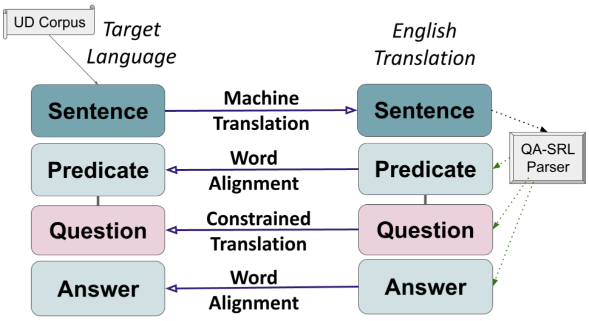

Figure 1: A schematic overview of our proposed methodology for transferring QA-SRL to new languages based on the English QA-SRL infrastructure.

Projection Pipeline Outline The pipeline operates in four stages (Figure 1):

1. **Sentence translation:** Starting from a tokenized, part-of-speech tagged corpus in $\mathcal{L}$, each sentence is translated into English.[2]
2. **English-side QA-SRL parsing:** The translated English sentence is passed through the QA-SRL parser to generate English question-answer pairs anchored to verbal and nominal predicates.
3. **Word alignment:** Using *word alignment* (§3.2) between the English translation and the original target-language sentence, we map English predicates to their target-language counterparts and project English answer spans onto $\mathcal{L}$.
4. **Constrained back-translation:** Generated English questions are translated back into $\mathcal{L}$ using a *predicate-preserving constrained translation* procedure (§3.3), ensuring the aligned predicate appears verbatim in the question.

An illustrative complete example of the projec-

[1] The Universal Dependencies project (UD) currently provides treebanks for over 150 languages (Nivre et al., 2020). Major parallel-data and translation resources such as OPUS (Tiedemann, 2022), NLLB (Costa-Jussà et al., 2022), and SeamlessM4T (Barrault et al., 2023) collectively cover between 100 and 200 languages, while large-scale MT systems such as Google Translate now support around 250 languages (see Wikipedia). Modern multilingual LLMs (e.g., GPT-4o) also exhibit strong zero-shot cross-lingual competence, further relaxing the need for language-specific models and annotations.

[2] We used open-source translation models from the Helsinki-NLP project for translating sentences from the target language into English: `opus-mt-tc-big-he-en` for Hebrew, `opus-mt-ru-en` for Russian, and `opus-mt-tc-big-fr-en` for French.

tion pipeline can be found in Appendix C. This process yields parallel QA-SRL annotations (sentence, predicate, question, answer) for any language with a basic linguistic resources and a translation system, enabling the creation of large-scale training sets for multilingual QA-SRL parsers.

3.2 Word Alignment

Word alignment links predicates and arguments between the English translation and the target-language sentence. We employ SimAlign (Jalili Sabet et al., 2020), an unsupervised mBERT-based method that frames alignment as bipartite graph matching. The alignment is used at two key modules:

1. **Predicate alignment:** For each English predicate captured by the QA-SRL parser, we locate its aligned token(s) in the target-language sentence. This keeps the scope of captured predicates roughly equivalent to the English parser's scope, covering most verbs and deverbal nominalizations in $\mathcal{L}$ (see Appendix D for full details regarding predicate identification).
2. **Answer span projection:** Each English answer span is mapped to the original sentence in $\mathcal{L}$ by locating the aligned tokens and extracting the smallest continuous span that contains them (see Appendix E for further details about the answer projection heuristics).

3.3 Predicate Preserving Constrained Translation

After generating English QA pairs with the QA-SRL parser and aligning the target-language predicate, we translate the questions into $\mathcal{L}$ while enforcing that the translated question explicitly contains the aligned predicate, preserving the QA-SRL format. We implement this using language-specific LLMs prompted with few-shot examples instructing the model to produce fluent translations of English questions confined to particular $\mathcal{L}$ predicates. This procedure is applied uniformly to verbal (QA-SRL) and nominal (QANom) predicates. An illustrative example is provided in Appendix F.

The resulting projected annotations form the basis for the multilingual datasets described in Section 4, which in turn are used to fine-tune and evaluate target-language QA-SRL parsers presented in Section 5.

4 Dataset Creation

To train and evaluate multilingual QA-SRL models, we construct automatically projected QA-SRL datasets for three languages and create a manually curated gold-standard evaluation set. This section describes the target languages, source corpora, and annotation pipeline used to build these resources.

4.1 Languages and Corpora

To evaluate our projection pipeline across diverse linguistic settings, we target three typologically distinct languages: Hebrew, Russian, and French. This selection stresses the method under varying structural and resource conditions — Hebrew as a medium-to-low resource, Semitic language with rich morphology and a revived modern profile; Russian with its flexible word order and extensive case system; and French as a resource-rich Romance language providing structural contrast.

For each language, we extract sentences from multiple Universal Dependencies (UD) corpora (Nivre et al., 2016; de Marneffe et al., 2021), which provide tokenization and part-of-speech tags essential for predicate identification (see Appendix D). Specifically, we use *HTB*, *IAHLTwiki*, and *IAHLTknesset* for Hebrew; *SynTagRus*, *GSD*, *Taiga*, and *PUD* for Russian; and *GSD*, *Sequoia*, *ParisStories*, and *PUD* for French. These corpora span diverse genres including news, Wikipedia, government proceedings, blogs, and spoken narratives, providing a broad linguistic sample for each language.

We apply our QA-SRL projection pipeline on the compiled corpora to attain linguistic specific QA-SRL annotations. Table 3 presents descriptive statistics of our projected datasets.

4.2 Evaluation set annotation

To evaluate projected QA-SRL annotations, we created a gold-standard dataset through manual human annotation. This served two purposes: (1) providing a reliable benchmark for QA-SRL models in the target languages, and (2) assessing the accuracy of our cross-lingual projection pipeline.

We sampled projected QA pairs from multiple corpora in each language and reviewed them via a multi-step manual process. Annotators corrected substantive errors (e.g., incorrect or missing questions/answers), removed invalid pairs caused by translation or alignment failures, and added missing but valid QAs. Minor stylistic edits were ap-

	Train			Dev			Test		
Language	**Sents**	**Pred.**	**QAs**	**Sents**	**Pred.**	**QAs**	**Sents**	**Pred.**	**QAs**
Hebrew	13,233	33,275	80,210	64	146	368	116	342	793
Russian	18,804	42,518	102,197	98	258	689	228	603	1,582
French	18,181	38,157	97,336	51	124	259	144	288	616

Table 3: Number of sentences, predicates, and QA pairs in the projected QA-SRL datasets for Hebrew, Russian, and French.

plied for clarity, while all substantial modifications (e.g., replacing or adding QA pairs) were explicitly flagged for later evaluation of the projection algorithm.

Hebrew and French datasets were annotated by the authors, while Russian annotation was carried out by a native speaker experienced in QA-SRL tasks, compensated at $15/hour. The average annotation rate was 90 QA pairs per hour. All annotators ensured grammaticality, faithfulness to the source sentence, and full coverage of the predicate's semantic arguments.

4.3 Quality Assessment of Data

4.3.1 Evaluation Criteria

We evaluate QA-SRL annotations using standard SRL metrics adapted to the QA-based format, focusing on two subtasks: **Unlabeled Argument Detection** and **Labeled Argument Detection**, corresponding to *Argument Detection* and *Label Assignment* in traditional SRL. These metrics have been previously adopted in QA-SRL and QANom evaluations (Roit et al., 2020; Klein et al., 2020) and are instantiated in our automatic setup (§5.1) as Argument Match and Question Match, respectively.

Unlabeled Argument Detection checks whether the predicted answer span covers a valid argument regardless of question phrasing, while Labeled Argument Detection additionally requires matching the role assignment conveyed by the gold question.

4.3.2 Evaluating the Training Set

Table 4 reports Precision, Recall, and F1 for Hebrew, Russian, and French. High Unlabeled F1 across all languages shows that the projection pipeline captures most predicate arguments, even in morphologically rich, structurally diverse settings. Since we focus on generating full QA sets for each predicate, this demonstrates effective reconstruction of core argument structures.

Recall remains high in both settings, indicating broad coverage of arguments and questions — critical for training, where missing information is more harmful than occasional noise. As expected, Labeled scores are lower due to stricter matching criteria but remain strong enough to support reliable supervision.

	Unlabeled AD			**Labeled AD**		
Language	**P**	**R**	**F1**	**P**	**R**	**F1**
Hebrew	67.8	94.7	79.0	57.4	93.8	71.2
Russian	70.3	87.9	78.1	47.3	83.3	60.3
French	81.3	93.5	87.0	60.8	91.6	73.1

Table 4: Manual evaluation results of the projected training sets.

By combining UD corpora, MT, word alignment, and constrained question translation, we built high-quality Hebrew, Russian, and French training data. Manual evaluation confirmed robust predicate-argument coverage, forming a solid foundation for developing multilingual QA-SRL parsers.

Error sources in the projection pipeline. To better understand the quality of the projected training data, we conducted a targeted analysis of error sources across the projection pipeline. Overall, we find that most errors originate from the English QA-SRL/QANom parser itself, which serves as an effective upper bound on projection quality and is known to be recall-limited (Roit et al., 2020; Klein et al., 2022). In contrast, errors introduced during back-projection are comparatively rare. Predicate alignment errors occur infrequently (4 misaligned predicates out of 161 examined in the manually curated gold set), and predicate-preserving constrained translation achieves a high success rate, effectively mitigating predicate drift. Remaining back-projection errors primarily involve minor answer-span boundary inaccuracies due to alignment gaps, rather than systematic semantic distortions. Taken together, this analysis indicates that the projected datasets preserve predicate–argument structure reliably, and that further gains in projection quality will depend primarily on advances in English QA-SRL parsing rather than on the cross-lingual transfer mechanisms themselves.

5 Models and Evaluation

In this section, we evaluate the quality of our QA-based semantic parsers across three target languages. We first describe the automatic evaluation framework (§5.1), followed by the experimental setup and models used (§5.2), and conclude with the main results and analysis (§5.3).

5.1 Automatic Evaluation

To enable large-scale evaluation of model predictions, we significantly adapt the standard protocol from prior QA-SRL work (Roit et al., 2020; Klein et al., 2020). We operationalize the two subtasks defined in Section 4 via a two-stage automatic procedure: **Argument Match**, aligning predicted and gold answer spans, and **Question Match**, assessing whether the predicted question expresses the same semantic role as the reference.

For Argument Match, we adopt the token-level alignment method of Roit et al. (2020). Predicted and gold answer spans are connected in a bipartite graph weighted by token-level *Intersection-over-Union (IOU)*; edges below a calibrated threshold $\tau = 0.5$ are discarded. Then, a maximal bipartite matching algorithm selects the best one-to-one alignments between predicted and gold spans. Spans that align above the threshold are treated as matches, while the rest are considered mismatches. Threshold tuning and further details are elaborated in Appendix G.

In the Question Match stage, we impose a stricter QA-to-QA match criterion and evaluate the questions associated with matched answer spans, which implicitly encode the argument's semantic role. Prior QA-SRL work has struggled to establish reliable question-level metrics, with existing approaches tied to English-specific templates and ill-suited for multilingual, free-form questions.

To address this, we introduce two complementary language-agnostic **Question Match** criteria: (1) **Exact Match**, requiring string identity with the gold question, and (2) **Semantic Match**, which scores a question as correct if its embedding is semantically equivalent to the gold according to a SentenceTransformers paraphrase model (cosine similarity ≥ 0.78). Appendix H details the model and threshold calibration.

This two-stage evaluation combines strict and relaxed signals for assessing predicate-argument structure, providing a scalable approximation to human judgment across languages.

5.2 Experimental Setup and Models

To assess the effects of scale, supervision, and language specialization, we evaluate three model families: (i) **instruction-tuned LLMs** (hundreds of billions of parameters) serving as few-shot no-projection baselines; (ii) a **multilingual 8B language model**, evaluated both in its in-context learning (ICL) mode and after LoRA fine-tuning (FT); and (iii) **language-specific 7B models**, likewise evaluated before and after fine-tuning.

This design enables us to isolate the effects of fine-tuning and to compare multilingual versus monolingual pre-training at a similar scale.

We employ **GPT-4o** (OpenAI, 2024) and **LLaMA-4-Maverick** (Meta, 2025) as few-shot instruction-tuned baselines, prompted with two language-specific examples for each predicate type (full prompts are provided in Appendix I). These LLMs, while requiring substantially more inference-time computation, provide a strong reference point without any task- or language-specific adaptation.

At the 8B scale, we employ **LLaMA-3-8B** (AI@Meta, 2024) as our multilingual backbone and compare it directly with monolingual language models of similar size. For both types, we evaluate (a) their in-context performance and (b) their LoRA-adapted variants, allowing us to examine how fine-tuning on projected data affects performance under a consistent architecture and parameter budget.

For the language-specific models, we use **DictaLM** for Hebrew (Shmidman et al., 2024), **SambaLingo** for Russian (Csaki et al., 2024), and **Claire** for French (Louradour et al., 2024). Their single-language pretraining provides specialized syntactic and lexical priors that we hypothesize to enhance QA-based semantic parsing. All models are LoRA finetuned on the our automatically projected datasets (§4), with hyperparameters selected on held-out development sets (full details in Appendix J).

To illustrate the task and expected outputs, Table 2 presents model-generated QA pairs for the sentence “The committee **approved** the new policy”, translated into Hebrew, Russian and French. We next present the evaluation results, comparing performance across languages and model types.

5.3 Results

Table 5 reports performance across Hebrew, Russian, and French for all model types, evaluated with Unlabeled Match, Exact Match, and Semantic Match.[3]

Large instruction-tuned models, particularly GPT-4o, achieve strong scores on Unlabeled Match in all languages, confirming their ability to detect salient argument spans without task-specific supervision. In Hebrew and Russian, their performance approaches or slightly exceeds that of fine-tuned models, though the gap is small. In French, however, the language-specific fine-tuned model surpasses all baselines, showing that targeted supervision can still improve even surface-level argument detection with a much smaller model.

For question-based metrics, the contrast is sharper: fine-tuned models, especially language-specific ones, consistently outperform few-shot prompting by a wide margin on both Exact and Semantic Match. This pattern holds across all languages, underscoring that while large instruction-tuned models can approximate argument boundaries, generating semantically appropriate questions benefits substantially from explicit supervision. Given the crucial role of question precision in conveying the argument's semantic role, these results highlight the limitations of in-context learning and reinforce the importance of our training-data projection pipeline in enabling cross-linguistic supervision. We note that Semantic Match is a deliberately conservative labeled metric, calibrated for high precision and therefore prone to false negatives (Appendix H); as a result, labeled argument performance is systematically underestimated by this metric.

Among fine-tuned systems, language-specific models perform best overall. Their consistent advantage over the multilingual LLaMA-3-8B variant underscores the value of language specialization, particularly for producing precise, semantically-oriented annotations. Overall, these findings highlight that language-aligned models are key to achieving robust QA-based semantic parsing across diverse languages.

Statistical significance. To assess whether the relatively small manually curated test sets affect the robustness of the observed trends, we conducted 10,000-iteration paired bootstrap significance tests by resampling predicates and recomputing F1 scores from their true positive, false positive, and false negative contributions (approximately 300–550 predicates per language). For Semantic Match F1, the advantage of the language-specific fine-tuned models over GPT-4o in-context learning is statistically significant in all three languages ($p < 0.001$). In contrast, for Unlabeled Match, differences between systems are not statistically significant in any language ($p = 0.08$–0.17), supporting the conclusion that the compared approaches yield broadly similar argument detection performance while diverging substantially in question quality.

Sensitivity to the Semantic Match threshold. To assess the sensitivity of our comparative results to the similarity threshold used by the Semantic Match metric, we re-evaluated all models across all languages using thresholds ranging from 0.70 to 0.90. As expected, absolute Semantic Match scores decrease gradually as the threshold becomes stricter. Crucially, however, the relative performance pattern remains stable throughout this range: the performance gap between fine-tuned models and GPT-4o is essentially unchanged for all three languages. This analysis indicates that while the automatic metric exhibits moderate sensitivity to threshold choice in absolute terms, our comparative conclusions regarding model performance are highly robust to this hyperparameter.

Manual error analysis of question generation. To better characterize labeled evaluation errors, we manually analyzed cases in which predicted answers passed the IOU-based Argument Match stage but failed Semantic Match. Across 50 such instances from the Hebrew test set, approximately half were semantically acceptable, corresponding to valid paraphrases or legitimate alternate questions induced by minor span differences. The remaining cases reflect genuine modeling errors, primarily involving incorrect predicate realization or role targeting. Overall, this analysis indicates that Semantic Match failures often reflect conservative matching rather than spurious predictions, leading to systematic underestimation of labeled performance. Detailed taxonomy and examples appear in Appendix K.

[3] For reference, the English QA-SRL parser used to generate the projected annotations reports 75.9 F1 on QA-SRL and 72.4 F1 on QANom under unlabeled argument detection (Cattan et al., 2024). These values provide a loose upper bound on parsers trained over annotations projected from predicted English QA-SRL/QANom.

Model	Metric	Hebrew	Russian	French
GPT-4o (ICL)	Unlabeled Match	**66.3**	**67.9**	63.2
	Exact Match	10.5	14.1	10.2
	Semantic Match	38.3	47.0	41.8
LLaMA-4-Maverick (ICL)	Unlabeled Match	61.6	60.1	52.6
	Exact Match	12.0	12.8	7.5
	Semantic Match	38.8	45.0	30.7
LLaMA-3-8B (ICL)	Unlabeled Match	43.3	51.8	47.8
	Exact Match	2.0	4.5	3.5
	Semantic Match	20.4	36.5	26.7
LLaMA-3-8B (FT)	Unlabeled Match	56.5	61.7	60.5
	Exact Match	21.2	19.2	19.1
	Semantic Match	42.7	53.1	42.6
Language-Specific (ICL)	Unlabeled Match	51.6	42.3	32.5
	Exact Match	8.1	4.5	2.9
	Semantic Match	31.9	31.5	16.4
Language-Specific (FT)	Unlabeled Match	63.0	65.7	**65.9**
	Exact Match	**30.1**	**25.4**	**23.1**
	Semantic Match	**51.7**	**58.1**	**57.2**

Table 5: F1 scores of baseline and fine-tuned models across Hebrew, Russian, and French. Double horizontal rules delineate three blocks: (1) The upper block lists strong large-scale baseline systems (GPT-4o and LLaMA-4-Maverick). (2) The middle block reports a mid-scale multilingual model before and after fine-tuning on our projected dataset. (3) The lower block presents language-specific models that are comparable in scale to the middle block, comparing their in-context and fine-tuned variants per language.

In sum, few-shot prompting with large general-purpose models remains competitive for identifying argument spans, but consistently falls short in generating semantically faithful questions. Across languages, evaluation metrics, significance testing, and threshold sensitivity analyses all converge on the same conclusion: fine-tuned models trained on language-specific projected data deliver substantially higher-quality semantic role representations. Notably, these gains are achieved by models that are orders of magnitude smaller than the instruction-tuned LLM baselines, reinforcing our central claim that targeted supervision and language adaptation remain essential for accurate, interpretable, and efficient QA-based semantic parsing in multilingual settings.

6 Conclusion

We introduced a scalable projection approach for producing QA-based predicate-argument annotations in new languages. By reusing an English QA-SRL parser within a pipeline of constrained translation and word alignment, our method generates high-quality training data for structurally diverse languages — demonstrated on Hebrew, Russian, and French. Fine-tuned models trained on this projected supervision achieve strong, language-specific predicate-argument parsing performance, approaching English accuracy on unlabeled argument detection and substantially outperforming few-shot prompted LLMs in question generation accuracy. These findings underscore the enduring value of explicit, task-specific supervision even in the era of powerful general-purpose models.

Beyond empirical results, our contribution lies in making QA-based predicate-argument analysis broadly attainable across languages with minimal human annotation. By turning QA-SRL into a practical projection vehicle, this work extends the reach of interpretable, semi-structured predicate-argument representations to low- and mid-resource languages — laying the foundation for multilingual applications such as fine-grained semantic analysis and content-based generation evaluation.

7 Limitations

Our core methodological contribution — the QA-SRL projection pipeline — relies on a sequence of model-dependent components, including translation, English QA-SRL parsing, word alignment, and constrained question translation. While this modular design enables scalable annotation transfer, it also introduces susceptibility to error propagation: inaccuracies at any stage (e.g., translation drift, parser errors, or alignment mismatches) can cascade and compound in the final projected annotations. As a result, the quality of the generated datasets and trained parsers is tightly linked to the robustness of each component, which may pose challenges when extending the approach to diverse languages or noisy real-world data.

While our experiments focus on Hebrew, Russian, and French, the method generalizes to any language with basic POS tagging, translation and alignment tools into English, and a pretrained language model capable of basic in-context tasks. Languages lacking these resources may pose quality challenges that are not addressed by the current proposed method. Thus, highly polysynthetic or low-resource languages remain an open direction for future work.

Acknowledgements

We would like to thank our reviewers for their constructive suggestions and insightful comments. This research has been funded by a grant from the Planning and Budgeting Committee (PBC) of the Israeli Council of Higher Education (CHE), and by a grant from the Israeli Science Foundation (ISF grant no. 670/23), for which we are grateful.

References

Omri Abend and Ari Rappoport. 2013. Universal conceptual cognitive annotation (ucca). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 228–238.

AI@Meta. 2024. Llama 3 model card.

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

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, and 1 others. 2023. Seamlessm4t: massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Abhidip Bhattacharyya, Cecilia Mauceri, Martha Palmer, and Christoffer Heckman. 2022. Aligning images and text with semantic role labels for fine-grained cross-modal understanding. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4944–4954, Marseille, France. European Language Resources Association.

Daniela Brook Weiss, Paul Roit, Ayal Klein, Ori Ernst, and Ido Dagan. 2021. QA-align: Representing cross-text content overlap by aligning question-answer propositions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9879–9894, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Avi Caciularu, Matthew Peters, Jacob Goldberger, Ido Dagan, and Arman Cohan. 2023. Peek across: Improving multi-document modeling via cross-document question-answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1989, Toronto, Canada. Association for Computational Linguistics.

Rui Cai and Mirella Lapata. 2020. Alignment-free cross-lingual semantic role labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3883–3894, Online. Association for Computational Linguistics.

Arie Cattan, Paul Roit, Shiyue Zhang, David Wan, Roee Aharoni, Idan Szpektor, Mohit Bansal, and Ido Dagan. 2024. Localizing factual inconsistencies in attributable text generation. *arXiv preprint arXiv:2410.07473*.

Simone Conia, Andrea Bacciu, and Roberto Navigli. 2021. Unifying cross-lingual semantic role labeling with heterogeneous linguistic resources. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 338–351, Online. Association for Computational Linguistics.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe

Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Zoltan Csaki, Bo Li, Jonathan Li, Qiantong Xu, Pian Pawakapan, Leon Zhang, Yun Du, Hengyu Zhao, Changran Hu, and Urmish Thakker. 2024. Sambalingo: Teaching large language models new languages. *arXiv preprint arXiv:2404.05829*.

Angel Daza and Anette Frank. 2020. X-srl: A parallel cross-lingual semantic role labeling dataset. *ArXiv*, abs/2010.01998.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Andrea Di Fabio, Simone Conia, and Roberto Navigli. 2019. VerbAtlas: a novel large-scale verbal semantic resource and its application to semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 627–637, Hong Kong, China. Association for Computational Linguistics.

Tomasz Dryjański, Monika Zaleska, Bartek Kuźma, Artur Błażejewski, Zuzanna Bordzicka, Paweł Bujnowski, Klaudia Firlag, Christian Goltz, Maciej Grabowski, Jakub Jończyk, Grzegorz Kłosiński, Bartłomiej Paziewski, Natalia Paszkiewicz, Jarosław Piersa, and Piotr Andruszkiewicz. 2022. Samsung research Poland (SRPOL) at SemEval-2022 task 9: Hybrid question answering using semantic roles. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1263–1273, Seattle, United States. Association for Computational Linguistics.

Jing Fan, Dennis Aumiller, and Michael Gertz. 2023. Evaluating factual consistency of texts with semantic role labeling. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 89–100, Toronto, Canada. Association for Computational Linguistics.

Hao Fei, Meishan Zhang, and Donghong Ji. 2020. Cross-lingual semantic role labeling with high-quality translated training corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7014–7026, Online. Association for Computational Linguistics.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. Large-scale qa-srl parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2060.

Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. 2024. Understanding cross-lingual Alignment—A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10922–10943, Bangkok, Thailand. Association for Computational Linguistics.

Luheng He, Mike Lewis, and Luke S. Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Ishan Jindal, Alexandre Rademaker, Michał Ulewicz, Ha Linh, Huyen Nguyen, Khoi-Nguyen Tran, Huaiyu Zhu, and Yunyao Li. 2022. Universal Proposition Bank 2.0. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1700–1711, Marseille, France. European Language Resources Association.

Paul Kingsbury and Martha Palmer. 2002. From treebank to propbank. In *LREC*, pages 1989–1993. Citeseer.

Ayal Klein, Eran Hirsch, Ron Eliav, Valentina Pyatkin, Avi Caciularu, and Ido Dagan. 2022. QASem parsing: Text-to-text modeling of QA-based semantics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7742–7756, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ayal Klein, Jonathan Mamou, Valentina Pyatkin, Daniela Stepanov, Hangfeng He, Dan Roth, Luke Zettlemoyer, and Ido Dagan. 2020. QANom: Question-answer driven SRL for nominalizations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3069–3083, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jérôme Louradour, Julie Hunter, Ismaïl Harrando, Guokan Shang, Virgile Rennard, and Jean-Pierre Lorré. 2024. Claire: Large language models for spontaneous french dialogue. In *Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1: articles longs et prises de position*, pages 530–548.

Meta. 2025. LLaMA 4: Leading intelligence. Unrivaled speed and efficiency. The most accessible and scalable generation of LLaMA is here. `https://www.llama.com/models/llama-4/`. Accessed: 2025-07-17.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinková, Dan Flickinger, Jan Hajic, and Zdenka Uresova. 2015. Semeval 2015 task 18: Broad-coverage semantic dependency parsing. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 915–926.

OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2025-07-17.

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2024. Contextual label projection for cross-lingual structured prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5738–5757, Mexico City, Mexico. Association for Computational Linguistics.

Leon Pesahov, Ayal Klein, and Ido Dagan. 2023. QA-Adj: Adding adjectives to qa-based semantics. In *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 74–88, Nancy, France. Association for Computational Linguistics.

Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. 2020. QADiscourse - Discourse Relations as QA Pairs: Representation, Crowdsourcing and Baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2804–2819, Online. Association for Computational Linguistics.

Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2020. Controlled crowdsourcing for high-quality QA-SRL annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7008–7013, Online. Association for Computational Linguistics.

Paul Roit, Aviv Slobodkin, Eran Hirsch, Arie Cattan, Ayal Klein, Valentina Pyatkin, and Ido Dagan. 2024. Explicating the implicit: Argument detection beyond sentence boundaries. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16394–16409, Bangkok, Thailand. Association for Computational Linguistics.

Janaki Sheth, Young-Suk Lee, Ramón Fernandez Astudillo, Tahira Naseem, Radu Florian, Salim Roukos, and Todd Ward. 2021. Bootstrapping multilingual AMR with contextual word alignments. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 394–404, Online. Association for Computational Linguistics.

Shaltiel Shmidman, Avi Shmidman, Amir DN Cohen, and Moshe Koppel. 2024. Adapting llms to hebrew: Unveiling dictalm 2.0 with enhanced vocabulary and instruction capabilities. *Preprint*, arXiv:2407.07080.

Oren Sultan and Dafna Shahaf. 2022. Life is a circus and we are the clowns: Automatically finding analogies between situations and processes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3547–3562, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jörg Tiedemann. 2022. From open parallel corpora to public translation tools: The success story of opus. In *LIVE and LEARN: Festschrift in honor of Lars Borin*, pages 133–138. University of Göteborg.

Maria Tseytlin, Paul Roit, Omri Abend, Ido Dagan, and Ayal Klein. 2025. Qa-noun: Representing nominal semantics via natural language question-answer pairs. *arXiv preprint arXiv:2511.12504*.

Sarah Uhrig, Yoalli Garcia, Juri Opitz, and Anette Frank. 2021. Translate, then parse! a strong baseline for cross-lingual AMR parsing. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 58–64, Online. Association for Computational Linguistics.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Nianwen Xue, Martha Palmer, Jena D. Hwang, Claire Bonial, Jinho Choi, Aous Mansouri, Maha Foster, Abdel aati Hawwary, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, and Ann Houston. 2013. Ontonotes release 5.0. Technical report, Linguistic Data Consortium, University of Pennsylvania / BBN Technologies. https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf.

Sangpil Youm, Brodie Mather, Chathuri Jayaweera, Juliana Prada, and Bonnie Dorr. 2024. Dahrs: Divergence-aware hallucination-remediated srl projection. In *International Conference on Applications* *of Natural Language to Information Systems*, pages 423–438. Springer.

Shiyue Zhang, David Wan, Arie Cattan, Ayal Klein, Ido Dagan, and Mohit Bansal. 2025. Qapyramid: Fine-grained evaluation of content selection for text summarization. In *Proceedings of the Conference on Language Modeling (COLM) (forthcoming)*. Association for Computational Linguistics.

A English QA-SRL/QANom Parser

For generating English annotations, we use the joint QA-SRL/QANom parser introduced by Klein et al. (2022), which models both tasks within a unified text-to-text formulation. The model is trained jointly on verbal and nominal predicates, but at inference time operates on a single marked predicate per sentence, generating natural language questions and corresponding answer spans for that predicate. This design leverages the natural alignment between QA-based semantics and sequence-to-sequence modeling while allowing consistent treatment of both QA-SRL and QANom.

We adopt the state-of-the-art implementation first released by Roit et al. (2024) and later improved by Cattan et al. (2024), which uses a larger T5 variant (T5-XL 3B) trained on the combined QA-SRL (FitzGerald et al., 2018) and QANom (Klein et al., 2020) corpora.

This parser[4] achieves state-of-the-art labeled and unlabeled argument detection on QA-SRL and QANom benchmarks and provides the English predicate-argument structures projected into Hebrew, Russian, and French in our cross-lingual pipeline.

B Illustrative Example: Limitations of Direct Surface Translation

A straightforward yet naive approach to obtaining QA-SRL annotations in a new target language is a direct round-trip translation method: the target sentence is translated into English, existing QA-SRL tools are applied to produce English question-answer pairs, and these QAs are then directly translated back into the target language. To demonstrate why this method is insufficient for cross-lingual QA-SRL projection, consider the following Hebrew sentence and its translation.

Source Sentence (Hebrew):
ולכן אני אומר לכל ילדי ישראל: סעו לירושלים.

English Translation :
And so I say to all the children of Israel: Go to Jerusalem.

English QA (QA-SRL parser on *English Translation*):
Q: *Who should go somewhere?*
A: ‘all the children of Israel’

Surface Hebrew Translation of *English QA*:
Q: מי צריך ללכת לאנשהו?
A: 'כל בני ישראל'

Issues observed:

- **Predicate drift:** The Hebrew predicate *”סעו“* (“go/plural imperative” — in the sense of a drive) from the original sentence is missing from Hebrew translation of the English QA, replaced by a different root *”ללכת“* (“to walk”). As a result, the semantics of the question is misaligned with the original context of the source sentence.

- **Argument mismatch:** The original Hebrew span 'כל ילדי ישראל' (“all the children of Israel”) is altered to 'כל בני ישראל' (“all the sons of Israel”), which is semantically different from the original argument.

- **Distorted QA structure:** In addition to the semantic drift of the back-translated QA, it also no longer corresponds precisely to the original Hebrew sentence surface form. QA-SRL is designed to provide a semi-structured representation of predicate–argument relations through the correspondence of answers to sentence spans (arguments) and the alignment of the question’s predicate with its occurrence in the sentence. Breaking these alignments hinders the downstream use of QA-SRL as a semantic representation or decomposition of target language sentences.

To conclude, this example highlights that naive direct QA translation introduces both **predicate replacements** and **argument span divergences**. Therefore, relying solely on English tools via machine translation cannot ensure faithful QA-SRL in new languages. Our approach instead incorporates constrained translation and word alignment to preserve predicate alignment and maintain accurate span mappings between the sentence and the QAs.

[4] Available at https://github.com/plroit/qasem_parser.

C Illustrative Example of the Projection Pipeline (French)

This appendix presents a step-by-step example of our multilingual QA-SRL projection pipeline for a French sentence.

Source sentence (French): *Je me suis finalement abstenue en ce qui concerne le vote pour un certain nombre de raisons.*

Translation to English: *Finally, I abstained from voting for a number of reasons.*

Step 1: Predicate identification (English) The English QA-SRL parser detects the predicate **"abstained"** (index = 3) as a verbal predicate.

Step 2: Predicate alignment Using *word alignment*, the corresponding French predicate is identified as **"abstenue"**.

Step 3: English QA-SRL output The parser produces the following English question–answer pairs:

Question	Answer
Who abstained from something?	I
What did someone abstain from?	voting
Why did someone abstain from something?	for a number of reasons

Step 4: Question translation Each question is translated into French using the *constraint translation*, ensuring that the predicate is faithfully preserved in the target language.

Step 5: Answer span alignment Using *word alignment* between English and French sentences, we identify the corresponding spans in the French text for each English answer. The aligned spans are then paired with the translated French questions, yielding the projected QA-SRL annotations for the target language:

Question (FR)	Answer
Qui s'est abstenu de quelque chose ?	Je
De quoi quelqu'un s'est-il abstenu ?	du vote
Pourquoi quelqu'un s'est-il abstenu de quelque chose ?	pour un certain nombre de raisons

Summary This example illustrates the three core operations of the projection pipeline:

- **Predicate alignment** between English and target-language verbs.
- **Constrained question translation** that preserves the original predicate in the question during translation.
- **Answer span alignment** ensuring that argument spans in the target-language sentence correspond accurately to the English span answer.

Together, these steps produce high-quality projected supervision for fine-tuning target-language QA-SRL parsers.

D Identifying Predicates in Target Language

Our projection algorithm transfers QA pairs generated on the English translation back to the original sentence in the target language $\mathcal{L}$. The QA-SRL and QANom parsers we train assume a pre-specified predicate, which must be either a verb or a deverbal nominalization. Section 5 reports QA generation performance under this setting with gold predicates provided. Because predicate detection is a prerequisite for training and inference yet introduces distinct challenges — especially for eventive nominalizations — this appendix details our strategy for identifying predicates in both the projected training data and the final parsers.

Challenges. Identifying predicates in the target language is non-trivial. While verbal predicates can be reliably detected via POS tagging, distinguishing deverbal eventive nominalizations from non-predicative nouns is substantially harder (Klein et al., 2020). This motivates our reliance on Universal Dependencies (UD) corpora, which supply both tokenization (necessary for alignment) and POS tags.

Predicate selection in training-set projection. For the projected training data, we rely on English-side predicates produced by the QA-SRL parser and their aligned tokens in $\mathcal{L}$. Since lexical category often shifts in translation (e.g., a verb in one language aligning to a nominalization in another), we take the union of QA pairs generated for both English verbs and nominalizations. Because the English parser incorporates a trained nominalization detector (Klein et al., 2020), this union covers most verbal and nominal predicates in the target language. We filter out English predicates whose aligned tokens in $\mathcal{L}$ are neither verbs nor nouns; for nouns we further require the aligned token to be identified as a deverbal nominalization by our classifier (described below).

Predicate identification at inference time. For evaluation against the manually corrected gold standard, we focus purely on QA generation and rely on gold predicates identified during the training-set projection stage using alignment and the nominalization classifier. For the released parsers, and for new sentences lacking UD annotations, we: (i) apply SpaCy for tokenization and POS tagging to detect verbs and nouns; (ii) use our nominalization classifier over candidate nouns to separate static entities from deverbal predicates; (iii) pass the identified predicates to the QA-SRL or QANom parser for question-answer generation.

Nominalization classification via in-context learning. To separate eventive nominalizations from non-predicative nouns, we employ a large language model in a few-shot in-context learning (ICL) setup. The prompt provides examples of both static entity nouns and deverbal predicates, for instance (examples shown in French):

- assiette: nom commun
- invitation: nom d'action
- comité: nom commun
- permission: nom d'action
- libération: nom d'action

Given a new noun, the model predicts whether it functions as an eventive nominalization or a static entity. This approach is effective in most cases but has inherent limitations: it does not fully capture context-sensitive uses. For example, the Hebrew noun"ארגון" ("organization") can be a static entity in ארגון הבריאות העולמי ("the World Health Organization"), but an eventive nominalization in בחדר ארגון החפצים ("the organizing of the room's items"). Context-aware extensions of this classifier are left to future work.

E Answer Span Postprocessing Heuristics

To improve answer span quality, we applied three lightweight heuristics during post-processing:

First, if the predicted span was noncontiguous (e.g., due to alignment gaps), we expanded it to the minimal contiguous span covering all tokens.

Example: For the Hebrew (tokenized) sentence: הוא נוצח השבוע ב הפרש של שני אחוזים ("He was defeated this week by a margin of two percent"), given the question איך מישהו נוצח? (How was someone defeated?), The align answer span was "ב של שני אחוזים" (by a of two percent), missing the word "הפרש`` ("margin"). Our heuristic filled the alignment gap and produced the corrected answer"בהפרש של שני אחוזים" (by a margin of two percent).

Second, for Hebrew only, we iteratively removed function words from the end of the span if they could not plausibly terminate a noun phrase (e.g., prepositions, conjunctions, and definite articles). This heuristic was designed to correct alignment artifacts that produced incomplete or ungrammatical span endings. Although the approach applies to other languages (for example, removing 'the' or 'a' in English), we applied it only to Hebrew in this work.

Example: For the Hebrew sentence: כך טען בנק ישראל ב מחקר ש פרסם באוגוסט 2015 ("Thus claimed the Bank of Israel in a study published in August 2015."), given the question מה פורסם? (What was published?), the predicted answer span was "מחקר ש" ("a study that"), which ends with the complementizer "ש`` ("that"), a function word that cannot end a grammatical phrase. Our heuristic removed it, yielding the corrected answer "מחקר`` ("study").

Third, if the span contained a sentence-internal period or included the predicate token itself, we split the span and retained the longer segment, excluding the problematic element.

These rules improved grammaticality and reduced noise in both training and evaluation.

F Predicate Preserving Constrained Translation

As discussed in Appendix B, naïve surface translation can lead to predicate drift, where the lexical root of the original verb is replaced during translation. In that example, the Hebrew predicate "סעו" ("go" — in the sense of "drive") was replaced by "ללכת" ("to walk") in the back-translated question, altering the semantics and breaking the alignment between the question and the original predicate.

To prevent such drift, we apply *predicate-preserving constrained translation*. When translating the English QA back into the target language, the LLM receives the English question along with the intended predicate from the source sentence as a lexical constraint. The prompt provided to the model is constructed as follows:

Who should go somewhere? | לנסוע

This explicitly instructs the model to preserve the original predicate root (לנסוע — "to drive / go"). As a result, the model produces the correctly aligned Hebrew question: מי צריך לנסוע לאנשהו?
(lit. "Who should go somewhere?" — using *go* in

the sense of “drive” matching the original Hebrew predicate)

As a result, the predicate remains faithfully preserved in the target-language question.

G Argument Matching Procedure and IOU Threshold Calibration

Argument Matching Procedure. To evaluate argument prediction quality, we construct a bipartite graph between predicted and gold answer spans for each predicate instance. Each edge is weighted by the token-level Intersection-over-Union (IOU) between the two spans. Edges with IOU below a threshold τ are discarded to avoid spurious partial matches. We then apply a maximal bipartite matching algorithm to select a one-to-one mapping between predicted and gold arguments that maximizes total IOU weight. Aligned pairs above τ are counted as true positives, while unmatched predicted arguments are treated as false positives and unmatched gold arguments as false negatives. This process follows the Unlabeled Argument Detection metric introduced by in Roit et al. (2020) and adopted by subsequent QA-SRL/QANom works (Klein et al., 2020, 2022; Roit et al., 2024; Cattan et al., 2024).

Threshold Calibration. Although prior QA-SRL studies (Klein et al., 2020, 2022; Roit et al., 2024; Cattan et al., 2024) typically use a relatively lenient threshold of $\tau = 0.3$, we adopt a stricter value of $\tau = 0.5$. This was determined via two complementary procedures on a manually annotated validation set: (i) selecting τ that maximized F1, and (ii) identifying the optimal cutoff from the ROC (Receiver Operating Characteristic) curve of true vs. false positive matches. Both analyses indicated that $\tau = 0.5$ yields the best balance between precision and recall, reducing false positives and improving robustness in morphologically rich and syntactically flexible languages. Figures 2 and 3 illustrate the calibration results.

H Semantic Similarity Model and Threshold Calibration

To evaluate labeled predicate–argument correctness in QA-SRL, we require a language-agnostic method for determining whether two questions express the same semantic role. We use the *paraphrase-multilingual-mpnet-base-v2* model from the *SentenceTransformers* library, which encodes questions into dense vector representations and computes semantic similarity via cosine similarity as a proxy for semantic equivalence.

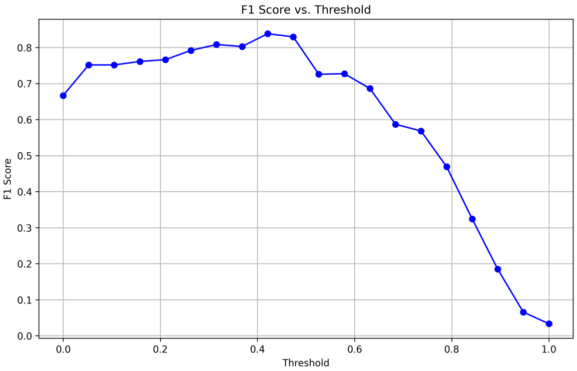

Figure 2: F1 score as a function of the IOU threshold

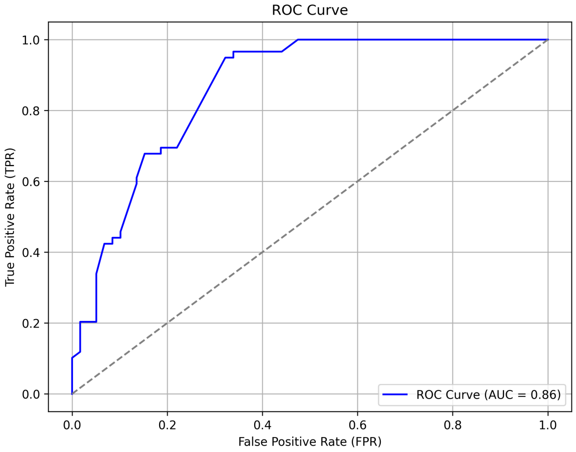

Figure 3: ROC curve for span matching decisions

Threshold calibration. To determine an appropriate similarity threshold for this semantic equivalence test, we conducted a calibration study on the Hebrew gold dataset. For a sample of predicted–gold question pairs, human annotators judged whether the predicted question was semantically equivalent to the reference question. We evaluated classifier behavior across cosine similarity thresholds ranging from 0.50 to 0.95 (in increments of 0.01), and measured agreement with human judgments.

Rather than optimizing raw accuracy or F_1, we selected the threshold that maximized the F_β score with $\beta = 0.5$, placing greater weight on precision. This reflects our preference for conservative labeled evaluation: when a predicted question is counted as correct, it should reliably express the same semantic role as the gold question. The resulting optimal threshold was **0.78**, yielding $F_{0.5} =$

0.90, with precision = 0.96 and recall = 0.70 on the annotated sample.

Precision–recall tradeoff and error profile. The calibrated Semantic Match criterion exhibits very high precision, indicating that false positives are rare. In contrast, recall is more moderate, implying that false negatives are substantially more common. To better characterize these false negatives, we manually inspected 50 Hebrew cases in which predicted answers passed the IOU-based Argument Match stage but failed the Semantic Match threshold. Half of these cases (25/50) were judged to be semantically acceptable by human inspection, typically corresponding to valid paraphrases or alternative but legitimate question formulations. This analysis indicates that the Semantic Match metric is intentionally conservative and tends to underestimate true labeled performance rather than overestimate it.

Cross-lingual application. Although threshold calibration was performed using Hebrew data, we apply the same threshold uniformly across all target languages. This choice is motivated by practical considerations and by the multilingual training of the underlying paraphrase model, under the assumption that cosine similarity scores are comparable across languages. Importantly, this threshold is a hyperparameter of the *evaluation metric* only: it does not affect the projection pipeline, model training, or deployment, and no ground-truth annotations are required at inference time.

I In-Context Prompt for QA Generation

Each model received few-shot demonstrations followed by the target sentence, with the predicate highlighted in bold. Two prompt templates were used: one for verbal predicates and one for nominal predicates. The structure and content were identical across all languages, with each model receiving the full prompt in its own language (Hebrew, Russian, or French). For clarity, the exact English versions of both prompts are shown below.

Verbal Predicate Prompt

> For a sentence with the predicate highlighted, create all questions and answers where the answers are found within the sentence. The answers must be a continuous fragment of the sentence, and the question must use the predicate as the main verb on which the question is asked.
>
> For example: for the sentence: "Zeev Revach and Hanna Laslo are well-known and beloved comedians with great energy, who **performed** last weekend"
>
> The questions and answers for the predicate "performed" are:
> Who performed? → "Zeev Revach and Hanna Laslo are well-known and beloved comedians with great energy"
> When did someone perform? → "last weekend"
>
> For this sentence and the predicate "**held**": "These people **held** high and important positions in politics, administration, and business"
>
> The questions and answers are:
> Who held? → "These people"
> Where did someone hold positions? → "in politics, administration, and business"
>
> Here is a sentence with the predicate highlighted: "<sentence with **predicate**>" Please generate all questions and answers where the answers are found within the sentence. The answers must be a continuous fragment of the sentence, and the question must contain the predicate.

Nominal Predicate Prompt

> For a sentence with the predicate highlighted, create all questions and answers where the answers are found within the sentence. The answers must be a continuous fragment of the sentence, and the question must use the predicate as an action noun, turning it into a verb, and ask a question on that verb.
>
> For example: for the sentence: "We see that **the understanding** by Euler of the algorithm as a synonym for a method of solving a problem is already very close to the modern one."
>
> The questions and answers for the predicate "understanding" are:
> What is understood? → "the algorithm"

How is something understood? → "as a synonym for a method of solving a problem"
Who understands something? → "by Euler"
How does someone understand something? → "very close to the modern one"

For the sentence: "The division commander Moshe Peled contested this and proposed instead to conduct an **attack** in the south of the Golan Heights."

The questions and answers for the predicate "attack" are:
Who can attack somewhere? → "The division commander Moshe Peled"
Where can someone attack? → "in the south of the Golan Heights"

For this sentence and the predicate "**appointment**": "In the Supreme Court, judges serve by **appointment** permanently (until age 70), with the president of the Supreme Court at their head."

The questions and answers are:
Where was someone appointed? → "In the Supreme Court"
Who was appointed? → "judges by"
What appointment? → "permanently (until age 70)"

Here is a sentence with the predicate highlighted: "<sentence with **predicate**>" Please generate all questions and answers where the answers are found within the sentence. The answers must be a continuous fragment of the sentence, and the question must contain the predicate.

J LoRA Adapter Configurations

All models were adapted using Low-Rank Adaptation (LoRA), with 4-bit NF4 quantization, linear adapter layers, dropout of 0.05, and gradient checkpointing. Optimization was performed using the AdamW optimizer with a linear learning rate schedule. Below we detail the adapter configurations and model identifiers used per language.

Hebrew. We trained LoRA adapters for **dicta-il/dictalm2.0-instruct**[5], a 7B Hebrew LLaMA 2 model, using:

- Rank: 16, LoRA alpha: 64
- Epochs: 25

As a multilingual baseline, we similarly trained LoRA adapters for **meta-llama/Meta-Llama-3-8B**[6] on the same Hebrew data and with identical hyperparameters.

Russian. We trained LoRA adapters for **sambanovasystems/SambaLingo-Russian-Base**[7], a 7B Russian-only model, with:

- Rank: 8, LoRA alpha: 32
- Epochs: 15

LoRA adapters were also trained for **meta-llama/Meta-Llama-3-8B**[2] on the Russian dataset using:

- Rank: 16, LoRA alpha: 64
- Epochs: 25

French. We trained LoRA adapters for **OpenLLM-France/Claire-7B-FR-Instruct-0.1**[8], a 7B French-only model, with:

- Rank: 16, LoRA alpha: 64
- Epochs: 10, Batch size: 64

Adapters were also trained for **meta-llama/Meta-Llama-3-8B**[2] on French using the same LoRA parameters with batch size 32.

K Manual Error Analysis of Semantic Evaluation

This appendix provides a detailed analysis of Semantic Match failures referenced in the main paper. We examine cases in which predicted answers passed the IOU threshold (≥ 0.5) but failed the semantic similarity check, focusing on error typology.

We manually analyzed 50 such cases from the Hebrew test set and assigned each predicted question to one of the following categories:

- **M (Paraphrase Model Error):** The predicted question is semantically equivalent to the gold question, but the automatic similarity model fails to detect the match.

[5] https://huggingface.co/dicta-il/dictalm2.0-instruct

[6] https://huggingface.co/meta-llama/Meta-Llama-3-8B

[7] https://huggingface.co/sambanovasystems/SambaLingo-Russian-Base

[8] https://huggingface.co/OpenLLM-France/Claire-7B-FR-Instruct-0.1

- **V (Valid Alternate Question):** The predicted question is valid for the predicted answer, but targets a different argument than the gold annotation, typically due to minor span differences.
- **P (Predicate Error):** The predicted question does not correctly correspond to the intended predicate.
- **R (Role Labeling Error):** The question targets an incorrect semantic role.

The analysis shows that 50% of the examined cases are semantically acceptable (30% paraphrase model errors and 20% valid alternate questions). The remaining cases correspond to predicate realization errors (30%) or role labeling mismatches (20%).

Overall, this breakdown clarifies the sources of Semantic Match failures and complements the robustness analyses reported in the main text.