Exploring Alignment in Shared Cross-lingual Spaces

Basel Mousi Nadir Durrani Fahim Dalvi
Majd Hawasly Ahmed Abdelali*
{bmousi,ndurrani,faimaduddin}@hbku.edu.qa
Qatar Computing Research Institute, HBKU Research Complex, Doha, Qatar

Abstract

Despite their remarkable ability to capture linguistic nuances across diverse languages, questions persist regarding the degree of alignment between languages in multilingual embeddings. Drawing inspiration from research on high-dimensional representations in neural language models, we employ clustering to uncover latent concepts within multilingual models. Our analysis focuses on quantifying the *alignment* and *overlap* of these concepts across various languages within the latent space. To this end, we introduce two metrics `CALIGN` and `COLAP` aimed at quantifying these aspects, enabling a deeper exploration of multilingual embeddings. Our study encompasses three multilingual models (`mT5`, `mBERT`, and `XLM-R`) and three downstream tasks (Machine Translation, Named Entity Recognition, and Sentiment Analysis). Key findings from our analysis include: i) deeper layers in the network demonstrate increased cross-lingual *alignment* due to the presence of language-agnostic concepts, ii) fine-tuning of the models enhances *alignment* within the latent space, and iii) such task-specific calibration helps in explaining the emergence of zero-shot capabilities in the models.[1]

1 Introduction

The emergence of multilingual contextualized embeddings has been a ground-breaking advancement, in the ever-evolving landscape of natural language processing. Adept at capturing the linguistic nuances across different languages, these embeddings have spurred a multitude of studies (Pires et al., 2019; Dufter and Schütze, 2020; Papadimitriou et al., 2021) seeking to understand the underlying mechanisms. How these models achieve multilinguality without explicit cross-lingual supervision during training is a particularly interesting question to answer.

Cross-lingual embeddings are designed to encode linguistic concepts that bridge equivalent semantic meaning across diverse languages. The question is: how well is this achieved in practice? When considering two arbitrary languages, *how well aligned are the embeddings of those languages?* and *how language agnostic are these multilingual embeddings in reality?* Addressing these questions necessitates a comprehensive approach.

In high-dimensional spaces, neural language models exhibit a capability to group words with shared linguistic associations, as highlighted by Mikolov et al. (2013). Expanding upon this foundational insight, recent research endeavors (Michael et al., 2020; Dalvi et al., 2022; Fu and Lapata, 2022) delved into conducting representation analysis within pre-trained models. Our objective, in this work, is to uncover encoded concepts within multilingual models and analyze their *alignment* and *overlap* across various languages within the latent space. We discover latent concepts by applying clustering to the underlying contextualized representations. The premise is that these clusters potentially signify latent concepts, encapsulating the language knowledge assimilated by the model. We build our work on top of this foundation to quantify concept *alignment* and *overlap* within multilingual latent space. We propose two metrics `CALIGN` and `COLAP` to quantify these two aspects and carry out analysis to study the following questions:

- To what extent do latent spaces across languages exhibit *alignment* and *overlap* in multilingual models?
- How does this change as the models are tuned towards any downstream NLP task?
- How do the multilingual latent spaces transform for zero-shot scenarios?

We conducted a study employing three multilingual transformer models: `mT5` (Xue et al., 2021), `mBERT` (Devlin et al., 2019), and `XLM-RoBERTa`

*Ahmed contributed to the project while he was at QCRI.
[1]The code is available at https://github.com/qcri/multilingual-latent-concepts

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6326–6348
August 11-16, 2024

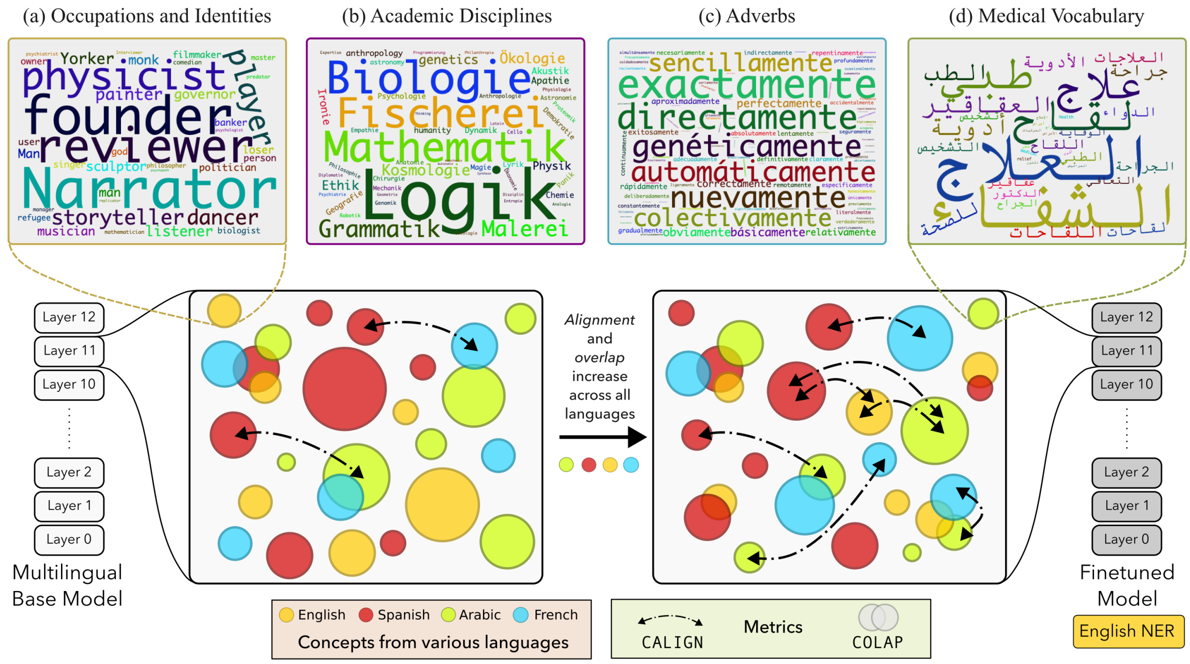


Figure 1: Overview of CALIGN and COLAP metrics in latent spaces of multilingual models, and how the space re-calibrates after fine-tuning. The top row shows concepts learned in mT5 across different languages: (a) English (b) German, (c) Spanish, (d) Arabic.

(Conneau et al., 2020). These models were fine-tuned for three downstream NLP tasks: machine translation, named-entity recognition and sentiment analysis, spanning sequence generation, labeling and classification respectively. Our analysis revealed intriguing insights, including:

- Deeper layers in multilingual models preserve semantic concepts, contrasting with language-dependent lexical learning in lower layers, resulting in a higher alignment.
- Fine-tuning calibrates the latent space towards higher alignment and the task-specific calibration of the latent space facilitates zero-shot capabilities.
- Divergent patterns emerge in the encoder and decoder latent spaces in seq2seq models. The final layers in the decoder tend to primarily retain language specific concepts.
- Closely related languages demonstrate higher overlap in latent space.
- The complexity of optimization function affects the extent of overlap in latent spaces
- While many model concepts exhibit multilingual traits, later layers post fine-tuning tend to retain primarily language-specific characteristics.

2 Methodology

The high-dimensional latent spaces learned within neural language models have been shown to encapsulate concepts formed by common linguistic attributes (Mikolov et al., 2013; Reif et al., 2019). Our approach is rooted in this foundational insight where we discover latent concepts for interpreting representational spaces in multilingual neural language models. More precisely, our study endeavors to gauge the degree of *alignment* and *overlap* of concepts across the latent spaces acquired through training models on a diverse array of languages. To this end, we introduce two metrics to quantify these phenomena. The first metric CALIGN *(Concept Alignment)* involves measuring alignment by identifying concepts that are semantically equivalent. This provides a nuanced understanding of how concepts in one language align with their counterparts in another, capturing the semantic coherence within the multilingual framework. Our second metric COLAP *(Concept Overlap)* delves into investigating the existence of overlapping cross-lingual latent spaces within the model's representation. This metric aims to highlight multilingual concepts that maintain multiple languages in a close latent space. By probing the shared latent spaces, we gain insights into the intricate relationships between concepts across languages, contributing to a more comprehensive understanding of multilingual model representations, and how they evolve when the model is trained for specific tasks. Figure 1 gives an overview of our approach. In the following sections, we detail each stage of our methodology.

2.1 Concept Discovery

Our investigation builds upon the work on discovering Latent Concepts in contextualized representations (Dalvi et al., 2022). At a high level, feature vectors (contextualized representations) are initially generated by performing a forward pass on a neural language model. The representations are then clustered to uncover the encoded concepts of the model. A concept, in this context, can be understood as a collection of words from one or more languages grouped together based on some linguistic relationship, such as lexical, semantic, syntactic, and morphological connections. Figure 1 illustrates concepts discovered within the latent space of the `mT5` model, where word representations are organized according to distinct linguistic concepts.

Formally, consider a pre-trained model $\mathbf{M}$ with L layers: $l_1, l_2, \ldots, l_L$. Using a dataset of S sentences totaling N tokens, $\mathcal{D} = [w_1, w_2, \ldots, w_N]$, we generate feature vectors: $\mathcal{D} \xrightarrow{\mathbf{M}_l} \mathbf{z}^l = [\mathbf{z}^l_1, \ldots, \mathbf{z}^l_N]$, where $\mathbf{z}^l_i$ is the contextualized representation for the word w_i from its sentence at layer l. A clustering algorithm is then employed in the per-layer feature vector space to discover layer-l encoded concepts.

2.2 Concept Alignment (CALIGN)

Multilingual neural language models are crafted to encode linguistic concepts that bridge equivalent semantic meaning across diverse languages. A key question guiding our exploration is how well this alignment is actually achieved in practice. Specifically, when considering two arbitrary languages, we seek to quantify how well the embeddings of those languages from the same neural model are aligned. We propose an alignment metric, denoted as `CALIGN` to quantify the correspondence of concepts across different languages within the latent space of multilingual models. Given a concept C_s (in language **s**) and a concept C_t (in language **t**), the number of aligned tokens $\mathcal{A}_{C_s}$ in C_s is:

$$\mathcal{A}_{C_s} = \sum_{w_s \in C_s} \mathbb{I}\left(\left(\sum_{w_t \in C_t} \mathcal{T}(w_s, w_t)\right) > 0\right)$$

where function $\mathcal{T}(w_s, w_t) = 1$ if w_s and w_t represent equivalent semantic meaning across the two languages. We simulate $\mathcal{T}(w_s, w_t)$ using a translation dictionary of N-best translations. We consider C_s and C_t to be θ_A-aligned (Λ_{θ_A}), if the following constraint is satisfied:

$$\Lambda_{\theta_A}(C_s, C_t) = \begin{cases} 1, & \text{if } \frac{\mathcal{A}_{C_s}}{|C_s|} \geq \theta_A \\ 0, & \text{otherwise} \end{cases}$$

We use a threshold θ_A to control the extent of alignment i.e. the percentage of words within a cluster required to satisfy the constraint. The alignment function proves valuable for identifying concepts that exhibit shared semantic meaning in multilingual latent spaces. Finally, `CALIGN` is the percentage of concepts from language **s** which are θ_A-aligned to some concept in another language.

2.3 Concept Overlap (COLAP)

While the alignment metric `CALIGN` helps to understand whether the model preserves encoded concepts (C_s, C_t) that can be aligned to each other, indicating their shared semantic meaning, it does not explicitly look at overlapping latent spaces across multiple languages in the same model. To investigate these overlapping latent spaces, we introduce another metric denoted as `COLAP` (`Concept Overlap`). This metric highlights concepts that encode words from multiple languages in a close latent space. Given k languages, and a set of tokens from language i as L_i, We identify a concept as overlapping if it satisfies the following constraint:

$$\mathcal{O}_C = \begin{cases} 1, & \sum_{i=1}^{k} \mathbb{I}\left(\frac{|C \cap L_i|}{|C|} \geq \theta_O\right) \geq 2 \\ 0, & \text{otherwise} \end{cases}$$

where θ_O defines the minimum threshold of words that must be present in the concept from at least two languages. `COLAP` is then computed as the percentage of total concepts that satisfy the above constraint.

Note that, the multilingual concepts may overlap while also being aligned. In such cases, both the `CALIGN` and `COLAP` metrics would identify these concepts. However, there are instances where an overlapping concept may contain related words that are not semantically equivalent, or where the concepts do not overlap but have semantic correspondence. In these scenarios, the two metrics capture distinct aspects.

3 Experimental Setup

3.1 Models and Tasks

We experimented with three multilingual transformer architectures namely: mT5, mBERT, and XLM-RoBERTa using the base versions (13 layers and 768 dimensions). The former is a state-of-the-art multilingual variant of the T5 (encoder-decoder Transformer) model and the latter two are the cross-lingual variants of the BERT and RoBERTa. To conduct the analysis, we tuned the mT5 model for the tasks of machine translation (*sequence generation*) using the TED corpus (Ansari et al., 2020). The mBERT and XLM-R models were tuned for NER-tagging (*sequence labeling*) with the Xtreme dataset (Hu et al., 2020) and Sentiment Analysis (*sequence classification*) with the SST-2 dataset (Socher et al., 2013). We experimented with English, German, French, Spanish, and Arabic.

3.2 Concept Discovery

We perform a forward pass through the models to generate contextualized feature vectors.[2] Subsequently, we apply K-means clustering[3] to the feature vectors, yielding K clusters (also referred to as encoded concepts) for both base and fine-tuned models. We set $K = 600$ and filter out representations that appear at least 10 times, following the settings prescribed by Dalvi et al. (2022).[4] We utilized the parallel data across languages to obtain the encoded concepts. This enables us to accurately compare the representational spaces generated by the same data across multiple languages. It also allows us to estimate the translation dictionary $\mathcal{T}(w_s, w_t)$. We computed word alignments using fast-align (Dyer et al., 2013) and then estimated lexical dictionaries using Moses toolkit (Koehn et al., 2007). The dictionary contains the N-best target translations of a source word. We used GPT-3.5 to annotate the latent concepts for our qualitative analysis (Mousi et al., 2023).

3.3 Thresholds

For CALIGN, we consider C_s (a concept in language **s**) to be aligned to C_t a concept in language **t**) if 80% of its types have a semantically equivalent word in C_t, i.e. $\theta_A = 0.8$. We use 10-best translations[5] of a word $w_s \in C_s$ to define this equivalence. We only consider concepts that have more than 5 word-types. Finally, we also only align concepts C_s/C_t if their sizes do not differ by more than 40%, to avoid aligning very small concepts in one language to a single large concept in another language. We also perform concept discovery independently across languages before aligning the concepts.

For computing COLAP, we perform concept discovery on multilingual data (mixed sentences from all languages). We deem a concept C to be multilingual or overlapping if all languages being considered form at least 30% ($\theta_O = 0.3$) of the concept.

While the choice of these parameters may seem arbitrary, we experimented with various configurations, such as using a $\theta_A = 0.7$–0.9 or using 5–20 best translations. The overall patterns of the results remained consistent across different configurations.[6] The selected thresholds were based on a qualitative examination of the concepts, allowing for some noise in the concept representations.

4 Results and Analysis

Cross-lingual representations are deemed to capture unified linguistic concepts across languages which enables them to generalize and to carry out the tasks for low resource languages and zero-shot scenarios. We use latent concept analysis of multilingual models to address the following questions: i) how latent space aligns and overlaps across languages in multilingual model? ii) how is the representation space calibrated as the model is tuned towards different downstream tasks? and iii) what impact does this re-calibration have on the alignment and overlap of concepts representing zero-shot languages? (which were not used for fine-tuning).

4.1 Concept Alignment

In Figure 2, we illustrate CALIGN across latent spaces in three models: mT5, mBERT, and XLM-R. Dotted lines represent base models, while solid lines denote fine-tuned models. Here the mT5 model is fine-tuned for the task of Machine Translation, mBERT for the NER-tagging and XLM-R for SST-2. The models are jointly trained using German and English samples. We discover latent concepts in both the base and fine-tuned

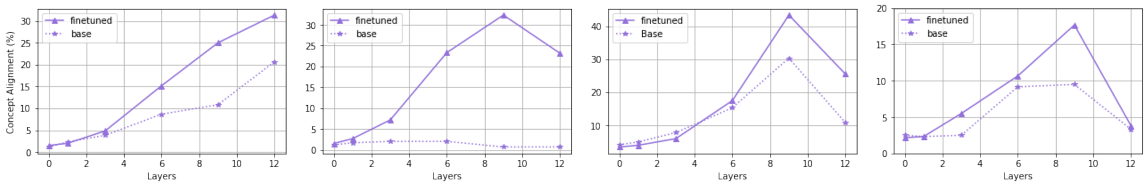
[2] We use NeuroX toolkit (Dalvi et al., 2023).

[3] Hawasly et al. (2024) showed K-means to be a viable alternative to the originally proposed agglomerative hierarchical clustering in studying latent spaces.

[4] The range of clusters (K) between 600 and 1400 yields consistent patterns, as also noted by Sajjad et al. (2022). We validated this observation in our initial experiments.

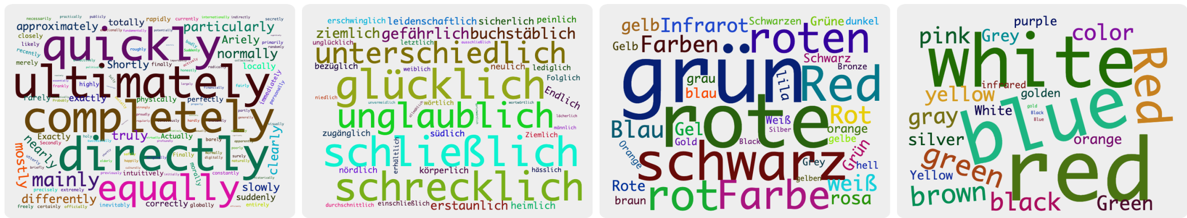
[5] A word may have many semantic meaning and translations based on different contexts.

[6] Please refer to Figure 30 in Appendix D.



(a) mT5 – encoder (MT) (b) mT5 – decoder (MT) (c) mBERT (NER) (d) XLM-R (SST-2)

Figure 2: Quantifying Concept Alignment CALIGN (%) in German–English Concepts: Dotted lines depict base models, while solid lines represent fine-tuned models across different multilingual models.



(a) Words ending with "ly" (b) Words ending with "lich" (c) Colors in German (d) Colors in English

Figure 3: Lower layers capture lexical concepts (a,b), while higher layers focus on semantic concepts (c,d).

models for English and German across different layers $(0, 1, 3, 6, 9,$ and $12)$,[7] plotting the number of aligned concepts (please refer to Section 2.2 for the definition of alignment). Here are some insights from the results:

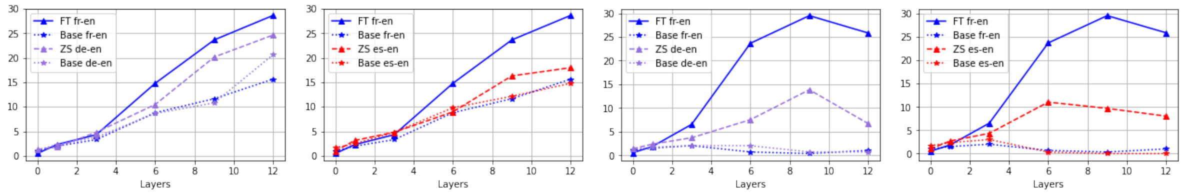
Deeper layers in multilingual models reveal increased alignment and preserve semantic concepts, contrasting with language-dependent lexical learning in lower layers. We observed a significant number of concepts that exhibited alignment within the latent spaces of these models. Notably, up to 42% of concepts demonstrated alignment within the German-English latent space of the mBERT-NER model. We noted an interesting trend where the number of aligned concepts increased with the depth of the network, reaching its peak in the higher layers of the model. In our qualitative analysis, we found that lower layers of the models are predominantly engaged in learning word morphology, including lexical concepts such as suffixation.[8] These aspects are often language-dependent, resulting in a comparatively lower alignment of latent spaces. However, as we go deeper into the network, we uncover more semantic concepts that are preserved across latent spaces in a language-agnostic manner. For example, Figures 3a and 3b present concepts in lower layers, depicting the learning of lexical concepts like derivational morphology. In contrast, Figures 3c and 3d showcase concepts learned in layer 12, highlighting the higher layers' focus on capturing similar semantic concepts (colors in this case). We found these results to hold consistently across other languages. Please refer to Appendix B for additional results.

Fine-tuning calibrates the latent space towards higher alignment. Comparing base models (dotted lines) to fine-tuned models (solid lines) revealed a notable increase in aligned concepts, particularly in higher layers. We posit that base models, trained with a multilingual MLM (mBERT and XLM-R) and "span-corruption" (mT5) objectives yield generic linguistic concepts that may not align fully across languages. However, fine-tuning models for specific tasks such as NER or translation leads to calibration of the latent space toward task-specific concepts. This aligns with prior research (Kovaleva et al., 2019; Merchant et al., 2020; Durrani et al., 2021, 2022), which indicates that higher layers of generic models become optimized for the downstream task.

We also observed that **task-specific calibration of the latent space facilitates zero-shot capabilities**. To substantiate this claim quantitatively, we extract latent concepts for zero-shot languages (not used during fine-tuning) and evaluate their alignment. Figure 4 illustrates concept alignment in the mT5 model tuned towards the task of French–English translation. We extract concepts

[7] We aimed to investigate the embedding layer, as well as the lower, middle, higher middle, and final layers.

[8] We also verified this quantitatively. See Figure 9 in Appendix B where we count the number of lexical and semantic concepts across different layers of the model.

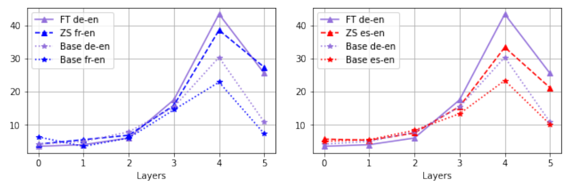


(a) zero-shot de (encoder) (b) zero-shot es (encoder) (c) zero-shot de (decoder) (d) zero-shot es (decoder)

Figure 4: Concept Alignment (%) in `mT5`. Dotted lines represent base models, solid lines denote fine-tuned French–English MT models, and dashed lines depict zero-shot alignment for German–English and Spanish–English.

	test11	test12	test13	test14
fr-en (ft)	49.0	43.8	40.8	42.7
de-en (ft)	39.9	36.4	36.9	35.5
de-en (zs)	28.2	18.9	23.1	21.7
es-en (ft)	43.3	35.9	44.7	44.5
es-en (zs)	32.0	26.7	24.0	28.2
*-en (bs)	0.01	0.02	0.10	0.20

Table 1: BLEU Scores for IWSLT tests: **ft** = the model fine-tuned for **fr–en** translation, **zs** = zero-shot performance of the pair using the **fr–en** tuned model and **bs** = the scores when using the base mT5 model.



(a) zero-shot fr (b) zero-shot es

Figure 5: Concept Alignment (%) in `mBERT`. Solid lines: fine-tuned German–English NER model. Dashed lines: zero-shot alignment for French and Spanish.

language	en	de	fr	es
	fine-tuned		zero-shot	
mbert (NER)	84.6	89.7	77.9	68.0
mbert (base)	3.0	4.9	5.8	4.1

Table 2: F1 scores for `mBERT-NER` (German,English). French and Spanish represent the zero-shot scenario.

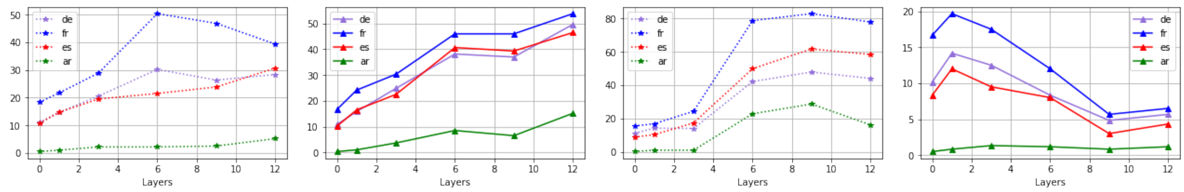
for French, English, German, and Spanish from these models on both the encoder and decoder sides, with the latter two representing zero-shot scenarios. The dashed lines indicate concept alignment for German and Spanish within these models. Notably, we observe a substantial increase in the percentage of aligned concepts, despite the model not being fine-tuned for German– or Spanish–English translation. This suggests that the presence of language-agnostic concepts within the latent space of these models facilitates performance in zero- and few-shot scenarios. Our findings correlate with the BLEU scores (Post, 2018), as shown in Table 1. Note that while the zero-shot German and Spanish translations show significantly lower performance compared to their respective models after fine-tuning, the model still performs reasonably well considering it was never explicitly trained for German- and Spanish-English translation tasks. We consistently observed these trends across various language settings in the `mT5` model[9] and in the `mBERT` model fine-tuned for the NER task for German and English. Notably the alignment improved in zero-shot French and Spanish languages (compare dashed lines (zs) to dotted lines (base) in Figure 5). Again, these findings correlate with the F1 Scores (see Table 2). We see similar results for `XLM-R` model fine-tuned for the SST2 task as well.

Divergent patterns emerge in the encoder and decoder latent spaces. Comparing our findings in `mT5`, as depicted in Figure 2, we noted disparities in alignment between the encoder and decoder spaces: i) while the base model demonstrated reasonable alignment on the encoder side (up to 20%), indicated by the dotted line in Figure 2a, alignment on the decoder side was minimal ($< 3\%$), as shown in Figure 2b. Decoders in transformer models are responsible for generating target language sequences based on the encoded input. We speculate that since its primary focus is on generating fluent and accurate translations, it may prioritize language-specific nuances and idiosyncrasies, leading to lesser aligned concepts across languages. This also explains a decrease in alignment observed in the final layers of the fine-tuned decoder.

[9] Please see Figures 21–23 in the Appendix B for results.

(a) Shared infix "olog" (de, en) (b) Anatomy & Senses (es, en) (c) Occupations (fr, en) (d) Names (ar, en)

Figure 6: Sample Overlapping Concepts in the mT5 model.



(a) mT5 encoder – Base (b) mT5 encoder – MT-tuned (c) mT5 decoder – Base (d) mT5 decoder – MT-tuned

Figure 7: Quantifying Overlapping Concepts in different languages in mT5 encoder and decoder

We see a similar dip in the last layer of encoder-only mBERT and XLM-R models for the NER and SST-2 tasks (refer to Figures 2c and 2d), which again can be attributed to the layers adapting to the task at hand instead of maintaining semantic alignment across languages.

4.2 Concept Overlap

CALIGN serves to assess whether the model captures concepts that exhibit alignment across languages, signifying shared semantic space. Our COLAP metric delves into this aspect further by exploring the presence of overlapping latent spaces within the model's representation. This sheds light on how the model effectively maintains multiple languages within a shared latent space. We demonstrate a selection of concepts demonstrating multilinguality. Figure 6a illustrates a concept at the lower layer where German and English intersect, sharing the common infix "olog". Various multilingual semantic concepts, including Anatomy & senses, Occupations and Names are depicted across different languages. Note that while CALIGN can identify the concept in Figure 6b because its constituent words are semantically equivalent, the cross-lingual words in Figure 6a are grouped based on lexical, rather than semantic similarity. COLAP helps us detect such concepts.

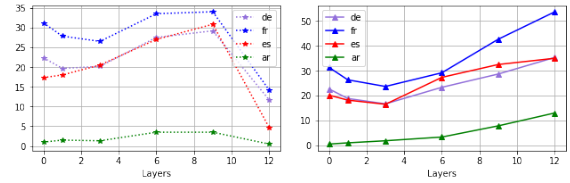
In Figures 7–8 we quantify overlap across latent spaces in various layers of mT5 and mBERT models. We note a significant number of concepts across layers with a high COLAP score in both mT5 and mBERT. The overlap typically peaks around 50% across most settings (refer to Figures 7 and 8). We draw the following insights from these results:

Closely related languages demonstrate higher overlap in latent space. We observe a spectrum of overlap across languages, with the highest degree found in French (peaking around 80%) and the lowest in Arabic (peaking around 25%) – please see Figure 7c. English and French showcase substantial overlap in their latent spaces, attributed to their shared linguistic roots within the Indo-European language family. Specifically, French stems from the Romance branch, while English belongs to the Germanic branch. This common linguistic heritage manifests in similarities in vocabulary and syntactic structures between the two languages. In contrast, Arabic exhibits notable differences in orthography and morphology when compared to English. As a Semitic language, Arabic presents unique linguistic characteristics absent in Indo-European languages like English and French. Its script diverges significantly from the Latin script, while its intricate root-and-pattern morphology stands in stark contrast to English morphology. These linguistic disparities contribute to a reduced degree of overlap in the latent space between English and Arabic compared to English and French.

The complexity of optimization function affects the extent of overlap in latent spaces While German and English share a closer linguistic relationship, and belong to the Germanic language branch within the Indo-European family, it exhibits a lesser overlap compared to French. The extent of their overlap in the latent space may be influenced by the differences in syntax, such as word order and grammatical structure, despite their linguistic closeness. Note that the base mT5 model employs span correction as its optimization function, which may primarily requires a focus on short-range dependencies. In contrast, the translation task requires the handling of long-range syntactic dependencies. Consequently, as the models are fine-tuned for machine translation tasks, we also observe a higher overlap for German in latent spaces of the fine-tuned models (See Figures 7b and 7d). We even notice an increase in overlapping concepts for Arabic-English in the higher layers post fine-tuning. A comprehensive investigation, however, is required to examine this further, and we defer this exploration to future studies.

While most of the concepts in a model exhibit multilingual traits, the later layers, post fine-tuning, tend to preserve predominantly language-specific characteristics. Although substantial overlap is evident across languages in general, the proportion of concepts that overlap diminishes to less than 20% (See Figure 7d) as the model undergoes fine-tuning for machine translation, dropping further below 5% in the final layers. This underscores that while the bulk of a model's concepts maintain multilingual attributes, the final layers within the decoder predominantly preserve language-specific traits. It's worth noting, however, that these concepts may still be semantically equivalent and satisfy CALIGN, as demonstrated in Section 4.1 (refer to Figure 4).

We do not observe a similar drop in the mBERT NER model (Figure 8b), where the consistently high overlap can be ascribed to concepts specific to output classes (e.g. location concepts), where semantic alignment may be less crucial than merely grouping locations from different languages closely together for prediction.



(a) mBERT Base (b) mBERT (NER)

Figure 8: Quantifying Concept Overlap in mBERT

5 Related Work

Numerous studies have explored the domain of multilingual embedding, investigating how deep neural language models encode knowledge across various languages without explicit supervision. Pires et al. (2019) demonstrated mBERT's ability to learn multilingual representations, enabling cross-lingual transfer even for languages with different scripts, provided they share topological similarities. Cao et al. (2020) employ a contextual word retrieval task where the model is tasked with finding corresponding words and sentences across parallel corpora. Dufter and Schütze (2020) identified critical architectural and linguistic properties for multilinguality, emphasizing the necessity of common positional embeddings, shared special tokens, and a restricted parameter space. Papadimitriou et al. (2021) investigated higher-order grammatical feature representation across languages using probing classifiers trained on mBERT embeddings. Their successful zero-shot cross-lingual transfer demonstrated parallel representation of grammatical features. Wen-Yi and Mimno (2023) conducted analysis on the embedding layer of mT5 and XLM-R, uncovering the diverse language encoding patterns within these models and highlighting the semantic encoding across languages. Xu et al. (2023) investigated the conceptual correspondence between structural concepts in linguistically diverse languages, emphasizing the correlation between conceptual alignment and cross-lingual transfer. They proposed a meta-learning approach to align these linguistic spaces, enabling zero-shot and few-shot generalization.

Our approach diverges from prior research methodologies by using an unsupervised approach to unveil multilingual concepts learned within the latent space of these models. We identify latent concepts across different languages and assess alignment across these concepts using our proposed metrics CALIGN and COLAP. Unlike previous approaches that focus on individual words and local alignment, our multilingual concept analysis provides insight into how different linguistic concepts align and overlap across multilingual spaces. We illustrate the alignment and overlap within these spaces and track their recalibration as the models undergo fine-tuning for downstream tasks. While prior research often examines if individual words have aligned counterparts in target languages, our work extends this by enforcing whether the latent spaces themselves are similarly constructed. This means that the neighbors of a word in one language correspond to neighbors of the target word in another language, introducing a stronger evidence of multilinguality at a fundamental level. Our findings suggest that this calibration of latent space enhances the model's performance in zero-shot scenarios, presenting a distinct analysis and revealing results that significantly differ from previous research.

6 Conclusion

The emergence of multilingual contextualized embeddings has sparked interest in understanding their mechanisms. We introduce two metrics, Concept Alignment (`CALIGN`) and Concept Overlap (`COLAP`), to quantify *alignment* and *overlap* within multilingual models. Our analysis reveals: i) deeper layers exhibit increased alignment due to presence of semantic concepts, ii) fine-tuning enhances alignment across cross-lingual concepts, facilitating zero-shot capabilities, iii) divergent patterns in encoder and decoder spaces and higher overlaps between closely related languages are observed. Our insights shed light on the dynamics of multilingual embeddings and lay the groundwork for a more comprehensive understanding of multilingual NLP models.

7 Limitations

We list below limitations of our work:

- While our approach effectively analyzes how multilingual models encode concepts across languages within their learned representations, it does not shed light on how these concepts are utilized by the model during prediction. Our results demonstrate a correlation between our metrics and the model's performance (as measured by BLEU and F1 scores) in the zero-shot scenarios. However, establishing causation from this correlation is not straightforward. In future research, we aim to integrate our method with ablation and knowledge attribution techniques to establish a direct connection between the encoded concepts and their impact on prediction.
- Due to the high dimensionality of contextual representations, only a restricted amount of data can be clustered to extract latent concepts. This limitation affects the goal of concept discovery, providing only a partial view of the spectrum of concepts that could be learned within the model. Our experiments were constrained by time and memory limitations. It is possible that with large-scale experimentation, we could uncover many other intriguing concepts. Additionally, time and memory constraints prevent us from exploring other clustering algorithms that may yield a superior hierarchy of concepts but are computationally infeasible.

References

Ebrahim Ansari, Amittai Axelrod, Nguyen Bach, Ondřej Bojar, Roldano Cattoni, Fahim Dalvi, Nadir Durrani, Marcello Federico, Christian Federmann, Jiatao Gu, Fei Huang, Kevin Knight, Xutai Ma, Ajay Nagesh, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Xing Shi, Sebastian Stüker, Marco Turchi, Alexander Waibel, and Changhan Wang. 2020. FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 1–34, Online. Association for Computational Linguistics.

Alexandra Birch, Matthias Huck, Nadir Durrani, Nikolay Bogoychev, and Philipp Koehn. 2014. Edinburgh SLT and MT system description for the IWSLT 2014 evaluation. In *Proceedings of the 11th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 49–56, Lake Tahoe, California.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations.

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*, pages 8440–8451. Association for Computational Linguistics.

Fahim Dalvi, Nadir Durrani, and Hassan Sajjad. 2023. Neurox library for neuron analysis of deep nlp models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 75–83, Toronto, Canada. Association for Computational Linguistics.

Fahim Dalvi, Abdul Rafae Khan, Firoj Alam, Nadir Durrani, Jia Xu, and Hassan Sajjad. 2022. Discovering latent concepts learned in BERT. In *International Conference on Learning Representations*.

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*, NAACL-HLT '19, pages 4171–4186, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Nadir Durrani, Hassan Sajjad, and Fahim Dalvi. 2021. How transfer learning impacts linguistic knowledge in deep NLP models? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4947–4957, Online. Association for Computational Linguistics.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Firoj Alam. 2022. On the transformation of latent space in fine-tuned nlp models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1495–1516, Abu Dhabi, UAE. Association for Computational Linguistics.

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

Yao Fu and Mirella Lapata. 2022. Latent topology induction for understanding contextualized representations.

Majd Hawasly, Fahim Dalvi, and Nadir Durrani. 2024. Scaling up discovery of latent concepts in deep NLP models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–806, St. Julian's, Malta. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. 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 4364–4373, Hong Kong, China. Association for Computational Linguistics.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online. Association for Computational Linguistics.

Julian Michael, Jan A. Botha, and Ian Tenney. 2020. Asking without telling: Exploring latent ontologies in contextual representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 6792–6812, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the ICLR Workshop*, Scottsdale, AZ, USA.

Basel Mousi, Nadir Durrani, and Fahim Dalvi. 2023. Can LLMs facilitate interpretation of pre-trained language models? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3248–3268, Singapore. Association for Computational Linguistics.

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532, Online. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Firoj Alam, Abdul Rafae Khan, and Jia Xu. 2022. Analyzing encoded concepts in transformer language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '22, Seattle, Washington, USA. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank.

In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Andrea Wen-Yi and David Mimno. 2023. Hyperpolyglot LLMs: Cross-lingual interpretability in token embeddings. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1124–1131, Singapore. Association for Computational Linguistics.

Ningyu Xu, Qi Zhang, Jingting Ye, Menghan Zhang, and Xuanjing Huang. 2023. Are structural concepts universal in transformer language models? towards interpretable cross-lingual generalization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13951–13976, Singapore. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Appendix

A Latent Concepts

In Figure 10, we present a selection of concepts learned within the latent space of the multilingual mT5 model. These figures showcase a diverse array of encoded concepts, encompassing lexical concepts (e.g., Figures 10a and 10d, which depict German and English words with affixes "ge" and "able" respectively), semantic concepts (e.g., Figures 10g – 10i, highlighting quantities, numbers and units of measurement in different languages), and more intricate semantic concepts illustrating fine-grained taxonomies (e.g., Figures 10b, capturing various scientific disciplines).

B Concept Alignment

In Section 4.1 we discussed several results. Here we demonstrate that our findings generalize to other languages.

Deeper layers in multilingual models reveal increased alignment and preserve semantic concepts, contrasting with language-dependent lexical learning in lower layers. We made this observation through qualitative analysis of concepts across different languages we studied in this paper. In Figures 14–16, we present lexical concepts learned within the lower layers of the multilingual

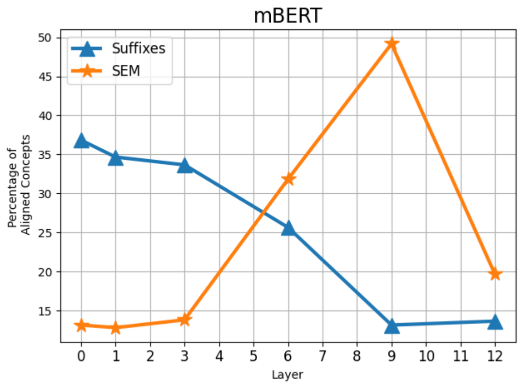


Figure 9: Layer-wise alignment of clusters to lexical and semantic properties in mBERT

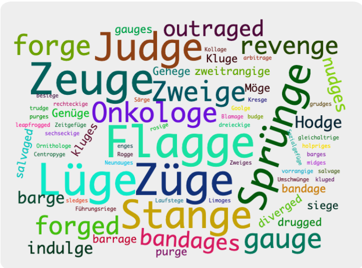
models, contrasting with the aligned semantic concepts found in the higher layers. To verify our hypothesis, we quantify the number of lexical (suffix-based concepts) and semantic concepts in English within the mBERT model. Please see Figure 9 for a layer-wise pattern of concepts.

Fine-tuning calibrates the latent space towards higher alignment We consistently higher alignment of concepts as the models were fine-tuned towards a downstream NLP task. Please refer to Figures 11–13 for results across different architectures and languages. We display alignment outcomes in base models (dotted lines) and after they were fine-tuned (solid lines). Please refer to Figures 17–20 for additional examples of concepts aligned across various languages.

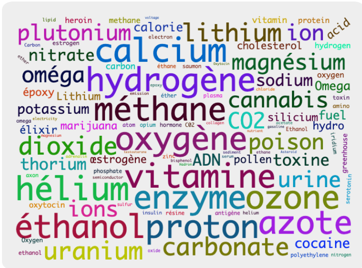
The task-specific calibration of the latent space facilitates zero-shot capabilities. In Figures 21–23, we display alignment outcomes using mT5 base models and after tuning them for the machine translation task. We examine language alignment within the encoder, decoder, and between the encoder and decoder. We observe that fine-tuning the models enhances the alignment of latent spaces. Interestingly, this increase in alignment also extends to other languages, despite the fact that the model was not specifically tuned for these zero-shot languages.

C Concept Multilinguality

In Section 4.2, we illustrated how both the base and fine-tuned models manifest concepts with overlapping latent spaces. Figure 24 showcases that these models display similar patterns even in the zero-shot scenario. Specifically, in this figure, we present the multilinguality of concepts in the mT5

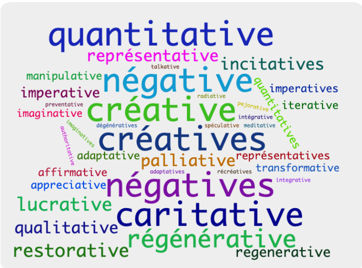


(a) “ge” infix

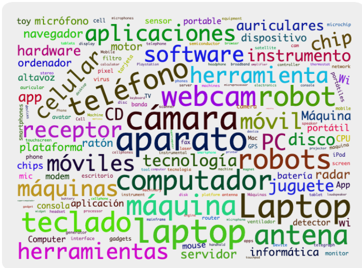


(b) Chemical Elements

(c) Modes of Transportation



(d) Words ending with “tive”

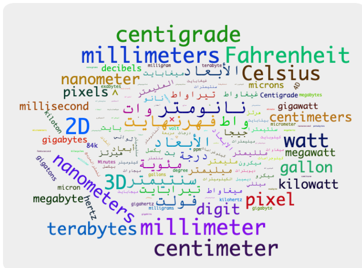


(e) Technological Devices and tools

(f) Family and Relationships

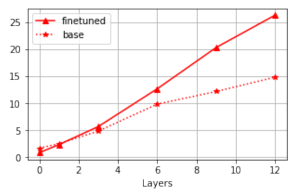
(g) Qualities and Numbers

(h) Units of Measurement

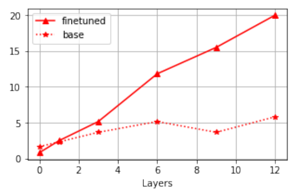


(i) Units of Measurment

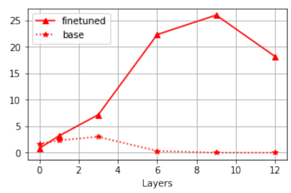
Figure 10: Sample Concepts learned in the mT5 model



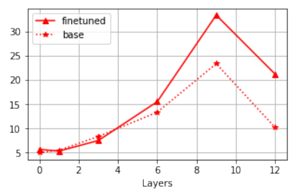
(a) mT5–encoder



(b) mT5–encoder-decoder

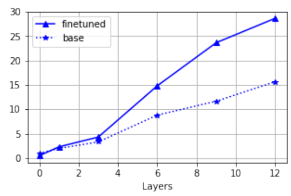


(c) mT5–decoder

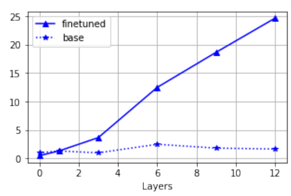


(d) mBERT

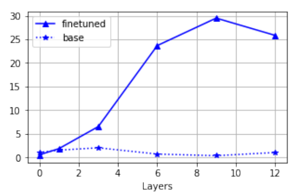
Figure 11: Quantifying Alignment Percentage in Spanish–English Concepts: Dotted lines depict base models, while solid lines represent fine-tuned models across different multilingual models.



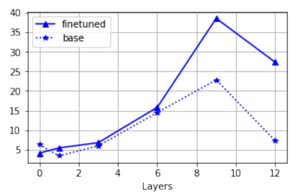
(a) mT5–encoder



(b) mT5–encoder-decoder

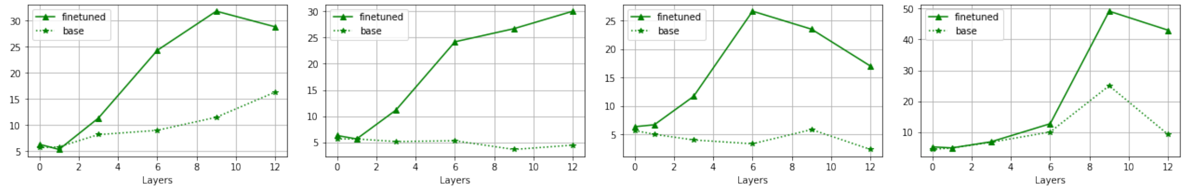


(c) mT5–decoder



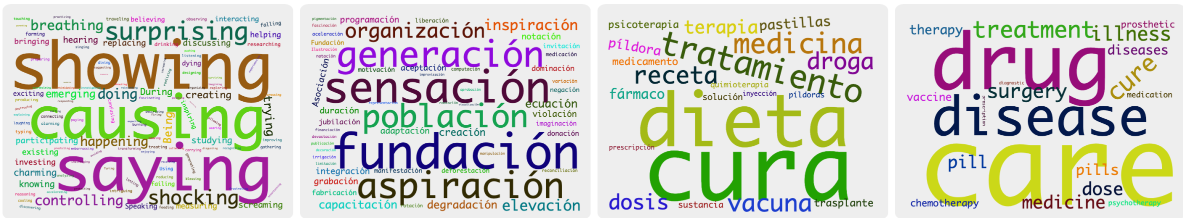
(d) mBERT

Figure 12: Quantifying Alignment Percentage in French–English Concepts: Dotted lines depict base models, while solid lines represent fine-tuned models across different multilingual models.



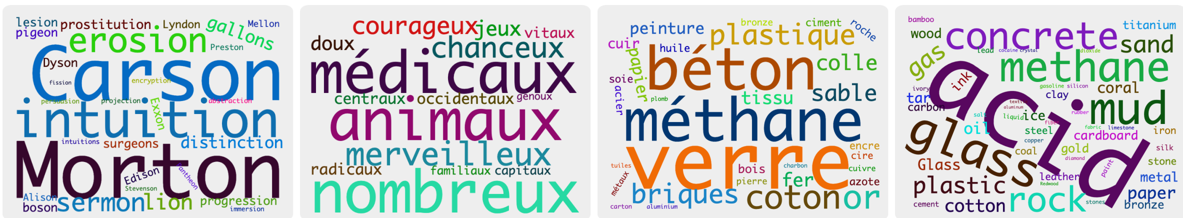
(a) mT5–encoder (b) mT5 encoder-decoder (c) mT5 decoder (d) mBERT

Figure 13: Quantifying Alignment Percentage in Arabic–English Concepts: Dotted lines depict base models, while solid lines represent fine-tuned models across different multilingual models.



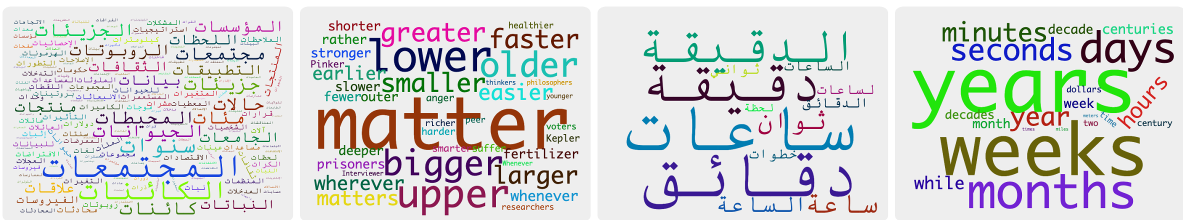
(a) Words ending with “ing” (b) words ending with “ión” (c) Medical terms in Spanish (d) Medical Terms in English

Figure 14: Spanish-English Concepts learned in the mT5 model: Lower layers (a and b) capture lexical concepts, while higher layers focus on semantic concepts (c and d).



(a) Words ending with “on” (b) Words ending with “ux” (c) Materials and Substances (d) Materials and Substances

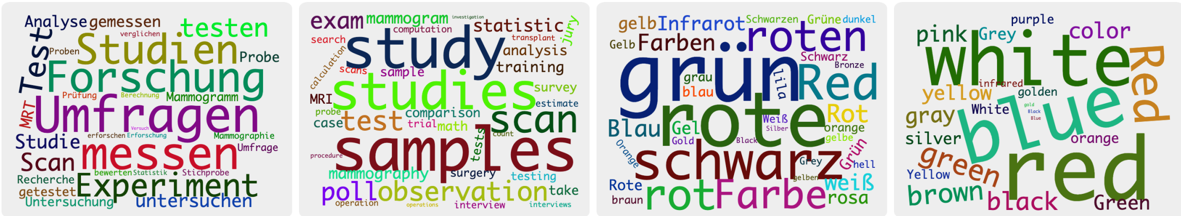
Figure 15: French-English Concepts learned in the mT5 model: Lower layers (a and b) capture lexical concepts, while higher layers focus on semantic concepts (c and d).



(a) Words ending with “At” (b) Shared infix “er” (c) Time phrases in Arabic (d) Time phrases in English

Figure 16: Arabic-English Concepts learned in the mT5 model: Lower layers (a and b) capture lexical concepts, while higher layers focus on semantic concepts (c and d).

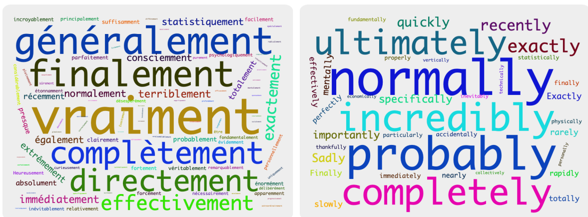
(a) Superlatives in German (b) Superlatives in English (c) Math related terms (de) (d) Math related terms (en)



(e) Study related (de) (f) Study related (en) (g) Colors in German (h) Colors in English

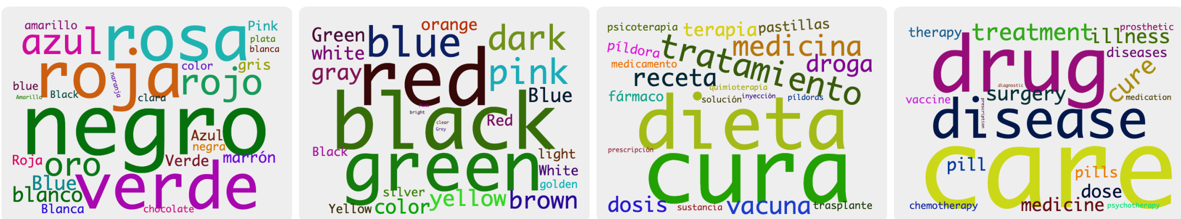
Figure 17: Pairs of Concepts in German-English mT5 model

(a) Nationality & Identity (fr) (b) Nationality & Identity (en) (c) Chemical Materials (fr) (d) Chemical Material (en)

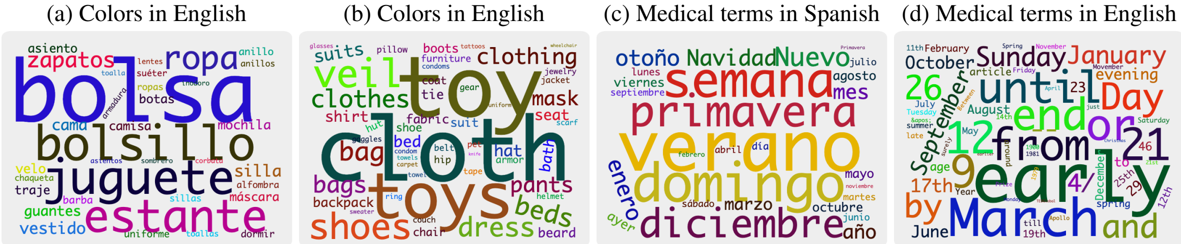


(e) Adverbs (fr) (f) Adverbs (en)

Figure 18: Pairs of Concepts in French-English mT5 model



(a) Colors in English (b) Colors in English (c) Medical terms in Spanish (d) Medical terms in English



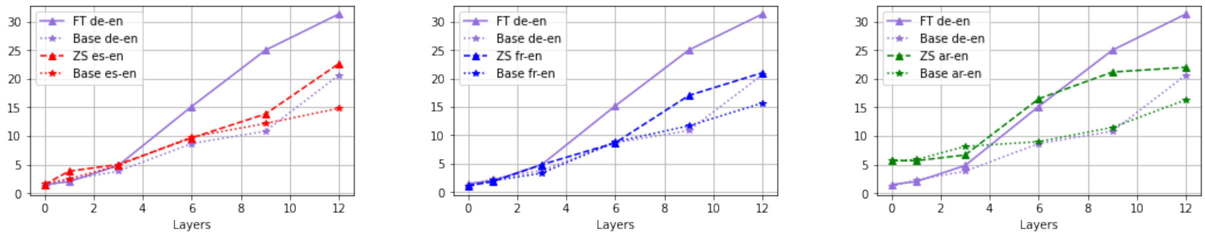
(e) Assorted Items (es) (f) Assorted Items (en) (g) Temporal terms in Spanish (h) Temporal terms in English

Figure 19: Pairs of Concepts in Spanish-English mT5 model

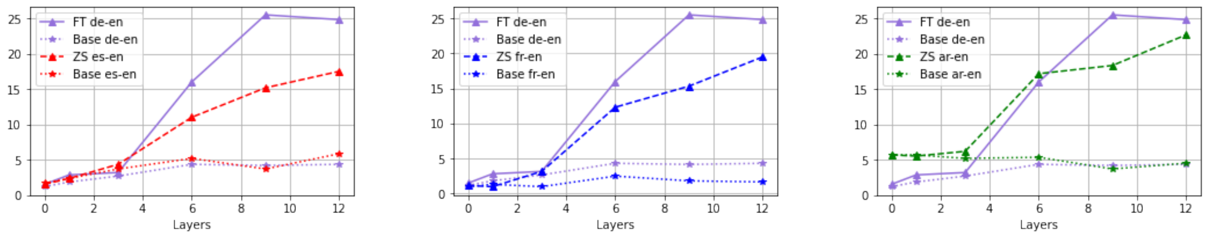
(a) Colors in Arabic (b) Colors in English (c) Time spans in Arabic (d) Time spans in English

(e) Geographical entities (ar) (f) Geographical entities (en) (g) Morphological variations (h) Verb transformations

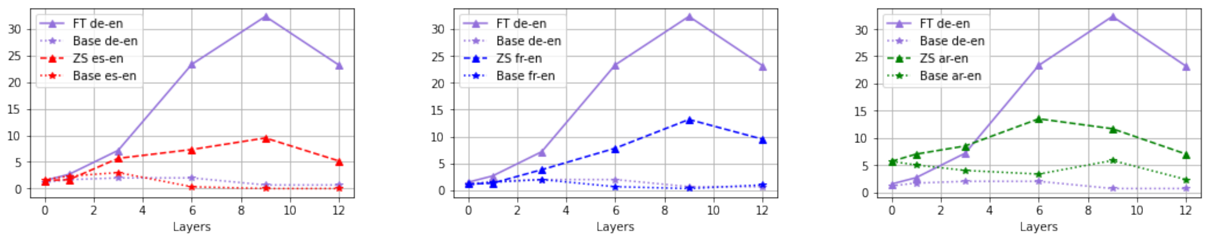
Figure 20: Pairs of Aligned Concepts in Arabic-English mT5 model



(a) zero-shot **es** on **de** encoder (b) zero-shot **fr** on **de** encoder (c) zero-shot **ar** on **de** encoder

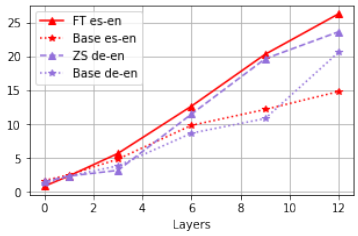


(d) zero-shot **es** on **de**↔**en** (e) zero-shot **fr** on **de**↔**en** (f) zero-shot **ar** on **de**↔**en**

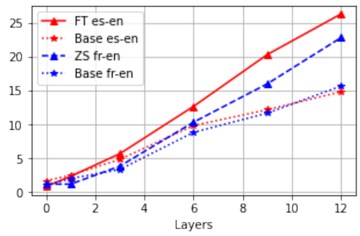


(g) zero-shot **es** on **de** decoder (h) zero-shot **fr** on **de** decoder (i) zero-shot **ar** on **de** decoder

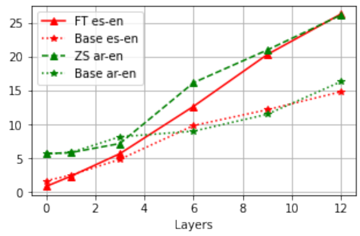
Figure 21: Percentage of Aligned Concepts: Dotted lines represent base models, solid lines denote fine-tuned German–English model, and dashed lines depict zero-shot alignment for spanish (left column), French–English (Middle column) and Arabic-English (right column); enc: Encoder, dec: Decoder



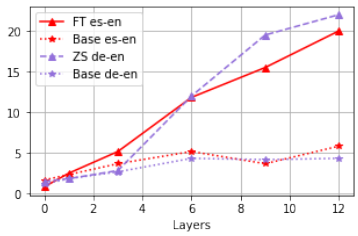
(a) zero-shot **de** on **es** encoder



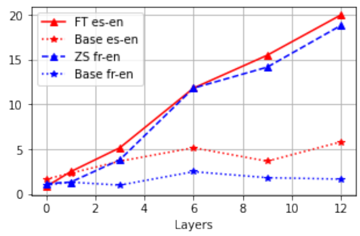
(b) zero-shot **fr** on **es** encoder



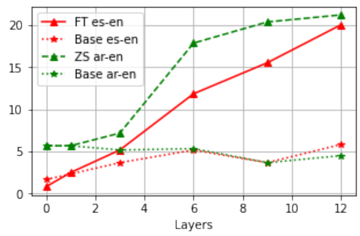
(c) zero-shot **ar** on **es** encoder



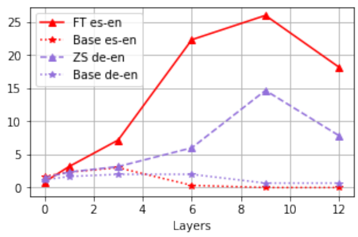
(d) zero-shot **de** on **es**↔**en**



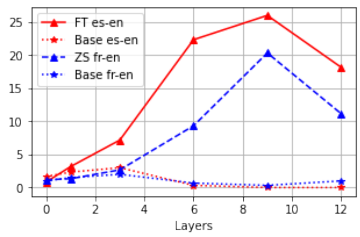
(e) zero-shot **fr** on **es**↔**en**



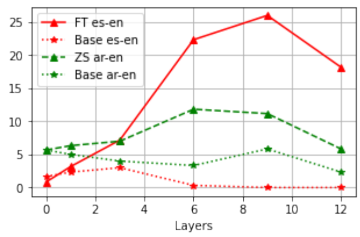
(f) zero-shot **ar** on **es**↔**en**



(g) zero-shot **de** on **es** decoder

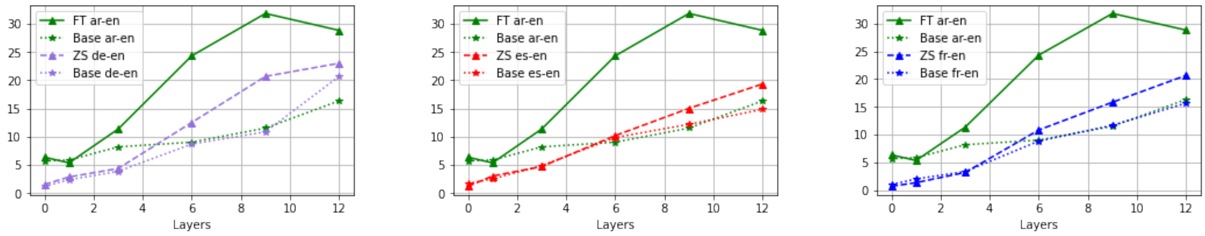


(h) zero-shot **fr** on **es** decoder

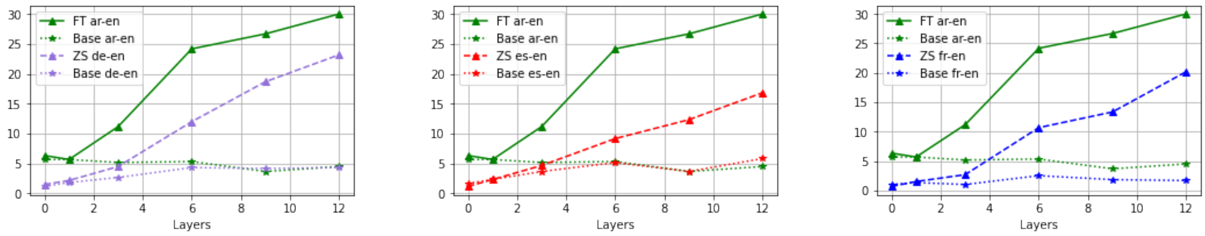


(i) zero-shot **ar** on **es** decoder

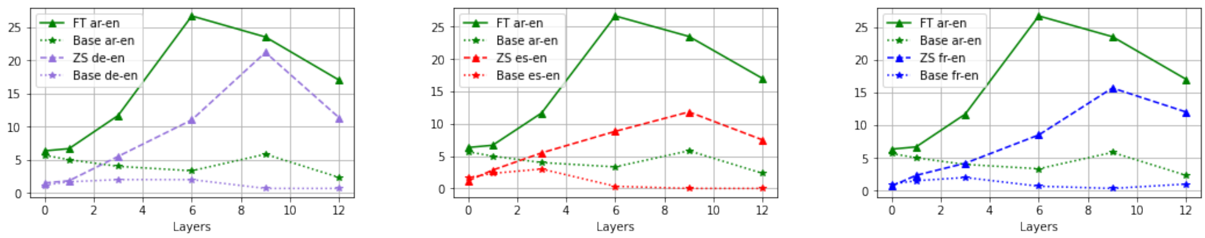
Figure 22: Percentage of Aligned Concepts: Dotted lines represent base models, solid lines denote fine-tuned Spanish–English model, and dashed lines depict zero-shot alignment for German-English (left column), French–English (Middle column) and Arabic-English (right column); enc: Encoder, dec: Decoder



(a) zero-shot **de** on **ar** encoder (b) zero-shot **es** on **ar** encoder (c) zero-shot **de** on **fr** encoder

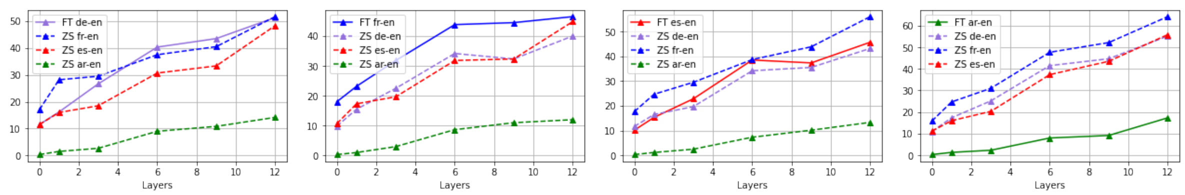


(d) zero-shot **de** on **ar↔en** (e) zero-shot **es** on **ar↔en** (f) zero-shot **fr** on **ar↔en**

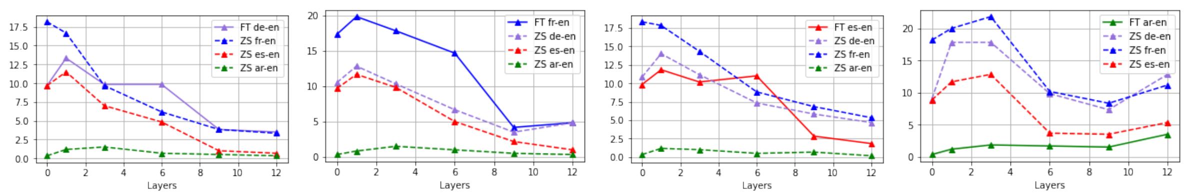


(g) zero-shot **de** on **ar** decoder (h) zero-shot **es** on **ar** decoder (i) zero-shot **fr** on **ar** decoder

Figure 23: Percentage of Aligned Concepts: Dotted lines represent base models, solid lines denote fine-tuned Arabic–English model, and dashed lines depict zero-shot alignment for German-English (left column), French–English (Middle column) and Arabic-English (right column); enc: Encoder, dec: Decoder

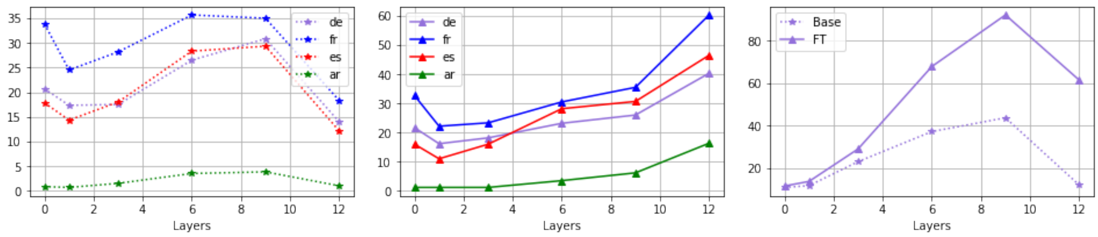


(a) 0-shot **fr, es, ar** on **de** enc (b) 0-shot **de, es, ar** on **fr** enc (c) 0-shot **de, fr, ar** on **es** enc (d) 0-shot **de, fr, es** on **ar** enc



(e) 0-shot **fr, es, ar** on **de** dec (f) 0-shot **de, es, ar** on **fr** dec (g) 0-shot **de, fr, ar** on **es** dec (h) 0-shot **de, fr, es** on **ar** dec

Figure 24: Quantifying Concept Overlap in different languages in mT5 encoder and decoders.



(a) XLM-R Base (b) XLM-R (NER) (c) XLM-R (SST)

Figure 25: Quantifying Concept Overlap in XLM-R

encoder and decoder models under zero-shot conditions. Notably, we observe that the zero-shot overlap (depicted by dashed lines) follows a comparable pattern to the overlap of latent spaces after fine-tuning (indicated by solid lines).

D Thresholds

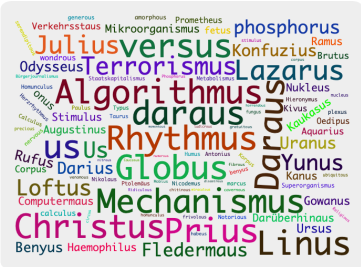
In Section 3.3 we mentioned the threshold we used for our experiments including the matching threshold, n-best translations to estimate $\mathcal{T}(w_s, w_t)$ and minimum number of types per concept. The choice of these parameters is arbitrary. We experimented with various configurations, such as using a 70–90% matching types, using 5–20 best translations. The overall patterns of the results remained consistent across different configurations (please refer to Figure 30). The selected thresholds were chosen based on a qualitative examination of the concepts, allowing for some noise in the concept representations.

E Data Statistics

In this section, we report the data statistics that we used for the experiment. Table 3 shows the number of sentences for the TED data (Birch et al., 2014) used for the machine translation experiments, Table 4 shows the statistics for the NER data used, and Table 5 shows the statistics for the sentiment analysis data used.

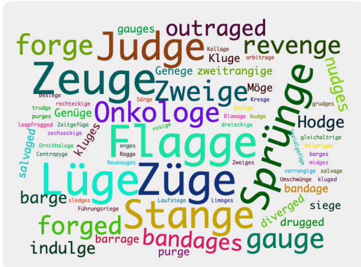
F Computing Budget

The extraction of the representations from a multilingual model requires 500GB of RAM memory. The clustering experiments for the extracted representations require 30GB of RAM memory each.



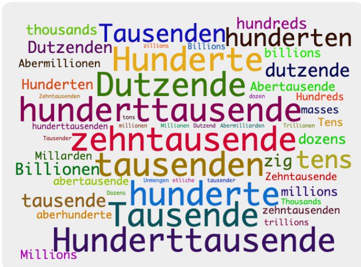
(a) Words ending with “us”

(b) Words containing “land”



(c) “ge” infix

(d) Conflict and competition



(e) Qualities and Numbers

(f) Landforms and Natural Features

(g) Furniture and Surfaces

(h) Medical and Scientific professions

(i) Commercial Establishments

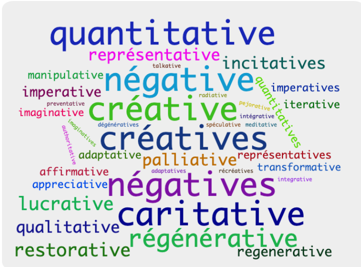
(j) Nationalities and Ethnicities

(k) Weather and Tempratures

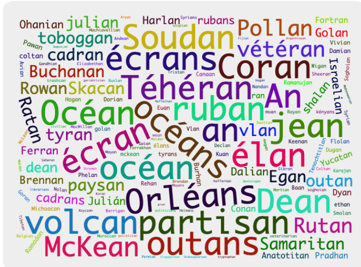
(l) Units of Measurement

Figure 26: Overlapping German-English Concepts in the MT-tuned mT5 model

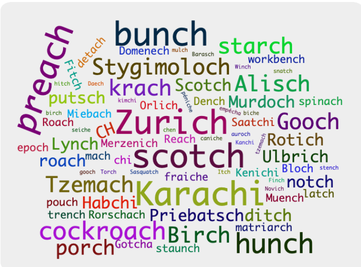
(a) Words ending with "able"



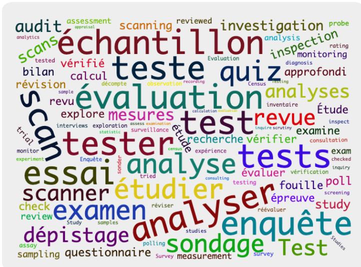
(b) Words ending with "tive"



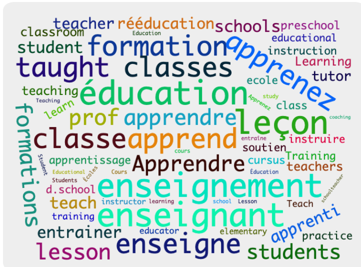
(c) words ending with "an"



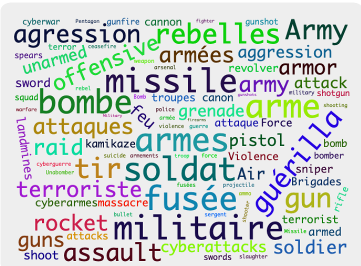
(d) words ending with "ch"



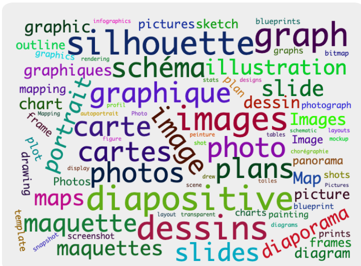
(e) Research Terminology



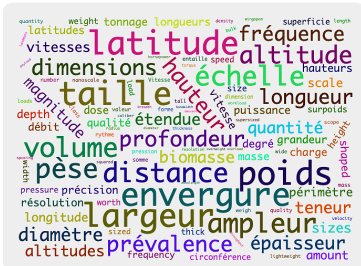
(f) Educational terms



(g) Military and Violence

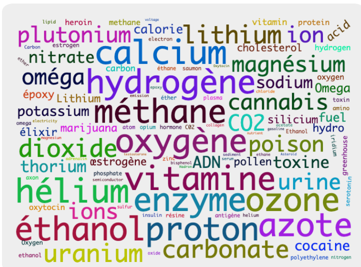


(h) Visual representation vocabulary



(i) Measurements Vocabulary

(j) Emotional Expression



(k) Chemical Elements

(l) Modes of Transportation

Figure 27: Overlapping French-English Concepts in the MT-tuned mT5 model

(a) Words containing “ve”

(b) words containing "able"

(c) "ch" infix

(d) Literature Writing and Vocabulary

(e) Family and Relationships

(f) Scientific Terms

(g) Chemical compounds

(h) Emotions and States of mind

(i) Technological devices and tools

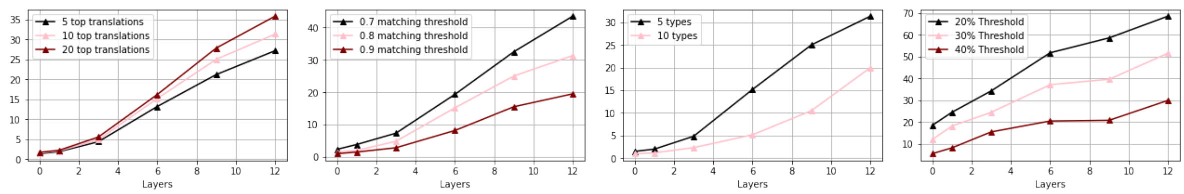
Figure 28: Overlapping Spanish-English Concepts in the MT-tuned mT5 model

(a) Recreational Sports and Activities (b) Anatomical Terminology (c) Adverbs of Emphasis and Certainty

(d) Geographical and Urban Terms (e) Time periods and Decades (f) Relationships and Connections

(g) Paths and Transportation (h) Nationalities and Ethnicities (i) Units of Measurment

Figure 29: Overlapping Arabic-English Concepts in the MT-tuned mT5 model



(a) N-best translations (b) Matching threshold (c) Minimum types per concept (d) Overlapping threshold

Figure 30: Varying different threshold parameters in CALIGN and COLAP

	de-en	es-en	fr-en	ar-en
train	209330	184724	234033	229194
test11	1433	1435	818	1199
test12	1700	1701	1124	1702
test13	992	1197	1026	1167
test14	1305	1305	1305	1107

Table 3: TED data statistics (number of sentences).

	de	es	fr	ar	en
train - sentences	20000	20000	20000	20000	20000
train - tokens	195387	129283	136788	129184	160394
validation - sentences	10000	10000	10000	10000	10000
validation - tokens	97805	64329	68220	64291	80536
test - sentences	10000	10000	10000	10000	10000
test - tokens	97646	64727	68754	64347	80326

Table 4: Xtreme NER data statisics

	en	de
train	67437	67437
validation	872	872
test	1821	1821

Table 5: SST2 data statistics