MorSeD: Morphological Segmentation of Danish and its Effect on Language Modeling

Rob van der Goot[1] **Anette Jensen** **Emil Allerslev Schledermann**[1]
Mikkel Wildner Kildeberg[1] **Nicolaj Larsen**[1] **Mike Zhang**[2] **Elisa Bassignana**[1]

[1]IT University of Copenhagen
[2]Aalborg University
robv@itu.dk

Abstract

Current language models (LMs) mostly exploit subwords as input units based on statistical co-occurrences of characters. Adjacently, previous work has shown that modeling morphemes can aid performance for Natural Language Processing (NLP) models. However, morphemes are challenging to obtain as there is no annotated data in most languages. In this work, we release a wide-coverage Danish morphological segmentation evaluation set. We evaluate a range of unsupervised token segmenters and evaluate the downstream effect of using morphemes as input units for transformer-based LMs. Our results show that popular subword algorithms perform poorly on this task, scoring at most an F_1 of 57.6 compared to 68.0 for an unsupervised morphological segmenter (Morfessor). Furthermore, evaluate a range of segmenters on the task of language modeling.[1]

Input:	frakkeskåner	lærte
MorSeD:	frakke-skån-er	lær-te
TinyBERT:	fra-kk-es-kan-er	l-æ-rte
BPE:	frakke-skå-ner	lærte
WordPiece:	fra-kke-sk-åne-r	lærte
Unigram:	fra-kke-skån-er	lærte
Morfessor:	frakke-skån-er	lært-e

Figure 1: Two examples from our dataset, with the input words, gold morpheme annotation (morsed), and the outputs of: a baseline English language model segmenter (TinyBERT), three Danish statistical segmenters, and a Danish unsupervised morphological segmenter (Morfessor).

1 Introduction

Although there is no exact consensus on the definition of morphemes (e.g. Nida, 1948; Bolinger, 1948), they are commonly described as the smallest meaning-carrying units in natural language (Sinclair, 1996). Morphemes are useful for linguistic analysis, language understanding, language learning and potentially as input units for NLP models. Traditionally, characters or words were used as inputs for NLP models, but contextualized Language Models (LMs) popularized subwords (Devlin et al., 2019), which are often based on a trained vocabulary obtained with statistical methods. Morphemes, however, are a promising alternative as they are of similar granularity but are linguistically motivated. In NLP, morphemes have been successfully used in machine translation models (Clifton and Sarkar, 2011; Popović, 2012), RNN LMs (Blevins and Zettlemoyer, 2019; Schwartz et al., 2020), for static word embeddings (Üstün et al., 2018), and as an auxiliary task in character-level models (Matthews et al., 2018).

Although there have been large multilingual benchmarking efforts for morphological tagging (Zeman et al., 2018) and reinflection (Cotterell et al., 2018), data for morphological segmentation is more scarce, Especially for mid-resource languages, like Danish (Joshi et al., 2020). Therefore, we create a small yet high-coverage benchmark to evaluate unsupervised segmenters for Danish morphological segmentation and provide an extensive evaluation of existing models.

There has been some work that incorporating morphemes as input to LMs. For English, Hofmann et al. (2021) showed that derivational segmentation aids LM interpretation of complex words, and Bostrom and Durrett (2020) showed that using units that closer resemble morphemes improves language modeling (although the mor-

[1]Data and code are available on https://bitbucket.org/robvanderg/morsed

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 223–229
March 3-4, 2025 ©2025 University of Tartu Library

TYPE	DESCRIPTION
Root Morphemes	The root of a word is its stem, the shortest meaning-bearing part. A root is also called a free morpheme, as it makes sense on its own and often has a concrete meaning.
Compounds	New words in Danish can be formed by combining existing words, creating new meanings. These are compound words and are considered complex. Many compounds are formed solely from root morphemes, which are often nouns, but also adverbs and adjectives.
Compounds with Linking	Some roots in compound words are connected using linking letters, commonly "-e" and "-s." Linking letters are often used when the first root is a verb.
Prefixes	A prefix is a derivative added to the beginning of a word, altering its meaning but not its word class. Prefixes cannot form words on their own.
Suffixes	A suffix is also a derivative, added to the end of a word, typically changing its word class. Like prefixes, they cannot form words on their own.
Inflections	Inflectional morphemes are mainly associated with nouns, verbs, and adjectives. They add information such as gender, definiteness, tense, and mood, but do not form words independently.

Table 1: Description of each type of morphological segmentation we use in our study.

phemes are of relatively low accuracy). Limisiewicz et al. (2024) use morphemes in a multilingual LM. They transforms unsupervised morphemes to byte sequences which are used as input sequence to an LM, but they do not evaluate the quality of the morphemes. Our work differs by focusing on Danish, including a wider range of morphemes, evaluating more segmenters, evaluating morpheme performance, and obtaining inputs closer to true morphemes.

Our contributions are: ① We present MORSED, an evaluation dataset for Danish morphological segmentation, including morpheme-level categories and labels. ② We evaluate various segmenters on the task of morphological segmentation: 3 subword algorithms and an unsupervised morphological segmenter. ③ We examine the impact of training data and vocabulary size on tokenizers by training them on 11 different data sources. ④ We assess our tokenizers for language model training using small discriminative transformer-based models.

2 MORSED

Here, we introduce MORSED, to the best of knowledge the first publicly available dataset annotated for morphological segmentation of Danish. We follow the guidelines and categories defined in Jensen (2021). Our main annotator (author of Jensen (2021)) has 35 years of experience as a Danish teacher, with a degree in Teaching and a postgraduate diploma in Adult Literacy Education. The dataset contains 800 words.[2] The words were selected by our main annotator, focusing on diversity and good coverage for each category. In Table 1, we describe each type of morphological segmentation.

A second native Danish annotator without a linguistic background annotated 300 words from MORSED by following the same guidelines. Since inter-annotator scores (e.g., Cohen's Kappa) are challenging to compute for segmentation tasks, we use F_1 score on the morpheme level for comparison. The resulting F_1 is 0.991, indicating well-defined guidelines and a clear task definition.

3 Setup

Segmentation Methods. We adopt (1) BPE (Shibata et al., 1999), which merges frequent character pairs into subwords until a fixed vocabulary size is reached; (2) WordPiece (Sennrich et al., 2016), which iteratively builds subwords based on likelihood, optimizing for unseen words; (3) Unigram (Kudo, 2018), which applies a probabilistic model to select the best subword units from an initial large set; and (4) Morfessor (Virpioja et al., 2013), which uses methods for unsupervised learning to perform morphological segmentation. We compare these segmenters to the Leave-As-Is (LAI) baseline, which simply returns the word unchanged.

Raw Text Data. For training the segmenters and the LMs, we use raw text data. We collect data from 8 different resources (Table 2). We filter the

[2]Morphological segmentation/labeling datasets are typically smaller than other NLP datasets, even for English. We believe that due to the diversity of selected words and the relatively morphological simplicity of Danish, the variety of phenomena within each category is well-represented in our data.

DATASET	DOMAIN	SOURCE
Bookshop	Books	Tiedemann (2012)
CC-100	Webscrape	Wenzek et al. (2020)
CulturaX	Webscrape	Nguyen et al. (2023)
Gigaword	Mixed	Strømberg-Derczynski et al. (2021)
OpenSubtitles[5]	Subtitles	Lison and Tiedemann (2016)
Reddit	Social	Chang et al. (2020)
Twitter	Social	`archive.org/details/twitterstream`
Wiki	Wiki	Attardi (2015)

Table 2: List of datasets. From the multi-lingual datasets, we only consider the Danish part.

data using the FastText language classifier (Joulin et al., 2017)[3] and shuffle the lines before taking the first 40M characters from each source. With these, we create two multi-domain datasets of 40M and 320M characters respectively by evenly mixing the 8 individual datasets.

Language model evaluation Due to computational constraints, we choose to train a model with the same architecture as TinyBERT (Jiao et al., 2020). We did a hyperparameter search with its default tokenizer on the English data from the BabyLM challenge (Warstadt et al., 2023) to find reasonable settings (details are available in the repository).[4] We use the Adam optimizer, with a learning rate of 1×10^{-3}, a batch size of 512, and 1 epoch over the mixed 320M dataset (Section 2), of which we keep 1% separate for evaluation.

We use a 15% masking strategy during training and evaluation, because perplexity is affected by the segmentation. We use Bits Per Character (BPC) to evaluate the language models. Bits per character represents the average number of bits needed to encode each character in the dataset. Furthermore, we use accuracy on the token level. Even though the accuracy is affected by the segmentation, it is highly interpretable, and since none of our models is tuned to optimize on this metric we expect it to correlate to language model quality.

4 Results

4.1 Morphological Segmentation.

Although there is a variety of metrics available for evaluating morphological segmentation (Virpioja et al., 2011), we opt for the interpretable precision, recall, and F_1 score based on found morphemes (not split points). We start with finding the best

[3] We keep all text with a confidence above .6 for Danish.

[4] We did this on English, as there is more consensus on which tokenizer/data to use.

[5] `http://www.opensubtitles.org/`

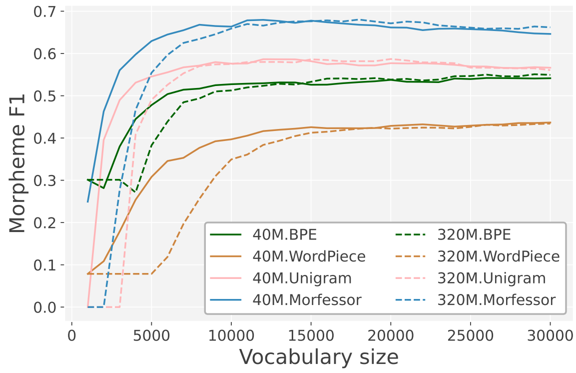

Figure 2: F_1 of each algorithm for different vocabulary sizes for the multi-domain dataset.

vocabulary size of each segmenter on the mixed datasets, as it has the broadest coverage, and then we compare the effect of training on each individual data source.

We evaluate a vocabulary size of 1K-30K subwords with intervals of 1K (Figure 2). Results show that performance for all algorithms follows a similar trend; performance improves strongly in the beginning (i.e., small vocabulary size), until a size of around 10K, after which performance remains in a similar range. For Morfessor and Unigram performance slowly drops, while for BPE and WordPiece it remains rather stable. Morfessor outperforms the other segmenters by a large margin, scoring a maximum F_1 of 67.96, showing that the task is still far from unsolved.[6] The segmenters trained on 320M characters often perform slightly worse compared to the 40M character training data (especially for smaller vocabulary sizes). In the following sections, we use 40M characters for training segmenters, and use the best vocabulary size for each method: BPE 26K, WordPiece 30K, Unigram 11K, Morfessor 12K.

Next, we compare the effect of the data source on the performance of the segmenters (Figure 3). Results show that while the mixed dataset leads to robust performance across segmenters , different segmenters have different best-performing datasets. As MorSeD is composed of well-formed, general-domain words, we would expect that corpora that resemble this (i.e., books, subtitles, wiki, subtitles corpora) would lead to better performance. This trend is loosely reflected in the scores, as the Twitter and Reddit dataset per-

[6] It should be noted that higher scores can be obtained in (partially) supervised settings (Kohonen et al., 2010).

Model	Root	Comp.	Link.	Pref.	MorSeD Suff.	Infl.	Prec.	Rec.	F1	Acc.	MELFO F1	Lang. Modeling ↓BPC	Acc.
TinyBERT	48.40	16.64	7.76	20.32	**29.43**	15.12	27.60	29.27	28.41	14.00	11.74	9.84	3.12
LAI	**100.00**	0.66	15.83	4.42	1.12	12.10	23.33	57.45	33.18	32.25	3.68		
BPE	90.42	45.85	24.45	30.93	10.80	9.37	47.91	62.39	54.20	46.50	25.79	5.25	4.11
WordPiece	83.23	23.93	9.85	13.96	8.94	8.35	38.88	49.81	43.67	26.00	12.87	**3.62**	27.37
Unigram	82.37	54.82	46.20	39.65	17.29	21.16	53.02	63.13	57.63	46.12	35.20	5.96	5.41
Morfessor	87.93	**68.41**	**50.09**	**56.86**	22.40	**44.03**	**65.00**	**71.20**	**67.96**	**59.75**	**44.06**	6.98	**54.04**

Table 3: Metrics for Language Modeling and Morphological Segmentation. For the language modeling experiments, we show BPC and accuracy (Acc.). For the morphological segmentation experiments on MorSeD, we show performance in F_1 on Root morphemes (Root), Compounds (Comp.), Linking elements (Link.), Prefixes (Pref.), Suffixes (Suff.), Inflections (Infl.) and average performance over the whole dataset: Precision (Prec.), Recall (Rec.), F1 on morphemes, Accuracy (Acc.) on the word level.

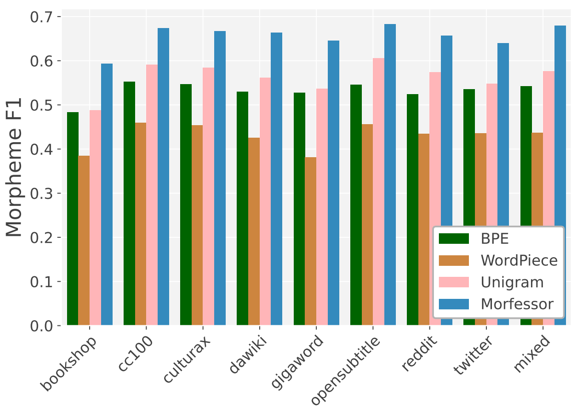

Figure 3: Comparison of the effect of data source, all with 40M characters, and the best vocabulary size for each algorithm.

form relatively poor. However, the Bookshop and FTSpeech also leads to quite low performance, which is probably due to topic bias, FTSpeech contains parlemental data and Bookshop contains quite some technical data (e.g., legal and political topics), which leads to a larger coverage of domain-specific words, but lower performance on MorSeD.

4.2 Language Modeling.

For each segmentation algorithm, we used the segmenter trained on the mixed dataset (40M) with the best size from the morphological segmentation results (Section 4) for evaluation on language modeling (Table 3, Language Modeling column). The BPC scores of the Danish tokenizers outperform the original TinyBERT tokenizer (9.2) trained on the Danish corpus. Across the Danish tokenizers, the BPC scores show minimal variance, with the WordPiece tokenizer achieving the best score of 3.62. Morfessor shows a higher BPC score than the other tokenizers (6.98). We hypothesize that, since BPC correlates directly with cross-entropy, Morfessor's more granular "subword" units (morphemes) lead to less probability mass being concentrated on the most likely token. This results in higher entropy, as the model distributes the probability mass across a larger set of possible tokens, reducing certainty in its predictions. Manual inspection of the output distributions revealed that the Morfessor based language model more often has the correct candidate ranked high, but its confidence scores are less well aligned (i.e. more often scores ¿0.5 for incorrect predictions, and lower scores for the best candidate when it is correct). Therefore, we also calculate the subword (i.e. morpheme) accuracy, where only the highest ranking candidate is used. Our results show that the Morfessor tokenizer achieves the highest accuracy by a large margin, indicating that it performs best among all models.

5 Analysis

Quantitative. Our results show that recall is higher than precision for all methods (Table 3). This indicates that most models under-segment. The difference between accuracy and F_1 score (between 6-8 absolute points) shows that there are cases where a word is segmented partially correct.

Models perform especially well on root morphemes, which are not segmented in our task definition (Section 2). A clear trend is that Morfessor and Unigram underperform on root morphemes, but perform better on the other categories. This is because of their smaller optimal vocabulary size (12,000 and 11,000 versus 26,000 for BPE and 30,000 for WordPiece), which leads to oversplitting on the root morphemes Overall, Morfessor outperforms all other segmenters on all classes except root morphemes and suffixes. For the latter, TinyBERT performs better on some word-endings that overlap with English (e.g. '-er', '-ing'), which are kept attached to the words by Morfessor.

Qualitative. To get a more fine-grained picture of the difficulties for the segmentation models, we spot-check cases where at least three of the segmenters were incorrect. Our analysis reveals that tokenizers frequently missegment in the categories *compounds* and *compounds with linking elements*. The segmentation of morphemes such as "-e" and "-s" is especially challenging, underscoring tokenizers' difficulties with complex morphological structures such as "sygeplejeskole" (syg-e-plej-e-skole; en: "nursing school"), "gulerod" (gul-e-rod; en: "carrot") and "landsholdstrup" (land-s-hold-s-trup; en: "national team"). Furthermore, as morpheme length increases, the error rate increases, highlighting the tokenizers' limitations in handling more complex word formations.

MELFO data After our experiments, we managed to get access to morphological segmentation data from the MELFO (Mobil e-læring for ordblinde) project[7]. This data is not publicly available, but we used it to evaluate the robustness of each segmenter on another dataset with different guidelines and annotators. Upon manual inspection, we found that the main difference between the datasets is the choice of words (there are 8 overlapping words) and that the segmentation of MORSED leads to more splits and smaller elements (e.g. fri-tid-s-hjem versus fritid-s-hjem). The results show a similar trend (i.e. ranking of models), but lower performances overall, which is partially due to tuning (of vocabulary size) on MORSED, but also due to the structure of the data: MELFO has a longer average word length (12 characters versus 8) and a larger average amount of morphemes per word (2.6 versus 1.9).

6 Conclusion

We introduced MORSED, a broad-coverage, expert-annotated dataset for subword segmentation in Danish. We used MORSED to show that an unsupervised segmenter outperforms statistical-based subword segmenters on the task of morphological segmentation for Danish by 10.3 points absolute F_1 score on our novel Danish benchmark. We also show that the tokenizer that performs best at morphological segmentation also performs well on language modeling (accuracy).

[7] https://laes.hum.ku.dk/centerets_forskning/melfo/

Acknowledgments

We would like to thank Arzu Burcu Güven for her feedback. We thank Bart Jongejan for sharing the MELFO data. We acknowledge the IT University of Copenhagen HPC resources made available for conducting the research reported in this paper. Mike Zhang is supported by a research grant (VIL57392) from VILLUM FONDEN. Elisa Bassignana is supported by a research grant (VIL59826) from VILLUM FONDEN.

References

Giuseppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Terra Blevins and Luke Zettlemoyer. 2019. Better character language modeling through morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1606–1613, Florence, Italy. Association for Computational Linguistics.

Dwight L Bolinger. 1948. On defining the morpheme. *Word*, 4(1):18–23.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.

Ann Clifton and Anoop Sarkar. 2011. Combining morpheme-based machine translation with post-processing morpheme prediction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 32–42, Portland, Oregon, USA. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the*

CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection, pages 1–27, Brussels. 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.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves BERT's interpretation of complex words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3594–3608, Online. Association for Computational Linguistics.

Anette Jensen. 2021. *Morfemer*. Gyldendal Uddannelse.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Oskar Kohonen, Sami Virpioja, and Krista Lagus. 2010. Semi-supervised learning of concatenative morphology. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 78–86, Uppsala, Sweden. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Tomasz Limisiewicz, Terra Blevins, Hila Gonen, Orevaoghene Ahia, and Luke Zettlemoyer. 2024. MYTE: Morphology-driven byte encoding for better and fairer multilingual language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15059–15076, Bangkok, Thailand. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Austin Matthews, Graham Neubig, and Chris Dyer. 2018. Using morphological knowledge in open-vocabulary neural language models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1435–1445, New Orleans, Louisiana. Association for Computational Linguistics.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *ArXiv preprint*, abs/2309.09400.

Eugene A Nida. 1948. The identification of morphemes. *Language*, 24(4):414–441.

Maja Popović. 2012. Morpheme- and POS-based IBM1 and language model scores for translation quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 133–137, Montréal, Canada. Association for Computational Linguistics.

Lane Schwartz, Francis Tyers, Lori Levin, Christo Kirov, Patrick Littell, Chi-kiu Lo, Emily Prud'hommeaux, Hyunji Hayley Park, Kenneth Steimel, Rebecca Knowles, et al. 2020. Neural polysynthetic language modelling. *ArXiv preprint*, abs/2005.05477.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Yusuxke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. 1999. Byte pair encoding: A text compression scheme that accelerates pattern matching. *Technical Report DOI-TR-161, Department of Informatics, Kyushu University*.

John Sinclair. 1996. The search for units of meaning. *Textus*, 9(1):75–106.

Leon Strømberg-Derczynski, Manuel Ciosici, Rebekah Baglini, Morten H. Christiansen, Jacob Aarup Dalsgaard, Riccardo Fusaroli, Peter Juel Henrichsen, Rasmus Hvingelby, Andreas Kirkedal, Alex Speed Kjeldsen, Claus Ladefoged, Finn Årup Nielsen, Jens Madsen, Malte Lau Petersen, Jonathan Hvithamar Rystrøm, and Daniel Varab. 2021. The Danish Gigaword corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 413–421, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ahmet Üstün, Murathan Kurfalı, and Burcu Can. 2018. Characters or morphemes: How to represent words? In *Proceedings of the Third Workshop on Representation Learning for NLP*, pages 144–153, Melbourne, Australia. Association for Computational Linguistics.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, Mikko Kurimo, et al. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline. *Aalto University publication series SCIENCE + TECHNOLOGY*.

Sami Virpioja, Ville T. Turunen, Sebastian Spiegler, Oskar Kohonen, and Mikko Kurimo. 2011. Empirical comparison of evaluation methods for unsupervised learning of morphology. *Traitement Automatique des Langues*, 52(2):45–90.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.