Spectral Probing

Max Müller-Eberstein and Rob van der Goot and Barbara Plank

Department of Computer Science, IT University of Copenhagen, Denmark
Center for Information and Language Processing (CIS), LMU Munich, Germany
Munich Center for Machine Learning (MCML), Munich, Germany
mamy@itu.dk, robv@itu.dk, b.plank@lmu.de

Abstract

Linguistic information is encoded at varying timescales (subwords, phrases, etc.) and communicative levels, such as syntax and semantics. Contextualized embeddings have analogously been found to capture these phenomena at distinctive layers and frequencies. Leveraging these findings, we develop a fully learnable frequency filter to identify spectral profiles for any given task. It enables vastly more granular analyses than prior handcrafted filters, and improves on efficiency. After demonstrating the informativeness of spectral probing over manual filters in a monolingual setting, we investigate its multilingual characteristics across seven diverse NLP tasks in six languages. Our analyses identify distinctive spectral profiles which quantify cross-task similarity in a linguistically intuitive manner, while remaining consistent across languages—highlighting their potential as robust, lightweight task descriptors.

1 Introduction

Analyzing the contextualized embedding representations of pre-trained language models (LMs) using lightweight probes (Hewitt and Liang, 2019; Voita and Titov, 2020) has identified latent features in the untuned encoders which are highly relevant to downstream NLP tasks at various layer depths (Tenney et al., 2019). Orthogonally, linguistic phenomena are also encoded at different timescales: i.e., rapidly changing (sub-)word-level information versus slower changing sentence or paragraph-level information. Decomposing contextualized embeddings into frequencies with different rates of change has yielded initial insights into the timescales at which these task-specific latent phenomena occur (Tamkin et al., 2020). These findings currently rely on handcrafted spectral filters and are limited to English. To enable more efficient analyses of finer-grained, continuous frequency spectra in contextualized representations covering more tasks and languages, this work contributes:

- A fully differentiable spectral probing framework for learning which frequencies are relevant for specific NLP tasks (Section 2).
- A multilingual probing study examining timescale characteristics of seven diverse NLP tasks across six languages (Section 3).
- An analysis of the relationships between the spectral profiles of different tasks and their consistency across languages (Section 4).

2 Probing for Spectral Profiles

Spectral Probing (Figure 1) builds on established signal processing methods (Ahmed et al., 1974) and recent findings on the manual frequency filtering of

![Figure 1: Visualization of Spectral Probing. Given a sequence of embedding values, decompose into composite frequency waves using DCT, apply the learned filter, retaining a subset of waves, for which IDCT returns the filtered sequence of values.](https://github.com/mainlp/spectral-probing)
contextual embeddings (Tamkin et al., 2020). The method automatically learns spectral profiles which measure the relevance of specific frequencies to a given task by amplifying or reducing contextual information with different rates of change.

**Discrete Cosine Transform** (Ahmed et al., 1974; DCT) is an invertible method for decomposing any sequence of real values \( \{x_0, \ldots, x_{N-1}\} \) (e.g., all values of an embedding dimension) into a weighted sum over cosine waves with different frequencies. The number of frequencies equals the sequence length \( N \), as the lowest frequency wave is a constant \( (k = 0) \) and the highest frequency wave completes one cycle every timestep \( (k = N - 1) \). The coefficient \( X_n^{(k)} \) for a wave at DCT index \( k \) at timestep \( n \) is calculated as:

\[
X_n^{(k)} = \sum_{n=0}^{N-1} x_n \cos \left( \frac{\pi}{N} \left( n + \frac{1}{2} \right) k \right). \tag{1}
\]

Inverting the DCT (IDCT) using all \( X_n^{(k)} \) will return the original sequence. However, weighting coefficients for some \( k \) by 0 will return a filtered version. Zeroing out all \( k \) above a threshold will only retain lower frequencies and make values oscillate with a slow rate of change. Vice-versa, zeroing out all \( k \) below a threshold will only retain higher frequencies—amplifying short-term changes.

**Fixed-band Filters** Applying frequency filters to a sequence of contextualized embeddings extracts linguistic information at different timescales. Within this formulation, the values across each embedding dimension are gathered into a real-valued sequence to which transformations such as the DCT can be applied. In seminal work, Tamkin et al. (2020) apply manually defined low \((k \in [0, 1])\), mid-low \((k \in [2, 8])\), mid \((k \in [9, 33])\), mid-high \((k \in [34, 129])\) and high frequency filters \((k \in [130, 511])\) to English BERT embeddings (Devlin et al., 2019) to investigate how accurately a linear probe can extract task-specific information within certain spectra. Capturing the full picture using manual, fixed-band filters is however not computationally feasible: Relevant frequencies might not lie in a contiguous band, and furthermore, frequencies can not only be turned on or off (i.e., weighted 0 or 1), but can actually be weighted continuously in \([0, 1]\).

**Learnable Filters** To capture the complete picture, we propose spectral probing which *learns* a continuous weighting of frequencies relevant to a task. In effect, the spectral filter is a vector \( \gamma \in \mathbb{R}^M \) for which each entry corresponds to the weight assigned to a particular frequency. Before inverting the DCT, each \( X_n^{(k)} \) is multiplied by the sigmoid-scaled weight \( \gamma^{(k)} \in [0, 1] \) which will then retain or filter out frequencies at index \( k \). As \( M \) depends on the sequence length \( N \), which changes across inputs, the spectral probe dynamically scales \( \gamma \) to the length at hand using adaptive mean pooling. In practice, we set \( M \) to the maximum input length for our given encoder (e.g., 512 for BERT) and shrink \( \gamma \) appropriately, as a wave cannot cycle more often than there are values. It would however be equally possible to set \( M \) smaller than \( N \) and interpolate the filter up to the length required. Overall, \( \gamma \) is a lightweight parameter which can be easily incorporated between the frozen encoder and probing head, and uses the existing training objective to jointly learn which frequencies to amplify or filter out.

### 3 Experiments

#### 3.1 Monolingual

**Setup** Initially, we compare spectral probing to previous fixed-band filters by reproducing the highest and lowest frequency experiments by Tamkin et al. (2020). These are the tasks of tagging parts-of-speech (PoS) in the Penn Treebank (Marcus et al., 1993; PTB) as well as classifying topics in the 20 Newsgroups corpus (Lang, 1995; 20News).

On the modeling side, we follow Tamkin et al. (2020) and train a linear probe (Alain and Bengio, 2017) on top of the frozen LM encoder to classify each manually/automatically filtered contextual embedding in an input sequence. This corresponds to probing and evaluating for the amount of task-relevant information in each sub-word across a sequence (e.g., underlying topic contextualization). The bands for the five manual filters follow the original definitions (see Section 2), and we compare them to unfiltered (ORIG) as well as automatically filtered (AUTO) embeddings from our spectral probe (details in Appendix B).

**Results** Figure 2 shows the accuracy (ACC) of the six prior filtering strategies in addition to the learned frequency weightings of the spectral probe. The unfiltered and manually filtered embeddings corroborate previous findings (Tamkin et al., 2020),
with high frequencies performing best on POS, and the lowest frequencies performing best on TOPIC.

The spectral probe achieves 95.9% ACC for POS, outperforming ORIG by a 0.1% margin and the best manual filter by 5.2%. The spectral profile in Figure 2a (right) sheds light on why this may be the case: While it also prioritizes high (sub-)word-level frequencies, the learned filter additionally includes surprising amounts of mid-high and lower frequencies, emphasizing the need for both local and global context to achieve high performance.

For TOPIC, the spectral probe achieves 72.1% ACC, outperforming both ORIG (41.3%) and the fixed low-band filter (71.2%). The learned filter (see Figure 2b, right) mirrors the fixed-band results: Only the lowest bands are active, while all higher ones are not. As mid-low frequencies still appear to contain weaker amounts of topic information, the soft inclusion of this region by the spectral probe could account for its performance boost. Overall, spectral probing confirms and refines frequency ranges from prior work while surfacing more detail and requiring no manual probe engineering, with only a single probing run instead of five.

### 3.2 Multilingual

Leveraging spectral probing, we extend timescale analyses beyond English and investigate spectral profiles across more diverse tasks and languages.

#### Setup
Each experiment covers German (DE), English (EN), Spanish (ES), French (FR), Japanese (JA) and Chinese (ZH). The tasks are POS-tagging and dependency relation classification (DEP) from Universal Dependencies (Zeman et al., 2021); named entity recognition (NER) from WikiANN (Pan et al., 2017); question answering (QA) from MKQA (Longpre et al., 2021); sentiment analysis (SENTI) and TOPIC classification from Multilingual Amazon Reviews (Keung et al., 2020); natural language inference (NLI) from XNLI (Conneau et al., 2018) and JSNLI (Yoshikoshi et al., 2020) for JA (details and examples in Appendix A).

For each language-task combination we train a linear probe on the unfiltered embeddings of multilingual BERT (Devlin et al., 2019; mBERT) and on the automatically filtered representations from our spectral probe. The remaining settings are identical to the monolingual setup (details in Appendix B).

### Table 1: Multilingual Results

| TASK    | ORIG         | AUTO         |
|---------|--------------|--------------|
| POS     | 92.4±1.9     | 92.5±1.8     |
| DEP     | 78.6±4.3     | 79.3±4.3     |
| NER     | 88.0±2.7     | 88.1±2.6     |
| QA      | 62.9±1.6     | 67.1±1.2     |
| SENTI   | 57.4±0.9     | 64.3±1.1     |
| TOPIC   | 27.1±8.1     | 37.2±8.2     |
| NLI     | 44.1±4.1     | 56.3±5.6     |

Means ± standard deviations over languages and random initializations (details in Appendix C).

![Figure 2](image2.png)

**Figure 2: Monolingual Results on PTB and 20News.** ACC of unfiltered (ORIG), low (L), mid-low (ML), mid (M), mid-high (MH), high (H), and the spectral probe’s automatic filters (AUTO) with frequency weightings.

![Figure 3](image3.png)

**Figure 3: Spectral Profiles** of all tasks (weight per frequency), with lower and upper bounds across languages.
we analyze more extensively next.

while QA, S

appears to benefit the least from both lower

and higher-frequency information. Instead, the

strong weight on mid-high frequencies matches the

and higher-frequency information. Instead, the

frequency information is most important to retrieve

PoS, but reaching the performance of the original

embeddings also requires some lower-frequency

information—most likely to disambiguate difficult

cases based on sentence-level context.

DEP appears to benefit the least from both lower

and higher-frequency information. Instead, the

strong weight on mid-high frequencies matches the

fact that dependency relations span multiple words

and benefit from information at the phrase-level.

NER sees a further decrease in high-frequency

information, coupled with an uptick in lower

frequencies. We hypothesize that phrase and sentence-

level information become more important for dis-

ambiguating certain entity types (e.g., ORG and

LOC). Across the token-level tasks this shift from

higher to lower frequencies is also reflected in filter

overlap which decreases from syntactic to semantic

token-level tasks, while their overlap with sentence-

level tasks increases (Figure 4a).

The sequence-level tasks share low-frequency

spectral profiles which overlap more with each

other than do the token-level tasks. In fact, Senti

and TOPIC overlap almost perfectly (although the

latter involves less mid-range frequencies). This

similarity is unlikely to be the result of the shared

underlying dataset as both tasks also overlap with

the unrelated XNLI and JSNLI datasets. At the

same time, the PoS and DEP tasks, which also

share datasets, have a lower overlap despite being

based on the exact same inputs. Overall, Senti,

TOPIC and NLI all appear to rely on information

which is consistent across a sequence—explaining

why simple methods such as mean-pooled sentence

embeddings can be effective in these scenarios.

QA provides an intermediate case: While it is

reliant on low frequencies it also includes more

mid-low and a small amount of higher frequency

information. This is reflected in Figure 4a, where

it shares more overlap with the token-level tasks

than all other sequence-level tasks. Since prob-

ing for the correctness of a question-answer pair

is dependent on finer-grained information than the

general sentiment, topic or semantic coherence of

a sequence, this inclusion of higher frequency in-

formation matches linguistic intuitions.

Cross-lingual Consistency Finally, we in-

vestigate the similarity of learned spectral profiles

across languages. While Figure 3 shows that there

is some variance between the filters of different

languages within a task, Figure 4b shows that ac-

tual quantitative overlap between languages is high,

ranging from 94%–98%. This holds even across

distinctive pairs such as JA-EN which differ sub-

stantially in factors such as sub-word length and

distance between syntactic dependents. This strong

consistency highlights the potential for spectral pro-

files to provide language-agnostic features for task

characterization and comparison.
5 Conclusion

Linguistic information at different timescales is an, as of yet, underexplored dimension in contextualized embeddings. We propose a fully differentiable spectral probe which automatically learns to weigh frequencies that are relevant to a specific task and improves over prior fixed-band filters by capturing continuous mixtures over frequencies (Section 2). This enables us to not only outperform the manual filters while using one probe instead of five, but to also identify that high-frequency tasks still benefit from low-frequency information (Section 3.1).

Extending spectral probing to seven tasks in six languages, we trained task-specific filters which outperformed the original, unfiltered embeddings. The resulting spectral profiles furthermore shed light on how linguistic information at different timescales relates to different task types (Section 3.2). They not only match the linguistic intuitions underlying each task, but also enable quantitative comparisons between them. The analysis of the filters’ overlap surfaced a clear dichotomy between token and sequence-level tasks, but also highlighted intersecting frequency ranges which contain information relevant across task types. Finally, the language-agnostic nature of these spectral profiles highlights future avenues towards more robust task descriptors (Section 4).

Limitations

Our experiments cover a diverse, but non-exhaustive set of NLP tasks and languages. While more extensive than prior related work (Tenney et al., 2019; Tamkin et al., 2020), we elaborate in the following regarding the motivation of the final setup: As the aim of our study was to investigate the cross-lingual properties of the underexplored timescale dimension of contextualized representations, the set of languages and tasks used in our experiments emphasizes consistency across languages. This limits us to high-resource languages for which datasets covering every task are available. However, with cross-lingual stability confirmed in our experiments, the study of lower-resourced languages is a clear avenue for future research.

Despite using a set of well-established datasets, it is important to keep data quality in mind when interpreting the results—even for these high-resource languages. In our initial exploratory data analyses, we identified and confirmed limitations known to the original dataset authors in that many include silver, or weakly filtered annotations driven by automatic matching and translation (e.g., WikiANN, XNLI, JSNLI). As we are less interested in benchmarking performance and rather focus on the feasibility and analysis of our spectral profiles, individual data instances of lesser quality should however be of limited concern. Appendix A details how each dataset was constructed originally, and also how it was pre-processed by us, such that results can be interpreted in the appropriate context.

In terms of modeling, we hope that future work will investigate spectral probes and their resulting task profiles across more encoder models with different architectures and pre-training strategies. Finally, while we have demonstrated spectral profiles to be suitable for characterizing different tasks consistently across languages, future research could supplement them with other descriptors such as embedding layer depth in order to identify even more distinctive profiles.

Ethics Statement

Given the theoretical nature and wide applicability of this work—both in terms of data domains and model architectures—it is difficult to anticipate broader impacts and future ethical implications. In general, benefits and harms in the field of probing originate from the information being investigated: While we are interested in linguistic timescale characteristics, probe-like methods have also been applied to protected attributes of data subjects in order to, for example, de-bias LMs (Ravfogel et al., 2020). Since this process involves personal information, any experiments extracting such characteristics should be sufficiently vetted for ethical acceptability. With spectral profiles being a relatively broad descriptor however, we do not expect them to identify frequencies exclusive to personal information or to replace existing, domain-specific probing methods.

Acknowledgements

Thanks to NLPnorth and MaiNLP Lab for their insightful feedback, particularly Elisa Bassignana, Joris Baan, Mike Zhang, as well as to ITU’s High-performance Computing Team. Additional thanks to the anonymous reviewers for their helpful comments. MME and BP are supported by the Independent Research Fund Denmark (DFF) Sapere Aude grant 9063-00077B. BP is supported by the ERC Consolidator Grant DIALECT 101043235.
References

Nasir Ahmed, T. Natarajan, and Kamisetty Ramamohan Rao. 1974. *Discrete cosine transform*. *IEEE Transactions on Computers*, C-23(1):90–93.

Guillaume Alain and Yoshua Bengio. 2017. *Understanding intermediate layers using linear classifier probes*. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017*, Workshop Track Proceedings. OpenReview.net.

Masayuki Asahara, Hiroshi Kanayama, Takaaki Tanaka, Yusuke Miyao, Sumire Uematsu, Shinsuke Mori, Yuji Matsumoto, Mai Omura, and Yugo Murawaki. 2018. *Universal Dependencies version 2 for Japanese*. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. *A large annotated corpus for learning natural language inference*. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. *TIGER*: Linguistic interpretation of a german corpus. *Research on language and computation*, 2(4):597–620.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. *XNLI*: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

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

Bruno Guillaume, Marie-Catherine de Marneffe, and Guy Perrier. 2019. Conversion et améliorations de corpus du Français annotés en Universal Dependencies. *Traitement Automatique des Langues*, 60(2):71–95.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Ziyang Hu. 2018. *Discrete Cosine Transform for PyTorch*. https://github.com/zh217/torch-dct.

J. D. Hunter. 2007. *Matplotlib*: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. *The multilingual Amazon reviews corpus*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. *Adam*: A method for stochastic optimization. *Computational Research Repository*, arxiv:1412.6980. Version 9.

Ken Lang. 1995. Newsweeder: Learning to filter news. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339. Retrieved from http://qwone.com/~jason/20Newsgroups/ on Feb 10th, 2022.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. *MKQA*: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. *Building a large annotated corpus of English*: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. *Cross-lingual
name tagging and linking for 282 languages. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems 32, pages 8024–8035. Curran Associates, Inc.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7237–7256, Online. Association for Computational Linguistics.

Mo Shen, Ryan McDonald, Daniel Zeman, and Peng Qi. 2016. UD_Chinese-GSD. https://github.com/UniversalDependencies/UD_Chinese-GSD.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14), pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Alex Tamkin, Dan Jurafsky, and Noah Goodman. 2020. Language through a prism: A spectral approach for multiscale language representations. In Advances in Neural Information Processing Systems 33, pages 5492–5504. Curran Associates, Inc.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 183–196, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, P случка Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Takumi Yoshikoshi, Daisuke Kawahara, and Sadao Kurohashi. 2020. Multilingualization of a natural language inference dataset using machine translation. Proceedings of the 244th Meeting of Natural Language Processing, pages 1–8. (Translated from Japanese original).

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noémi Aepli, Hamid Aghaie, Željko Agić, Amir Ahmad, Lars Ahrenberg, Chika Kennedy Ajede, Gabriele Aleksandravičaitė, Ika Alfina, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Maria Jesus Aranzabe, Bilge Nas Arıcan, Hórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Deniz Baran Aslan, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Virginiya Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmoin, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Ke Pa Kengo txea, Gózde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biaggi Ecker, Freidick Bick, Agné Bielinskiene, Kristín Bjarnardóttir, Rogier Blokland, Victoria Bobicev, Loic Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gospa Bouma, Sam Bowman, Adrianne Boyd, Anouck Braggaar, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Lauren Cassidy, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomir Čepšl, Néslhan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabri- cio Chalub, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Mihaela Cristescu, Phile- mon Daniel, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oğuz De- rin, Elvis de Souza, Arantza Diaz de Ibarra, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Do- brovoljč, Timothy Dozat, Kira Droaganova, Puncet Dwivedi, Hanne Ebbing, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Abdurahman Esfandiar, Mohammad Esfandiar, Valérie Eudebert, Aimee Evers, Luigi Fabiano, Joo Lee, Frits Eikema, Cédric Faure, Théo Favre, Katarina Gajdošová, Daniel Galbraith, Marcos García, Moa Gärdenfors, Sebastian Garza, Fabrizio Fer- raz Gerardi, Kim Gerdes, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guino-
Appendix

A Data Setup

In the following, we provide details about the versions, splits and pre-processing of each dataset. Additionally, we present example instances together with their token/sequence-level annotations in Table 2 (in English, where available). In our experiments, each model is tuned on the training split and only evaluated on the validation split as we are not interested in obtaining state-of-the-art results, but rather aim to analyze overall performance patterns across tasks. We use the original splits where provided and generate our own otherwise.

Penn Treebank (Marcus et al., 1993) We use Penn Treebank version 2 (PTB) as published in OntoNotes 4.0. Sections 02-21 were used for training, section 22 for validation, and section 23 for test, totaling 30,060, 1,336 and 1,640 instances respectively. The label space covers 48 part-of-speech tags. Note that Tamkin et al. (2020) use PTB version 3 in their experiments which we were unable to obtain due to licensing constraints. As such the exact data and splits may differ.

Universal Dependencies (Zeman et al., 2021) From Universal Dependencies version 2.9 (UD), we select the following treebanks: German-GSD (Brants et al., 2004), English-EWT (Silveira et al., 2014), Spanish-GSD (McDonald et al., 2013), French-GSD (Guillaume et al., 2019), Japanese-GSD (Asahara et al., 2018), Chinese-GSD (Shen et al., 2016) with standard splits, totaling 66,040 training and 6,683 validation instances. The label set comprises the 17 UPOS classes and the 36 dependency relations which can occur between a word and its head.

WikiANN (Pan et al., 2017) This dataset contains silver NER data for 282 languages which was extracted from Wikipedia using URL references as a proxy for named entities. It contains the entity types location (LOC), person (PER) and organization (ORG) which are annotated in BIO-format. Our experiments use the existing data splits with 20,000 training and 10,000 validation instances.

MKQA (Longpre et al., 2021) Multilingual Knowledge Questions and Answer (MKQA) is an open-domain question answering dataset which covers 10,000 questions and their corresponding answers in an aligned corpus spanning 26 languages. After removing unanswerable questions, we use each correct QA pair to generate an additional incorrect pair for the same question, yielding a total set of 13,516 instances used in our experiments. To generate an incorrect answer, we sample an alternative answer of the same type (e.g., time, number) which does not equal the correct answer. Finally, we randomly split the data 80/20 into training and validation portions for which the instances are aligned across languages (i.e., the same questions and answers). The final task is a binary classification task for whether a QA pair is true or false, with a random baseline of 50%.

Multilingual Amazon Reviews (Keung et al., 2020) MAR are used for both sentiment analysis and topic classification. For Senti, we convert the 1–5 star rating into $\{1, 2\} \rightarrow \text{negative}, \{3\} \rightarrow \text{neutral} \text{ and } \{4, 5\} \rightarrow \text{positive}$. For Topi, we consider the 30 product categories as topics. All original splits are kept, resulting in 200,000 training and 5,000 validation instances per language.

20 Newsgroups (Lang, 1995) This dataset contains English emails from 20 newsgroups and their corresponding topics. We use the bydate-version which is sorted by date and removes duplicate entries and email headers (which give away the topic). Of the official training and testing data, we subdivide the former 11,314 instances into an 80/20 train/validation split. Note that there may differences to the version used in Tamkin et al. (2020) due to alternative splitting strategies.

XNLI (Conneau et al., 2018) The Cross-lingual Natural Language Inference (XNLI) dataset covers 15 languages translated from and including English (as it lacks Japanese data, we supplement it with JSNLI). The task is to identify the relation between a premise and a hypothesis as: contradiction, entailment or neutral. Our setups use the original training and validation splits with 392,702 and 2,490 input pairs respectively.

JSNLI (Yoshikoshi et al., 2020) This dataset contains premise-hypothesis pairs from the Stanford Natural Language Inference corpus (Bowman et al., 2015) which were translated automatically into Japanese and filtered for correctness. It contains 533,005 training and 3,916 validation instances with the same three classes as XNLI.
**TOKEN-LEVEL TASKS**

|   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|
| PTB | In | Tokyo, | trading is halted during lunchtime. |
| UD | Can rabbits and chickens live together? |
| WikiANN | The Zeros formed in Chula Vista in 1976. |

**SEQUENCE-LEVEL TASKS**

|   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|
| MKQA | when did love become a part of marriage? | 18th century | 1 (true) |
| AMR | All socks had large holes after a few months. | apparel negative |
| 20News | [...] Does anyone know how to size cold gas roll control thruster tanks for sounding rockets? [...] | sci.space |
| XNLI | I’ve got more than a job. | I don’t have a job or any hobby. | contradiction |
| JSNLI | 地下鉄を待っている間に本を読む男。 | 男は地下にいる。 | entailment |

|   |   |   |
|---|---|---|
| The man reads a book while waiting for the subway. | The man is underground. |

Table 2: Example Dataset Instances annotated with respective token/sequence-level labels.

**B Experiment Setup**

**Models** In the monolingual English experiments, we use bert-base-cased (Devlin et al., 2019; BERT) following Tamkin et al. (2020). For the multilingual experiments we use bert-base-multilingual-cased (Devlin et al., 2019; mBERT). For both LMs, we use respective checkpoints from the Transformer library’s model hub (Wolf et al., 2020).

Manual, fixed-band filters as well as the automatically learned filters are applied to the contextualized embeddings produced by the last layer of either model. As visualized in Figure 1, we decompose the sequence of values from each embedding dimension (768 in both LMs) using the DCT (Ahmed et al., 1974; DCT-II), weight the appropriate $k$ by a fixed amount or by the learned weight in $\gamma$, before applying the IDCT to reconstruct a sequence of real values. These make up each dimension of the filtered embeddings.

Following Tamkin et al. (2020), the original/filtered embeddings are passed to a linear probe (Alain and Bengio, 2017) consisting of two parameters: a transformation $W \in \mathbb{R}^{E \times C}$ and a bias $b \in \mathbb{R}^C$, where $E$ is the embedding dimension and $C$ is the number of classes specific to each task.

**Training** As we run probing experiments, neither the 108M-parameter BERT, nor the 178M-parameter mBERT are fine-tuned. We only train the linear probe which has 1,538–36,912 parameters depending on the task, plus the 512 parameters of the learned spectral filter $\gamma$. As in Tamkin et al. (2020), we use the Adam optimizer (Kingma and Ba, 2014) with a learning rate of $10^{-3}$ which decays by 0.5 every time the loss plateaus. Updates are applied in batches of size 32 across a maximum of 30 epochs, with an early stopping patience of 1. Each setup is run with the five random seeds: 1932, 2771, 7308, 8119, 9095. On our hardware consisting of an NVIDIA A100 GPU with 40GBs of VRAM and an AMD Epyc 7662 CPU, training a probe takes approximately 10 minutes.

**Evaluation** In order to probe a sequence of contextualized embeddings for information at different timescales, it is necessary to apply each filter at the sub-word level. To measure the effect of different frequencies, we follow Tamkin et al. (2020) and evaluate all tasks using accuracy (ACC) on the sub-word level. Note that for token-level tasks each token label is therefore repeated across all of its sub-words, while for sequence-level tasks, each sub-word is classified with the label of its sequence.

7739
Implementation  All models are implemented using PyTorch v1.10 (Paszke et al., 2019) and NumPy v1.22 (Harris et al., 2020). Additionally, we use a modified version of the torch-dct package (Hu, 2018) to perform the DCT and IDCT. Visualizations are generated using matplotlib v3.5 (Hunter, 2007). Further, the code for reproducing our experiments is available at https://github.com/mainlp/spectral-probing.

C  Detailed Results

The following supplements the results presented in Section 3 with more detailed scores. Table 3 lists the exact scores for the monolingual English experiments on PoS and T0PIC using the ORIG embeddings, the fixed-band filters and the learned AUTO filter. Table 4 lists the detailed scores for the ORIG and AUTO-filtered embeddings per language, in addition to the cross-lingual mean and standard deviation, across our seven tasks.

While the scores across random initializations never exceed a standard deviation of 1.0, it is important to note that scores may have higher variance across languages. This is to be expected due to different data across languages as well as pre-training availability. However we note that overall performance patterns (i.e., higher AUTO and relative task performance) are consistent across languages.
Table 3: **Detailed Monolingual Results** (ACC) for unfiltered (ORIG), low (L), mid-low (ML), mid (M), mid-high (MH), high (H), and automatically learned filters (AUTO), on the tasks of POS-tagging and TOPIC classification. Reported are the mean over five random initializations ± standard deviations. The same results plus the spectral profiles (frequency weightings) learned by AUTO are plotted in Figure 2.

| TASK | ORIG | LOW | MID-LOW | MID | MID-HIGH | HIGH | AUTO |
|------|------|-----|---------|-----|----------|------|------|
| POS  | 95.8±0.1 | 21.9±0.0 | 21.8±0.1 | 26.2±0.1 | 48.6±0.1 | 90.6±0.0 | 95.9±0.0 |
| TOPIC | 41.3±0.2 | 71.2±0.4 | 18.4±0.3 | 5.6±0.3 | 5.6±0.3 | 5.6±0.4 | 72.1±0.3 |

Table 4: **Detailed Multilingual Results** (ACC) for unfiltered (ORIG) and automatically learned filters (AUTO) on the tasks of POS-tagging, dependency relation classification (DEP), named entity recognition (NER), question answering (QA), sentiment analysis (SENTI), TOPIC classification, and natural language inference (NLI). Each task covers the languages German (DE), English (EN), Spanish (ES), French (FR), Japanese (JA) and Chinese (ZH). Reported are the mean over five random initializations ± standard deviations as well as the mean over languages (AVG) ± the standard deviation across languages. The latter results are reported in Table 1, in addition to the spectral profiles (frequency weightings) learned by AUTO in Figure 3.

| TASK | EMB | DE | EN | ES | FR | JA | ZH | AVG |
|------|-----|----|----|----|----|----|----|-----|
| POS  | ORIG | 92.0±0.0 | 91.6±0.1 | 93.8±0.0 | 95.1±0.1 | 92.5±0.0 | 89.5±0.1 | 92.4±1.9 |
| AUTO | 92.1±0.1 | 91.6±0.0 | 93.9±0.0 | 95.1±0.0 | 92.7±0.1 | 89.8±0.1 | 92.5±1.8 |
| DEP  | ORIG | 79.0±0.1 | 78.4±0.1 | 81.2±0.1 | 83.0±0.1 | 79.6±0.1 | 70.6±0.2 | 78.6±4.3 |
| AUTO | 79.5±0.2 | 78.4±0.1 | 81.8±0.1 | 83.8±0.1 | 80.8±0.1 | 71.3±0.2 | 79.3±4.3 |
| NER  | ORIG | 90.3±0.0 | 85.3±0.1 | 90.4±0.0 | 88.1±0.1 | 84.1±0.1 | 89.5±0.0 | 88.0±2.7 |
| AUTO | 90.4±0.0 | 85.5±0.1 | 90.5±0.0 | 88.3±0.0 | 84.4±0.1 | 89.7±0.0 | 88.1±2.6 |
| QA   | ORIG | 63.2±0.2 | 64.5±0.1 | 64.1±0.2 | 63.9±0.3 | 61.0±0.8 | 60.7±0.8 | 62.9±1.6 |
| AUTO | 66.8±0.1 | 68.1±0.5 | 67.9±0.2 | 68.1±0.2 | 65.1±0.1 | 66.1±0.4 | 67.1±1.2 |
| SENTI| ORIG | 56.0±0.2 | 57.1±0.2 | 58.7±0.2 | 57.1±0.2 | 57.2±0.2 | 58.0±0.2 | 57.4±0.9 |
| AUTO | 64.0±0.2 | 63.5±0.5 | 65.4±0.2 | 64.7±0.5 | 65.4±0.5 | 62.7±0.3 | 64.3±1.1 |
| TOPIC| ORIG | 22.7±0.1 | 26.8±0.4 | 22.9±0.3 | 24.0±0.3 | 22.9±0.5 | 43.3±0.4 | 27.1±8.1 |
| AUTO | 34.3±0.7 | 39.8±0.4 | 30.2±0.2 | 30.7±0.4 | 35.8±0.5 | 52.3±0.5 | 37.2±8.2 |
| NLI  | ORIG | 41.5±0.2 | 43.6±0.3 | 43.2±0.2 | 42.7±0.2 | 52.3±0.3 | 41.7±0.2 | 44.1±4.1 |
| AUTO | 51.3±0.8 | 56.4±0.7 | 54.3±0.7 | 54.5±0.5 | 67.2±0.4 | 53.8±1.0 | 56.3±5.6 |