Subword Segmentation in LLMs: Looking at Inflection and Consistency

Marion Di Marco[1] and Alexander Fraser[1,2]
[1] School of Computation, Information and Technology, Technische Universität München (TUM)
[2] Munich Center for Machine Learning
{marion.dimarco|alexander.fraser}@tum.de

Abstract

The role of subword segmentation in relation to capturing morphological patterns in LLMs is currently not well explored. Ideally, one would train large models like GPT using various segmentations and evaluate how well word meanings are captured. Since this is not computationally feasible, we group words according to their segmentation properties and compare how well a model can solve a linguistic task for these groups. We study two criteria: (i) adherence to morpheme boundaries and (ii) the segmentation consistency of the different inflected forms of a lemma. We select word forms with high and low values for these criteria and carry out experiments on GPT- 4o's ability to capture verbal inflection for 10 languages. Our results indicate that in particular the criterion of segmentation consistency can help to predict the model's ability to recognize and generate the lemma from an inflected form, providing evidence that subword segmentation is relevant.

1 Introduction

The linguistic abilities of large language models have been studied to a large extent, with many new abilities emerging as language models become ever larger and more powerful. While areas such as lexico- syntactic understanding, text generation and reasoning abilities have received much attention, morphology has only played a minor role, despite being of great interest to the NLP community.

Conceptually, the morphological abilities of a model are tightly linked to the internal representation of subwords: LLMs do not operate on complete words, but instead, most words are broken into subword pieces for better computational efficiency and to handle unknown words. Subword segmentation strategies typically rely on frequency statistics and are not linguistically guided. This suggests that such segmentation strategies do not provide a suitable basis to fully capture morphology, e.g. Park et al. (2021); Hofmann et al. (2021).

Morphology relates to the construction of words, and thus represents the basis of understanding natural language. Depending on the language, morphology can play a more or less relevant role, but even in a language with rather simple morphology such as English, morphology is indispensable, whether for rare words, or for more common ones. For instance, a botanizer is a person that botanizes, a baker's workplace is a bakery and a mathematician cares about mathematics. Morphological processes are typically defined by general patterns, and, critically, understanding these patterns enables both the generation of novel words and the interpretation of previously unknown words.

Understanding the meaning of word parts in the larger context of a word, as well as the underlying patterns to compose new word forms is essential to fully comprehend language. This is particularly true for languages with complex morphology, where a larger proportion of information is encoded morphologically, leading to a comparatively high number of inflected forms that have insufficient coverage in the training data, and in the worst case do not occur in the training data at all. Despite the impressive language capabilities of LLMs, the impact of the underlying segmentation is not clear. Generally, LLMs are capable of modeling morphology and accessing morphological information, but presumably not on an ideal basis, because segmentation strategies, such as WordPiece or BPE (Schuster and Nakajima, 2012; Sennrich et al., 2016) rely on frequency- based heuristics that do not optimally capture morphological patterns.

In the following, we study two criteria, adherence to morpheme boundaries and segmentation consistency of inflected forms of a lemma. We first analyze how well these criteria are met in existing LLMs, and then investigate to what extent words which have high or low values for these two criteria affect the performance of the LLM on a linguistically interesting task.

1.1 Segmentation Problems and Criteria

There is no obvious way to determine the quality of sub- word segmentation. A simple and straightforward idea is the number of splits per word, with the underlying assumption that fewer splits suggest a "good" segmentation in contrast to a segmentation into many very short pieces. While this assumption is intuitively plausible, and an overly aggressive segmentation likely results in basically meaningless pieces, the mere number of segments does not take into account the ability to generalize and how the segmentation of one word relates to the segmentation of related words of the same inflection paradigm. For example, consider GPT4's segmentation of different forms of the German verb (ein)pfianzen: 'to plant (in)':

word	GPT
ling. sound	einpfianzen
einpif|lan|zen	ein|pfianzen
eingepfianzt	eing|ep|f|an|zt
ein|ge|pfianzt	pfianzte
pf|lan|zte	pfianzt|te
pfianzen	pf|lan|zen
pfianzen	pfianztet
pf|lan|zt|et	pfianzt|tet

The segmentation does not adhere to linguistic boundaries as, for example, neither the particle ein nor the inflectional suffixes are separated from the verb stem pfianz (plant). Another problem is that of inconsistency: inflectional variants of the same word are split differently, and thus lead to different internal representations. The table shows a linguistically sound segmentation into verb stem and the respective inflectional morphemes (the particle ein- , - ge- as part of the past participle, and different inflectional suffixes). A segmentation as proposed above is not realistic, as a comparatively small vocabulary needs to accommodate a high amount of words of different languages, and thus, lexical units cannot always be preserved. However, we can still formulate conceptually language- independent criteria, namely (i) a consistent representation for variants of closely related words and (ii) the adherence to word or morpheme boundaries; any further segmentation between these points becomes, theoretically, less relevant as the subwords of a complete word can be recomposed to obtain its representation.

text[[127, 830, 472, 929], [493, 76, 838, 206]] Intuitively, the advantages of a linguistically sound segmentation are obvious: adherence to morpheme boundaries enables an internal representation that can be shared across all observed occurrences. Similarly, the separation of inflectional affixes aims at making generalization across inflectional variants easier, which is particularly important for morphologically rich languages. A consistent representation of related words tries to achieve the same effects and is a more robust formulation: while a linguistically sound segmentation is per design consistent, the slightly simpler criterion of consistent segmentation is language- independent and less resource- intensive.

While there is a growing interest in the morphological abilities of LLMs, there is no data on the segmentation quality of existing large- scale LMs: in this work, we (i) study the segmentation of 10 different languages in GPT- 4o with respect to the two criteria outlined above and (ii) assess the impact of segmentation quality by contrasting the performance of words grouped according to these criteria on the tasks of lemma prediction and the generation of inflected forms.

2 Related Work

There is a large body of research concerning the representation of the training data of language models and translation systems: while the typical segmentation strategies are frequency- based such as WordPiece or BPE (Schuster and Nakajima, 2012; Sennrich et al., 2016), there is also evidence that these segmentation approaches are not optimal for morphologically rich languages and fail to fully capture the morphological complexities of words (Klein and Tsarfaty (2020), Park et al. (2021)). Hofmann et al. (2021) show that a linguistically grounded segmentation can improve a model's performance. Hou et al. (2023) explore the effect of subword segmentation by training Bert and GPT models on different segmentation algorithms, namely BPE and two morphological segmentation strategies. Their experiments show that morphologically guided segmentation leads to lower perplexity and faster convergence during training; their models trained on morphologically segmented data reach a similar or better performance than models trained on BPE, depending on the task. Furthermore, they find that models of smaller size trained

on morphologically segmented data can perform comparably to models of larger size trained with BPE. While not specifically studying the impact of segmentation, but instead the multilingual capabilities of English- centric LLMs, Armengol- Estapé et al. (2022) assume that the quality of subword segmentation plays a part in the performance for languages different from English, as the segmen

2023) propose to use the number of splits per word as an indicator for splitting quality, assuming that few splits per word suggest a "good" segmentation in contrast to a segmentation into many short pieces. To the best of our knowledge, there is no study that looks at segmentation criteria as outlined in this paper in combination with a linguistic task.

tation is mostly based on the predominant English vocabulary and thus not representative of many other languages. Their findings indicate that languages with more subword tokens per word tend to perform worse.

There are many variants of language- specific PLMs trained on representations to accommodate the properties of a language, (e.g. Antoun et al. (2020); Nzeyimana and Niyongabo Rubungo (2022)), mostly in a monolingual setting. Jabbar (2024) proposes a linguistically- informed representation of the training data that relies on canonical forms instead of concatenable pieces. This makes the generation step less straightforward as the pieces cannot just be concatenated, but have to be reconstructed into inflected forms. The idea to combine linguistically guided segmentation with frequency- based segmentation has also been applied to machine translation, for example Tamchyna et al. (2017); Banerjee and Bhattacharyya (2018); Mager et al. (2022), and often found to be preferable to just frequency- based segmentation. A further task linked with the representation of sub words is that of morphological reinflection (e.g Kann et al. (2017)), where an inflected form needs to be generated for a given pair of word and morphological features.

There is a growing interest in the quality of the underlying segmentation: Beinborn and Pinter (2023) look at the semantic plausibility of subword tokens; the segmentation strategy in Yehezkel and Pinter (2023) aims at incorporating context information to obtain more meaningful splits. With regard to morphology, Weissweiler et al. (2023) study the ability to create inflected forms for nonce words for typologically different languages, finding that GPT does not perform as well as systems specifically trained for morphological tasks. Soler et al. (2024) study the impact of segmentation on the quality of word representation by comparing words that are segmented with those having a dedicated embedding, i.e. unsplit words, in a word similarity task. In general, they find that the representation of split words is often worse than for non- split words. Interesting in the context of our work, their results show that a morphologically sound segmentation tends to lead to a better representation. With regard to oversplitting, their findings are mixed, but indicate that for split words, a higher number of tokens does not necessarily decrease representation quality.

Beinborn and Pinter (2023) and Weissweiler

3 Methodology

We study the quality of GPT- 40's segmentation for 10 languages (English, French, German, Spanish, Italian, Portuguese, Finnish, Swedish, Czech, Hungarian). We look at the segmentation quality from two angles: first, we examine how well inflection suffixes are separated from the stem of the word, i.e. a linguistically- oriented criterion. Second, we look at the segmentation consistency, i.e. whether all words from an inflection paradigm are segmented in a cohesive way. We assess whether the segmentation has an impact on the model performance.

In previous work on subword segmentation, either on language modeling or on machine translation, the typical approach is to compare the performance of a model trained on a baseline subword segmentation with that of a model trained on a contrastive segmentation. Working with an LLM such as GPT, this strategy is not feasible due to the immense expense to train such a model. Instead, we compare the outcome on a downstream task for words of different levels of segmentation quality, by selecting words with high and low values according to the criteria outlined previously. Assuming that (i) the segmentation quality has an effect on the particular task and that (ii) the proposed criteria are suitable to capture the segmentation quality, we should be able to see a performance difference between the two sets.

The linguistic task is that of predicting the lemma of an inflected verb form, which is applicable to every language in our data set; in a second experiment, we also generate inflected forms given the lemma and a morphological tag. We chose verbal morphology as it provides more variety than the inflection of nouns and adjectives.

3.1 Data Set

We use the morphological database in MorphyNet[1] (Batsuren et al., 2021), which contains inflectional and derivational morphology for 15 languages. For our experiments, we only consider languages with

1emma slozit form slozeny features V;PFVIV.PTCP;PASSIFEM;PL segmentation slozitnly

Table 1: Inflectional morphology in MorphyNet for the Czech word slozeny ('composed'). The segments correspond to the morphological features.

Latin script and selected 10 languages of different language families. To annotate the separation of inflection suffixes and stem, we use MorphyNet's inflectional information, where entries for an inflected form list the lemma, the morphological features and the canonical representation of the morphological segmentation (cf. table 1)

Some entries in the data set do not correspond to modern standard spelling (for example poynted as English verb); thus we applied a filtering step based on two conditions: first, the lemma of the word needs to occur in a dictionary and second, the inflected word form needs to occur at least once in a text corpus for the respective language. For this purpose, we obtained a Wikipedia dump for every language. The filtering is designed to be rather conservative such that the word forms are valid forms of contemporary language, which is important when assessing the impact of the segmentation quality, where we want the test set to be as clean as possible. Table 8 (in the appendix) shows the number of entries after the filtering.

Additionally, the Wikipedia data is used to get an idea about a word's frequency. While the frequencies in this text corpus do not correspond to those in the pre- training data, they still allow to approximately distinguish between high- frequency and low- frequency words.

4 Separation of Stem and Inflection

In this first experiment, we apply a linguisticallyoriented criterion and study whether and how inflection suffixes are separated from the stem. We start from the hypothesis that a clean separation of inflectional suffixes allows for a better representation with regard to generalization due to separating the lexical content in the stem from the morphosyntactic information in the inflectional parts.

text[[129, 855, 475, 902], [496, 76, 837, 108]] We define five categories, as illustrated in table 2, to describe the segmentation status of a word. Given the gold analysis, we compare how the word is segmented in the LM. The five categories are defined as follows:

EXACT: the word is split into exactly two parts, the stem and the inflection suffix SINGLE: the inflection suffix consists of one piece; the stem is further split CONCAT: the inflection suffix consists of several pieces; the stem is or is not further split OVERLAP: there is no clear separation between the stem and the inflectional suffix UNSPLIT: the word remained unsplit

The categories EXACT, SINGLE and CONCAT all met the condition of a split at the stem- inflection boundary, for the categories OVERLAP and UNSPLIT, the stem cannot be clearly separated from the stem. In practice, we find that the category UNSPLIT is comparatively infrequent, with a majority of the words falling into the groups EXACT, SINGLE, CONCAT and OVERLAP.

The segmentation analysis in MorphyNet is in canonical notation, thus the concatenation of the segmentation analysis does not result in the inflected form itself, but in a sequence of the lemma and the inflectional suffix(es), for example (FR) rembrunissons → rembrunirlissons ((we) darken). As inflection suffixes, we consider all parts of the segmentation except for the first one, which is the lemma3. As we ignore the lemma part of the gold segmentation, there are no problems with irregular verb forms or stem changes between lemma and inflected form. Many languages only have one suffix part, others like Finnish or Hungarian can have more. In the case of several suffixes, we only consider words where the concatenation of the suffixes in the gold segmentation also corresponds to the right side of the inflected word, but not forms like (ES) abramonos → abrilamoslnos (let's open up) where the suffixes are represented in the canonical form and thus can deviate from the surface form.

The GPT segmentation was obtained for the target word without surrounding sentence context4. Figure 1 shows the distribution of the segmentation categories for verbs. Overall, the category

[1] 1

Table 1: Segmentation categories derived from MorphyNet for French verbs (inflectional suffixes are highlighted).

lemma	form	morph. features	gold segm.	GPT-4o-segm.	category
commander	commandait	V IIND;PST;IPFV;3;SG	commanderlait	command ait	EXACT
canaliser	canalisent	V V IIND;PRS;3;PL	canaliserlent	can alis ent	SINGLE
commander	commanderait	V V IIND;3;PL	commanderlerait	command era ient	CONCAT
commander	commandaient	V IIND;PST;IPFV;3;PL	commanderlait	comm anda ient	OVERLAP
commander	commande	V IIND;PRS;1;SG	commanderle	commande	UNSPLIT

range	segm.	EN	DE	SE	FR	IT	ES	PT	FI	HU	CS
low:	OVERLAP	485	483	483	496	455*	491	493	452	360	483
f $\leq$ 10	NO OVERLAP	469	488	306/328	491	490	494	497	468	370	485
mid:	OVERLAP	493	493	487	489	470*	489	494	462	394	495
10<f	NO OVERLAP	495	494	454/481	492	491	493	496	471	398	481
high:	OVERLAP	499	496	470	489	485*	488	493	489	408	493
f > 500	NO OVERLAP	496	498	165/170	494	496	494	495	408/415	397	458/485

Table 2: Number of correctly predicted lemmas in a set of 500 randomly selected verb forms.
: the no-overlap system is significantly better than the overlap system (x-square test with a significance level of $\alpha = 0.05$ -

IMAGE NOT PROVIDED

Figure 1: Segmentation categories per language.

OVERLAP is dominant in most languages. This is particularly striking for English, which has the highest amount of training data by far, while also being a morphologically poor language. The English inflectional suffixes are generally rather short (e.g. - s for the plural of nouns or the third person for verbs), but many subword pieces tend to be longer (- izing, - ated, - lated, - ating, - ized, ...). While some of them are close to morphemes, the segmentation is not systematic in a linguistic sense. In particular the amount of the category CONTACT is to a certain extent language- dependent as only languages with generally longer inflectional suffixes can have the suffix split into several pieces. However, even though many verbs are not split at the boundary between stem and inflectional suffix, there is still often some form of systematicity.

4.1 Task: Verb Lemma Prediction

In this experiment, we investigate whether segmentation at the boundary between stem and inflectional suffixes has an effect on the task of predicting the lemma. As the frequency might be a relevant factor, we define 3 frequency ranges (cf. table 3) based on the observed frequency in the Wikipedia data. We compare verbs of the splitting category OVERLAP with verbs where inflection and stem are clearly separated (EXACT, SINGLE, CONTACT), with the hypothesis that verbs of the set OVERLAP should

perform worse than verbs of the set NOT OVERLAP, as a clear separation between stem and inflection conceptually allows for a better generalization, in particular for words of lower frequency.

We randomly select 500 verbs per group; as common irregular verbs are typically listed in abundance in grammatical resources and thus are likely leaked in the pre- training data, we excluded the ten most common irregular verbs (according to gpt- 4o) per language. Furthermore, we excluded verb forms that have the same surface form as the lemma, as the frequency of the word used as inflected form might differ considerably from the frequency of the form used as lemma.

We use the model gpt- 4o with a relatively low temperature of 0.1 for a more stable outcome; the prompt is formulated in English for all languages:

Answer with one word.

The lemma of the (French|...) verb "v" is

The prompt clearly states that we look for the verb lemma and also explicitly mentions the target language, which is important in case of an ambiguous part- of- speech and verbs that can occur in different languages, for example mentions which can also be an inflected form of the French verb mentir (to lie), in addition to the English form.

Table 3 shows the results grouped according to language families: there is no clear difference in the performance between the two sets, indicating that the separation of inflectional suffixes and stem is not a sufficient criterion for segmentation quality. Only for Italian, we can observe a better performance for the NO OVERLAP set.

A general factor might also be that the OVERLAP set represents the majority group in most languages, and thus, even in combination with frequency information, is not fine- grained enough to be discriminative of segmentation quality, while at the same time, the condition to segment at the inflection boundary is hard to meet, especially when considering that the segmentation has to work for many languages at once. This result does not necessarily say that linguistically sound segmentation in general is not better, but we can only conclude that the criterion of segmentation at the inflection boundary is not sufficient to measure segmentation quality.

5 Segmentation Consistency

The criterion in the previous section was based on linguistic well- formedness; here, we look at segmentation quality from the angle of consistency, which also aims at capturing generalization abilities, but is formulated more robustly. We pursue the question whether a consistent segmentation across the inflected forms of a lemma provides a better basis for the representation than an inconsistent segmentation. The underlying assumption is that an internally coherent representation of different surface realizations of the same word should result in an overall better representation of that word, and thus provide a better basis for generalization and the modeling of potentially unseen words. Table 4 shows some examples, ranging from a generally consistent representation of the stem part of the verb to a largely inconsistent segmentation6.

dram atis ieren
v inc ereras en
dram atis ieren
v intoras en
dram atis ieren
vin cirasen
dram atis ieren
vincerast
dram atis ieren
vin conoras est
dram atis ieren
vin cese erras te
dram atis ieren
vin cesso eror as
tento dramatize
to win
to speed

Ideally, a good segmentation should provide a consistent splitting of the stem part, with more necessary variation towards the end of the word. We use the Overlap Coefficient to measure the similarity between the sets of segments of two different verb forms, which is defined as the size of the intersection divided by the size of the smaller one of the two sets:

$$overlap(A, B) = \frac{|A \cap B|}{min(|A|, |B|)}$$

The scores range between 0 (no overlap) and 1 (perfect match). A particular characteristic of this metric is that if A is a subset of B , then the coefficient is 1: this has the effect of comparing rather the segments of the stem part while disregarding suffixes that add to the overall length of the word, assuming that the stem part does not change much, whereas we expect comparatively more variation in the suffixes. In contrast, the Jaccard index (ratio of intersection over union) might be less practical when the two compared forms are of different lengths, and we do not expect a subset to be similar.

Below are some examples for forms of the Italian verb sorprendere (to surprise), and the respective overlap scores between lemma and form:

lemma	form	overlap score
s or pr endere	s or pr endere bbe	1
s or pr endere	s or pr end iamo	0.75
s or pr endere	s or pre se	0.5

The splitting in the first line is linguistically questionable, but the lemma's segments are an exact subset of the inflected form's segments, which is good in terms of consistency (overlap=1). For the other two words, the segments only partially match between the forms, and thus have a lower score. To obtain the overlap coefficient of an inflection paradigm, we computed the average of the overlap of every possible pair of forms. Table 2 shows an overview for all languages: for most languages, average overlap scores of 0.5 - 0.7 are dominant.

IMAGE NOT PROVIDED

Figure 2: Distribution of overlap scores per inflection paradigm for verbs. The scores are rounded to the nearest decimal, resulting in ranges of 0.1

In the following, we look at two variants of the lemma prediction task: (i) the average overlap of a verb paradigm, and (ii) the segmentation similarity of only the verb form and the lemma while also distinguishing between inconsistencies at the beginning vs. elsewhere in the word.

5.1 Paradigm Segmentation Overlap

In this experiment, we contrast verb forms from paradigms with high vs. low overlap coefficients: The underlying assumption is that the internal representation of verb forms with a less overlapping segmentation is sub- optimal as the forms cannot be well linked, whereas verbs with a high overlap coefficient are expected to be better connected within the paradigm. A further factor is the similarity of the segmentation of the inflected form to that of the lemma, i.e. the expected answer: a segmentation similar to the lemma is likely beneficial, thus further adding to the high/low overlap scenario. Note that we do not always have the full inflection paradigm of a verb at our disposition due to limitations of the dataset and our various filtering steps in the pre- processing; as inflection paradigm we thus define all observed forms of a verb lemma (with a minimum number of 5 forms per observed paradigm). We apply the following criteria to select 200 verbs per group:7

Average paradigm overlap select verb paradigms with the highest/lowest average overlap coefficients per language

Overlap to lemma: from those paradigms, select one form each with the highest/lowest overlap to the lemma (select at random if there are several forms with equal overlap)

Frequency: additionally, we look at two frequency bands and consider forms with frequencies below 10 or above 500.

Based on this definition of high/low overlap, we select sets for the tasks of lemmatization and generation of inflected forms8.

5.1.1 Lemmatisation Task

The experimental settings are identical to that in section 4.1. Table 5 shows the result: There is a general tendency for the low- overlap sets to perform worse; this effect is most pronounced for Hungarian and low- frequency Finnish words.

With regard to errors, the proposed lemma is often orthographically close (for example, (DE) order/nordnen (to order/organize)). We also observed errors traceable at the semantic level, for example (DE) lost ((he) casts) → verlieren (to lose) instead of losen, presumably due to the (unsplit) form lost, i.e. the past participle of to lose.

5.1.2 Generation of Inflected Forms

The generation task consists in finding the correctly inflected form given the lemma and a tag specifying the morphological features, which is more challenging than the previous task of predicting the lemma of an inflected form. One difficulty is that of the prompt formulation and the terminology used for the respective grammatical features. To keep the prompting as simple as possible and equal across languages, the prompt is simply derived from the morphological tag provided by MorphyNet, where the abbreviations are replaced by their full terms, according to the UniMorph documentation (cf. table 9 in the appendix). We compare a zero- shot and a one- shot variant, where the example is randomly selected from a set of 50 additional items from the same category (high/low overlap) and frequency range. The (one- shot) prompt format9 is as follows:

Table 5: Number of correctly predicted lemmas (N=200) contrasting segmentation consistency. marks significant difference between high/low overlap sets (χ -square test with a significance level of $\alpha = 0.05$

	DE	SV	FR	IT	ES	PT	FI
freq $>$ 500							
highOverlap zero shot	197	190	196	193	200	200	186
lowOverlap zero shot	189	175*	184*	191	191*	188*	180
highOverlap one shot	191	194	194	189	200	200	186
lowOverlap one shot	185	185	187	195	191*	197	180
freq $\leq$ 10							
highOverlap zero shot	188	174	187	196	198	192	180
lowOverlap zero shot	166*	131*	161*	171*	169*	160*	130*
highOverlap one shot	189	175	184	195	199	193	185
lowOverlap one shot	172*	140*	172	172*	177*	176*	122*

Generate the inflected form given a German lemma and the morphological features. Answer with one word. lemma: "wagen", tag: verb, indicative, past, first person, plural form: "wagten" lemma: "biegen", tag: verb, indicative, present, third person, singular

Table 6 shows the results: for the zero- shot variant, the sets of less consistently split verbs perform worse, in particular for the low- frequency words. Overall, the one- shot variant does not improve much over the zero- shot variant, but reduces the difference between the two groups of consistently vs. inconsistently split verbs. These results indicate that the segmentation consistency is relevant, in particular for low- frequency words.

5.2 Positional Segmentation Differences

Here, we focus on consistent segmentation between the verb form and the lemma, and further assume that consistent segmentation at the beginning of the word, i.e. the lexical part of the word, is more important than at the end of the word, where the model is likely more robust due to observed variations with different inflections. We apply the following conditions to select verbs for three contrasting sets:

- Similarity verb forms with a similarity to the lemma below 0.35 or above 0.7- Position of difference verb forms with low similarity are grouped into subsets where the first subword token is the same for both words (same_1st) or different (diff_1st)10

Frequency forms with a frequency below "low- freq") or above 50 ("high- freq")

Table 7 shows the results: while we see the hypothesis that inconsistent segmentation at the beginning of a word has a negative effect confirmed, though not for all languages, we also have the somewhat surprising result that a matching first token, even with an otherwise low similarity, performs as well

as the high- similarity group. One possible interpretation is that the relevant semantic information is already mostly contained in this first token.

6 Conclusion and Future Work

We proposed two criteria to capture the quality of subword segmentation in LLMs and evaluated to what extent words which score high or low for these criteria affect the performance of the LLM on a linguistic task for ten diverse languages. Both criteria are targeted at the generalization abilities of the language model; the first one is more linguistically inspired and aims at a clear separation of stem and inflectional suffixes, whereas the second one rewards consistent segmentation within an inflection paradigm. The design of the criteria is in principal language- independent, but requires language- specific information: morphologically annotated data for the first one, and information about sets of inflection paradigms for the second one.

The results of our experiments indicate that the subword segmentation does influence the behaviour

7: Number of correctly predicted verb lemmas out of N = 100 contrasting positional segmentation differences.
marks significant difference between high/low similarity sets (x- square test with a significance level of $\alpha = 0.05$

of the model. In particular for the criterion of segmentation consistency, we could observe a better performance for the sets with higher segmentation overlap. In contrast, the morpheme- boundary criterion was found to be less suitable. With a view to linguistic resources, the consistency- based criterion departs from a more minimal point, as no morphological analysis is needed other than knowing the inflection paradigm of a word.

The underlying segmentation is not only relevant for the representation within one language, but might also improve the multilingual competence of a model. Conceptually, when adhering to morpheme boudaries, the resulting segmentation can separate between lexical parts and functional components, which might benefit multi- lingual and structure learning; for instance, through supporting the learning of lexical equivalents of words sharing the same (or close) orthographic forms of the stem with different inflections, such as symbol - icEN, - ischDE, - iqueFR, - icznePOL. Moreover, inflectional affixes contain context information such as tense or number, and an accessible and consistent representation can potentially contribute to the learning of syntactic structure across languages.

Finally, with view to the current efforts to include less- resourced languages into LMs, segmentation strategies that promote a consistent representation and maximize the generalization abilities are a relevant and interesting research field.

7 Limitations

In this section, we briefly discuss the limitations of the presented work.

text[[138, 796, 483, 926], [500, 219, 846, 268]] Linguistic Tasks An obvious limitation are the simple linguistic downstream tasks that we used to evaluate the performance of the model. In our experiments, we mainly focus on predicting the lemma of a given verb form, which is arguably not the most exciting task, but has the advantage of being applicable to all languages in our data set. We extend this task to the generation of inflected forms based on lemma and morphological tags, which is more challenging and thus might be more affected by the underlying segmentation quality.

A general issue with the generation task is that it is to a certain extent language- dependent due to the different morphological features per language, and consequently the optimal terminology to describe these features in the prompt. This makes this task likely more dependent on the prompt formulation than the lemmatization task.

In a certain way, both the prediction of lemmas and the generation of inflected forms are not necessarily natural tasks for the LLM. However, to better understand the impact of the underlying segmentation, we wanted a direct link between the investigated form and the linguistic task. This is more difficult to model in more complex tasks such as translation where more factors come into play.

Prompting We did not explore several prompt options, but used a simple and straightforward one. Similarly, we did not explore different prompt languages, but kept the English prompt for all investigated languages.

In general, we are primarily interested in the performance difference between the sets of differently segmented words, but less in obtaining the best possible performance. Thus, to keep the conditions as

simple as possible, we only looked at a zero shot scenarios in most experiments.

Non- Concatenative Morphology With regard to linguistic soundness in segmentation, a crucial factor that cannot be satisfactorily modeled by subword segmentation is non- concatenativity, such as irregular word forms (e.g. go - went), but also semi- regular variations such as an Umlaut in specific contexts, such as (DE) Apfel $s_g - \tilde{A}pfel_{pl}$ (apple s_g/pl). To fully capture these phenomena, one approach that has been proposed for both language modeling and machine translation is the representation of canonical forms in combination with morpho- syntactic information (e.g. Tamchyna et al. (2017), Antoun et al. (2020); Nzeyimana and Niy

ongabo Rubungo (2022), Jabbar (2024)), which however needs an additional step to generate inflected forms when generating, namely the generation of inflected forms based on the canonical representation in combination with the respective morphological features, which is not trivial.

In our study, we mostly ignored the problems of non- concatenative operations, in particular in the second part focusing on the segmentation consistency within a verb paradigm where phenomena such as stem changes between lemma and inflected form necessarily lead to lower segmentation similarity. Our main reason is that regular segmentation strategies operating on surface words cannot handle such phenomena, and thus a linguistically sound modeling is out of reach with this method.

Languages and their Representation in the Training Data Finally, the amount of training data per language is also likely to have an influence on the segmentation quality for the respective languages, as suggested in Armengol- Estapé et al. (2022). With English making up the majority of the trainig data for GPT models, we would assume a distribution of subword tokens that best represents English, but not necessarily other languages, in particular if a language's words differ considerably from English. While this is not a central point of our investigation of segmentation criteria in general, finding an optimal representation across languages is nonetheless a relevant factor that deserves attention in segmentation strategies for multilingual language models.

8 Acknowledgements

The work was supported by the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (grant agreement No. 101113091) and by the German Research Foundation (DFG; grant FR 2829/7- 1).

References

References

[1] Wissam Antoun, Fady Baly, and Hazem Hajj. AraBERT: Transformer- based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open- Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9- 15, Marseille, France, 2020. European Language Resource Association.

[2] Jordi Armengol- Estapé, Ona de Gibert Bonet, and Maite Melero. On the multilingual capabilities of very large- scale English language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3056- 3068, Marseille, France, 2022. European Language Resources Association.

[3] Tamali Banerjee and Pushpak Bhattacharyya. Meaningless yet meaningful: Morphology grounded subword- level NMT. In *Proceedings of the Second Workshop on Subword/Character Level Models*, pages 55- 60, New Orleans, 2018. Association for Computational Linguistics.

[4] Khuyagbaatar Batsuren, Gabor Bella, and Fausto Giunchiglia. MorphyNet: a large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39- 48, Online, 2021. Association for Computational Linguistics.

[5] Lisa Beinborn and Yuval Pinter. Analyzing cognitive plausibility of subword tokenization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4478- 4486, Singapore, 2023. Association for Computational Linguistics.

[6] Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 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, 2021. Association for Computational Linguistics.

[7] Jue Hou, Anisia Katinskaia, Anh- Duc Vu, and Roman Yangarber. Effects of sub- word segmentation on performance of transformer language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7413- 7425, Singapore, 2023. Association for Computational Linguistics.

[8] Haris Jabbar. Morphpiece : A linguistic tokenizer for large language models. Preprint, arXiv:2307.07262, 2024.

[9] Katharina Kann, Ryan Cotterell, and Hinrich Schütze. Neural multi- source morphological reinfection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 514- 524, Valencia, Spain, 2017. Association for Computational Linguistics.

[10] Stav Klein and Reut Tsarfaty. Getting the #life out of living: How adequate are word- pieces for modelling complex morphology? In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 204- 209, Online, 2020. Association for Computational Linguistics.

A Data

Table 8 lists the number of inflected verb forms per language in our data set.

Lang	Verbs
EN	23342
FR	57650
DE	21567
ES	15924
IT	49349
PT	27727
FI	22152
SV	14432
CS	20029
HU	37780

Table 3: Overview of the number of verbs (inflected forms) per language after the filtering step.

B Tags and Abbreviations

Table 9 lists the abbreviations used in MorphyNet's tags and the respective feature names used in the prompt formulation, based on the documentation in https://unimorph.github.io/doc/unimorph-schema.pdf

V	verb
V.PTPC	participle
IND	indicative
SBJV	subjunctive
IMP	imperative
COND	conditional
POT	potential
PST	past
PRS	present
FUT	future
SG	singular
PL	plural
1	first person
2	second person
3	third person
PFV	perfective
IPVF	imperfective
PROG	progressive
PRF	perfect
FORM	formal
INFM	informal

Table 4: Abbreviations and features used in the generation experiment.