Enhancing Lexical Complexity Prediction in Italian through Automatic Morphological Segmentation

Laura Occhipinti[1]

[1]University of Bologna, Italy

Abstract

Morphological analysis is essential for various Natural Language Processing (NLP) tasks, as it reveals the internal structure of words and deepens our understanding of their morphological and syntactic relationships. This study focuses on surface morphological segmentation for the Italian language, addressing the limited representation of detailed morphological information in existing corpora. Using an automatic segmentation tool, we extract quantitative morphological parameters to investigate their impact on the perception of word complexity by native Italian speakers. Through correlation analysis, we demonstrate that morphological features, such as the number of morphemes and lexical morpheme frequency, significantly influence how complex words are perceived. These insights contribute to improving automatic lexical complexity prediction models and offer a deeper understanding of the role of morphology in word comprehension.

Keywords

Morphological segmentation, Lexical complexity prediction, Italian language

1. Introduction

Morphological analysis is crucial for various NLP tasks, as it provides insights into the internal structures of words and helps us better understand the morphological and syntactic relationships between words [1].

The Italian language, with its rich morphology and extensive use of inflection and derivation, presents unique challenges and opportunities for morphological segmentation.

Automatic segmentation, a key component of morphology learning, involves dividing word forms into meaningful units such as roots, prefixes, and suffixes [2]. This task falls under the broader category of subword segmentation [3] but is distinct due to its linguistic motivation. Computational approaches typically identify subwords based on purely statistical considerations, which often results in subunits that do not correspond to recognizable linguistic units [4, 5, 6, 7]. Making this task more morphologically oriented could enable models to generalize better to new words or forms, as basic roots or morphemes are often shared among words, and it could also facilitate the interpretation of model results.

When discussing morphological segmentation, we can refer to two types: (1) Surface segmentation, which involves dividing words into morphs, the surface forms of morphemes; (2) Canonical segmentation, which involves dividing words into morphemes and reducing them to their standard forms [8].

For instance, consider the Italian word *mangiavano* (they were eating). The resulting surface segmentation would be *mangi-* + *-avano*, where *mangi-* is a morph derived from the root of the verb *mangiare*, and *-avano* is the suffix indicating the third person plural of the imperfect tense. In contrast, the canonical segmentation would yield *mangiare* + *-avano*, with *mangiare* as the canonical morpheme and *-avano* as the suffix[1].

In this study, we focus on surface morphological segmentation for the Italian language. Morphological features are often not adequately represented in available corpora for this language, or they refer exclusively to morphosyntactic information, such as the grammatical category of words and a macro-level descriptive analysis mainly related to inflection. Information about the internal structure of words, such as derivation or composition, is often lacking.

The primary objective of this work is to use an automatic segmenter to extract a series of quantitative morphological parameters. We believe that our approach does not require the detailed analysis provided by canonical segmentation, which could entail longer processing times.

[1]It's important to note that the segmentation process is not always straightforward, as it involves various linguistic criteria that may not be immediately clear. For example, one of the challenges lies in deciding whether to detach or retain the thematic vowel—a vowel that appears between the root and the inflectional suffix, especially in Romance languages. In the case of *mangiavano*, the thematic vowel *-a-* could either be considered part of the root or treated as a separate morph. Similarly, other segmentation criteria might involve distinctions between compound forms, derivational affixes, or fused morphemes that do not have clear boundaries. As a result, the segmentation criteria can vary based on linguistic theory, the specific task (e.g., computational vs. linguistic analysis), or even the intended application of the segmentation (e.g., for syntactic parsing or machine learning).

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

laura.occhipinti3@unibo.it (L. Occhipinti)

0009-0007-8799-4333 (L. Occhipinti)

In addition to examining classic parameters reported in the literature that influence complexity [9], such as word frequency, length, and number of syllables, we aim to explore how morphological features integrate with these factors to affect word complexity perception. Specifically, we seek to understand how the internal structure of words contributes to the cognitive load that speakers experience when processing more complex lexical items.

Our premise is that words with more morphemes are more complex because they contain more information to decode [10]. For example, consider the word *infelicità* (unhappiness). To decode it, one must know the word *felice* (happy), from which it is derived, as well as the prefix *in-*, which negates the quality expressed by the base term, and the suffix *-ità*, which transforms the adjective into an abstract noun. Therefore, to fully understand the meaning of *infelicità*, the reader or listener must be able to correctly recognize and interpret each of these morphemes and their contribution to the overall meaning of the word.

The main contributions of this work are: (1) Providing a tool capable of automatically segmenting words into linguistically motivated base forms; (2) presenting the dataset constructed for training our model; (3) evaluating the impact of different linguistic features on speakers' perception of word complexity, with a particular focus on morphological features.

2. Related Works

The study of morphological segmentation has evolved from classical linguistics to advanced machine learning techniques [11, 12]. The main approaches include **lexicon-based** and **boundary-detection-based** methods [2]. Lexicon-based methods rely on a comprehensive database of known morphemes [13, 14, 15], while boundary-detection methods identify transition points between morphemes using statistical or machine learning techniques [16, 17, 18].

Another significant distinction is between **generative** models and **discriminative** models. Generative models, suited for unsupervised learning, generate word forms and segmentations from raw data [19, 20, 21]. In contrast, discriminative models, which require annotated data, predict segmentations based on learned relationships from labeled examples [22, 23].

Unsupervised methods do not require labeled data, making them attractive for leveraging vast amounts of raw data. They trace back to Harris (1955), who used statistical methods to identify morphological segments. Notable systems include Linguistica [24, 25] and Morfessor [26, 27], which employ the Minimum Description Length (MDL) principle to identify regularities within data. Despite their utility, unsupervised methods often suffer from oversegmentation and incorrect segmentation of affixes [19, 28]. These challenges arise due to the complex interplay of phonological, morphological, and semantic factors in natural languages.

Semi-supervised methods leverage both annotated and unannotated data, enhancing model performance with minimal manual annotation [29]. These methods are effective in scenarios with limited labeled data[30, 31], using initial labeled datasets to hypothesize and validate patterns across larger unlabeled corpora [32]. While beneficial, semi-supervised methods depend on the quality of initial labeled datasets and may struggle with languages exhibiting extensive morphological diversity [2].

Supervised methods, relying on annotated datasets, typically achieve higher accuracy due to learning from explicitly labeled examples. Techniques include neural networks, Hidden Markov Models (HMM), and Convolutional Neural Networks (CNNs) [33, 34, 35, 23]. Despite their high performance, supervised methods are limited by the need for extensive annotated corpora, which can be costly and time-consuming to create.

Given access to a large annotated dataset for the Italian language, on which we made semi-manual corrections, our study primarily adopts a supervised approach.

2.1. Resources available for the Italian language

Several computational resources and tools have been developed to manage Italian morphological information [36, 37, 38, 39, 40, 41]. These resources are essential for improving the accuracy of text processing and supporting advanced linguistic research. However, many of them focus primarily on morphological analysis, without providing detailed support for morphological segmentation, which limits their usefulness in tasks that require fine-grained word structure analysis. Even those tools that offer segmentation often approach it with different methods and objectives than ours.

Morph-it! [37] is an open-source lexicon that contains 504,906 entries and 34,968 unique lemmas, each annotated with morphological characteristics that link inflected word forms to their lemmas. While valuable for lemmatization and morphological analysis, it is not suited for morphological segmentation, as it primarily focuses on inflected forms rather than decomposing words into their individual morphemes.

MorphoPro [39] is part of the TextPro suite and is designed for morphological analysis of both English and Italian. It uses a declarative knowledge base converted into a Finite State Automaton (FSA) for detailed morphological analysis. However, MorphoPro's output is geared towards global morphological analysis and lacks support for internal word segmentation into morphemes, limiting its applicability for more granular tasks.

MAGIC [36] provides a lexicon of approximately 100,000 lemmas and performs detailed morphological and morphosyntactic analysis. However, similar to other resources, MAGIC does not focus on morphological segmentation. Instead, it provides morphological and syntactic information about word forms, making it more useful for general morphological analysis rather than segmenting words into individual morphemes.

Getarun [38] offers a lexicon of around 80,000 roots and provides sophisticated morphosyntactic analysis. However, like MAGIC, it is designed primarily for syntactic parsing and lacks functionality for detailed morphological segmentation, focusing instead on morphological and syntactic relationships.

DerIvaTario [41] is another resource that provides significant support for morphological segmentation, particularly in the context of derivational morphology. It offers detailed information on derivational patterns in Italian, mapping out how words are formed through derivational processes, which is especially useful for studying word formation in a structured manner. However, DerIvaTario focuses primarily on canonical segmentations and does not always recognize smaller morphemes, such as final morphemes. This limitation means it may miss finer-grained morphological elements, making it more suitable for analyzing larger, derivational units rather than capturing all inflectional components.

AnIta is an advanced morphological analyzer for Italian, implemented within the FSA framework [40]. It supports a comprehensive lexicon with over 120,000 lemmas and handles inflectional, derivational, and compositional phenomena. AnIta's segmentation occurs on two levels: superficial segmentation of word forms and derivation graphs. Although derivation graphs are incomplete, the tool's focus on superficial segmentation aligns with our research needs. For the segmentation of lemmas related to derivational phenomena, AnIta adopts two main rules: (1) affixes are kept unchanged; (2) lexicon entries are segmented only if their base is a recognizable independent Italian word.

3. Methods

In this study, we trained three models, originally developed for other languages, using an Italian dataset that was manually created and verified with morphological segmentations. After evaluating the performance of the models, we selected the most effective one and used it to extract morphological parameters from the words in the MultiLS-IT dataset, a resource designed for lexical simplification in the Italian language [42, 43].

The dataset comprises 600 contextualized words, annotated for complexity and accompanied by substitutes perceived as simpler than the target word. Each word was evaluated by a group of native speakers with a perceived complexity score ranging from 1 to 5. In the dataset, the aggregated and normalized complexity value is between 0 and 1, where 0 indicates very simple words and 1 indicates very complex words[2]. The morphological traits extracted by the selected model were then integrated with other linguistic features typically considered influential in the perception of word complexity [9]. These combined features were analyzed in a correlation study with the perceived complexity values of MultiLs-IT to assess their impact on predicting linguistic complexity. By examining the relationships between these variables, we aim to determine whether morphological measures can be effectively used in systems designed to automatically identify word complexity.

3.1. Dataset

The primary reference for this work is the AnIta dataset, which includes data annotated with morphological segmentations based on specific rules. One rule excludes bases derived from Latin, Greek, and other languages. Since Italian, especially in technical and specialized fields, contains many such words, we modified the dataset to include these forms to ensure accurate representation.

The initial dataset consisted of numerous entries automatically generated by AnIta, often including over-generated word-forms (possible words [44]), especially in evaluative morphology. This resulted in a comprehensive dataset with approximately two million entries.To adapt the AnIta dataset for our research needs, we undertook several steps.

1) Due to the extensive size, we reduced the sample, retaining one-third of entries for each letter, resulting in approximately 728,814 word-forms (35% of the original dataset). This sample maintains a fair representation of all linguistic categories[3]. 2) We systematically identified and addressed prefixes and suffixes, prioritizing longer affixes to preserve more informative morphological structures. This semi-automatic approach facilitated manual verification while enhancing segmentation quality. 3) We manually reviewed the segmented words, ensuring accuracy and consistency, preserving prefixes in their original forms as per AnIta's rule number one. 4) The final dataset was divided into training (80%) and test (20%) sets, comprising 583,051 and 145,763 words respectively. This split allowed effective training and validation of our models without needing a separate validation set, as no parameter tuning was performed. This streamlined

[2]The resource is available at https://github.com/MLSP2024/MLSP_Data.

[3]Initially, we aimed to manually review the entire dataset to address any inconsistencies and overlooked segments. However, due to time constraints, we opted to reduce the dataset by randomly selecting 30% of the entries for each letter.

Automatic segmentation systems	Precision	Recall	F1	Accuracy
Neural Morpheme Segmentation	**0.9879**	**0.9806**	**0.9892**	**0.9793**
MorphemeBERT	0.9868	0.9199	0.9522	0.9581
Morfessor FlatCat	0.7974	0.3676	0.5033	0.7399

Table 1
Results of models on morphological segmentation.

methodology ensured a robust dataset for implementing and evaluating our automatic segmentation system.

3.2. Segmentation Models

Given the extensive dataset at our disposal, we selected models within the domain of supervised or semi-supervised learning. The models considered include: **Morfessor FlatCat** [31]: a semi-supervised model that utilizes a HMM approach for morphological segmentation. It is efficient in handling languages with complex morphological structures. The model's flat lexicon and the use of semi-supervised learning make it particularly suited for scenarios where annotated data is scarce.

Neural Morpheme Segmentation [33]: a supervised model based on CNNs, designed to segment morphemes by treating the task as a sequential labeling problem using the BMES scheme (Begin, Middle, End, Single). This model is noted for its ability to capture local dependencies within textual data. Its architecture includes multiple convolutional and pooling layers, enhancing its capability to identify and segment complex morphological patterns.

MorphemeBERT [45]: an advanced model that integrates BERT's characters embeddings with CNNs to enhance morphological segmentation. BERT provides deep, context-rich linguistic representations, which can significantly improve the model's accuracy in identifying morphemic boundaries.

3.3. Evaluation

After constructing the dataset and selecting the previously described models, we proceeded with the training. Table 1 presents a comparative evaluation of the three models using precision, recall, F1 score, and accuracy. These metrics are standard for assessing the performance of boundary detection models, providing a comprehensive overview of each model's effectiveness in identifying and segmenting morphemes accurately.

Neural Morpheme Segmentation demonstrates the highest performance among the three systems across almost all metrics, particularly excelling in precision and F1 score. The high precision (0.9879) indicates that the model is very accurate in identifying correct morpheme boundaries, minimizing false positives. In other words, when the model segments a word, it reliably places the boundaries at the correct points. Its F1 score (0.9892), which balances precision and recall, underscores the model's ability not only to accurately segment morphemes but also to capture the majority of them with minimal oversight. The high recall (0.9806) confirms that the model rarely misses morphemes, making it particularly well-suited for handling complex or less frequent morphological patterns. This balance between high precision and recall showcases the robustness of the CNN-based architecture, which can effectively model both local dependencies between segments and the global morphological structure of words[4].

MorphemeBERT demonstrates a high level of precision, indicating that when it identifies a morpheme, it is likely correct. However, its recall is noticeably lower than that of Neural Morpheme Segmentation, which suggests that while it makes fewer errors, it also fails to detect a significant number of morphemes. This trade-off between precision and recall points to a more conservative approach in morpheme segmentation, where the model prioritizes accuracy over coverage. The F1 score of 0.9522, though still strong, highlights this imbalance between precision and recall, meaning the model performs well but lacks the comprehensive identification that would elevate its overall performance. The accuracy of 0.9581 reflects that the model is quite reliable in general, but its inability to capture as many correct morphemes as Neural Morpheme Segmentation affects its overall segmentation capability. This limitation might be due to how MorphemeBERT integrates BERT embeddings, which are optimized for context-rich predictions but may struggle with identifying morphemic boundaries in less straightforward or ambiguous cases, leading to more missed segments.

Morfessor FlatCat shows a considerably weaker performance compared to the other two models. While its precision score of 0.79744 is decent, meaning that the morphemes it identifies are mostly accurate, its recall is notably low. This indicates that the model misses a substantial number of morphemes, failing to capture the full complexity of word segmentation. The low recall suggests that Morfessor FlatCat struggles to identify many valid morphemic boundaries, which results in incomplete or inaccurate segmentations. Consequently, its F1 score (0.5033) and accuracy (0.7399) are signifi-

[4]This model is available upon request. Please contact the author directly to access to the model and relevant references.

cantly lower, suggesting that this system is less reliable for applications requiring high fidelity in morpheme segmentation.

4. Selection of Linguistic Features

Based on a thorough review of the literature on lexical complexity prediction [9, 46], we selected several linguistic features to analyze their impact on complexity. In addition to common surface characteristics, such as the number of letters, syllables, and vowels in words, commonly used in complexity studies and readability calculations, we identified other relevant parameters. One key factor is the frequency of a word, as more frequent words tend to be perceived as more familiar and thus less complex. We calculated it using the ItWac corpus [47]. Another important parameter is the number of senses a word has, measured using the lexical resources ItalWordnet [48]. Lastly, the presence of stop words, calculated with Spacy model, which are common words that often carry little inherent meaning, can influence the perceived complexity of a sentence or text. Given the focus of this study on morphological features' impact on lexical complexity, we concentrated on several key aspects related to the internal structure of words. These features could show how morphological traits contribute to word intricacy:

Number of morphemes: Morphemes are the smallest units of meaning in words, including affixes (prefixes and suffixes) and roots. The number of morphemes gives an indication of the information load of a word. Lexical items with more morphemes typically require more decoding effort from readers. We used our Convolutional Neural Model for automatic morphological segmentation and morpheme counting.

Morphological density: This quantitative metric is defined as the ratio of the number of morphemes to word length, offering a measure of how densely packed meaningful units are within a word. Higher morphological density can indicate more cognitive load, as each unit contributes distinct information, potentially raising the complexity of the word.

Frequency of the lexical morpheme: Lexical morphemes carry the core meaning of the word. Employing our morphological segmentator on the ItWac corpus [47], enabled us to dissect the word into segments and aggregate the frequencies of individual morphemes. This frequency, transformed using a logarithmic scale, helps predict complexity by leveraging the familiarity of frequently occurring morphemes. The use of lexical morpheme frequency as a complexity indicator is based on the idea that even if a word is unfamiliar as a whole, its component morphemes may be common in the language and more recognizable [49].

By integrating these morphological features with other linguistic traits typically considered influential in speakers' perception of complexity, we aim to assess their impact on predicting linguistic complexity[5].

5. Analysis and discussion

Through studying the correlations between these variables, we seek to determine whether morphological measures can be effectively used to develop systems capable of automatically identifying word complexity. To achieve this, we conducted a correlation and significance analysis between the features discussed earlier and the perceived complexity values for the 600 words included in MultiLs-IT.

Feature	Correlation	p-value
Length	0.082	0.045*
Number of vowels	0.097	0.018*
Number of syllables	0.091	0.026*
Number of Morphemes	0.112	0.006*
Senses_ID	-0.277	0.000*
Stopword	-0.124	0.003*
Lemma Frequency	-0.467	0.000*
Morphological Density	0.036	0.381
Lexical morpheme frequency	-0.333	0.000*

Table 2
Spearman correlation coefficients and p-values for features and complexity. Note: * indicates statistical significance.

Table 2 presents the Spearman correlation coefficients and their statistical significance for the features calculated[6]. The correlation analysis reveals several important insights.

Word length, number of vowels, and number of syllables all have small but statistically significant positive correlations with complexity. This suggests that, as expected, longer words with more vowels and syllables tend to be perceived as more complex. These factors are typical in readability studies, where more phonologically complex words are generally harder to process.

The number of morphemes also shows a positive correlation with complexity, reinforcing the idea that words with more morphemes are perceived as more complex. This feature is statistically significant as well.

Negative correlations for senses_ID, stopword presence, and lemma frequency suggest that words with more senses, those that are stopwords, or those that are more

[5]For a detailed analysis of how these parameters were processed, refer to Occhipinti 2024.

[6]Spearman's rank correlation was chosen because it does not assume a linear relationship between variables, making it more suitable for our dataset, where the relationships between features like word length, number of morphemes, and word complexity may not follow a strictly linear pattern. Spearman's correlation measures whether an increase in one variable tends to be consistently associated with an increase (or decrease) in another, which is more appropriate given the nature of our linguistic features.

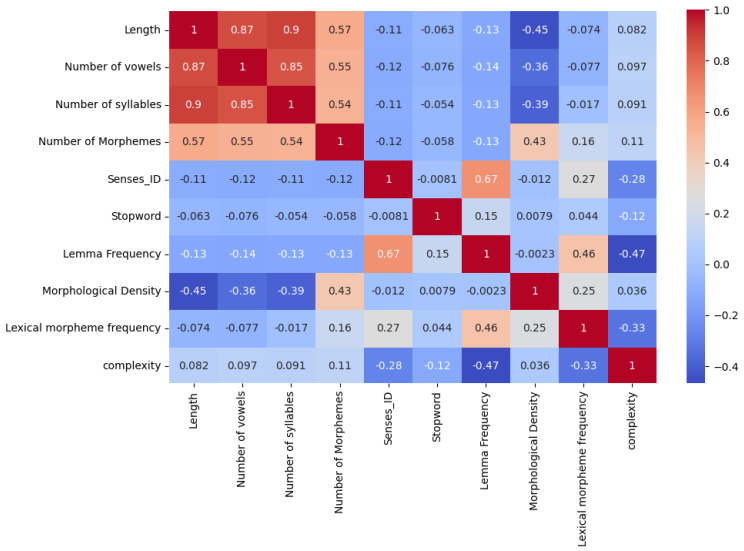

Figure 1: Correlation of complexity values.

frequently used are perceived as less complex. These features are also statistically significant. It is noteworthy that the number of senses (senses_ID) is inversely proportional to complexity. This could be attributed to the incompleteness of ItalWordNet, potentially leading to unreliable predicted values.

Morphological density, however, does not show a statistically significant correlation with complexity, suggesting that the ratio of morphemes to word length may not be a strong predictor of perceived complexity.

The lexical morpheme frequency shows a significant negative correlation with complexity, indicating that more frequently occurring morphemes contribute to lower perceived complexity. This supports the notion that familiar morphemes, even within otherwise complex words, aid in comprehension.

These findings underscore the importance of considering a range of linguistic features, including morphological traits, when assessing lexical complexity. By integrating these features into computational models, we can enhance their ability to accurately predict word complexity and, subsequently, improve lexical simplification.

6. Conclusion

This study highlights the significance of integrating morphological features into automatic models to enhance the comprehension and prediction of lexical complexity. The high performance of the Neural Morpheme Segmentation model demonstrates the efficacy of convolutional neural networks in capturing the detailed patterns of morphological segmentation in the Italian language.

The correlation analysis reveals that while traditional metrics like word length and frequency are valuable predictors of complexity, incorporating morphological features provides additional insights that enrich our understanding of lexical complexity. Notably, the positive correlation between the number of morphemes and perceived complexity suggests that words with more morphemes are inherently more complex. Conversely, frequent lexical morphemes tend to reduce perceived complexity, highlighting the importance of familiarity in complexity perception. Our study also emphasizes the need for diverse linguistic features, including both surface characteristics and morphological traits, to create more robust and accurate models for predicting word complexity. The statistically significant correlations for most features validate their relevance in complexity prediction. However, it is important to note that our findings are based on a relatively small dataset of annotated complexity perceptions. To obtain more robust and generalizable results, it would be highly beneficial to have access to a larger and more diverse dataset of complexity annotations. Expanding the dataset to include a wider variety of texts and contexts would enhance the reliability of the correlations observed and improve the training and evaluation of automatic complexity prediction models.

Future research should focus on gathering more extensive annotated datasets and exploring additional linguistic features that may influence complexity perception. By doing so, we can further refine our models and develop more effective tools for lexical simplification and other applications aimed at improving text accessibility.

References

[1] J. T. Devlin, H. L. Jamison, P. M. Matthews, L. M. Gonnerman, Morphology and the internal structure of words, Proceedings of the National Academy of Sciences 101 (2004) 14984–14988.

[2] T. Ruokolainen, O. Kohonen, K. Sirts, S.-A. Grönroos, M. Kurimo, S. Virpioja, A comparative study of minimally supervised morphological segmentation, Computational Linguistics 42 (2016) 91–120.

[3] S. J. Mielke, Z. Alyafeai, E. Salesky, C. Raffel, M. Dey, M. Gallé, A. Raja, C. Si, W. Y. Lee, B. Sagot, et al., Between words and characters: A brief history of open-vocabulary modeling and tokenization in nlp, arXiv preprint arXiv:2112.10508 (2021).

[4] R. Sennrich, B. Haddow, A. Birch, 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), 2016, pp. 1715–1725. doi:10.18653/v1/P16-1162.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[6] K. Bostrom, G. Durrett, Byte pair encoding is suboptimal for language model pretraining, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 4617–4624.

[7] X. Song, A. Salcianu, Y. Song, D. Dopson, D. Zhou, Fast wordpiece tokenization, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 2089–2103.

[8] R. Cotterell, C. Kirov, J. Sylak-Glassman, D. Yarowsky, J. Eisner, M. Hulden, The sigmorphon 2016 shared task—morphological reinflection, in: Proceedings of the 14th SIGMORPHON workshop on computational research in phonetics, phonology, and morphology, 2016, pp. 10–22.

[9] K. Collins-Thompson, Computational assessment of text readability: A survey of current and future research, ITL-International Journal of Applied Linguistics 165 (2014) 97–135.

[10] W. U. Dressler, Ricchezza e complessità morfologica, Ricchezza e complessità morfologica (1999) 1000–1011.

[11] S. Scalise, Morfologia, il Mulino, 1994.

[12] J. A. Goldsmith, Segmentation and morphology, in: The handbook of computational linguistics and natural language processing, Wiley Online Library, 2010, pp. 364–393.

[13] J. G. Wolff, The discovery of segments in natural language, British Journal of Psychology 68 (1977) 97–106.

[14] C. G. Nevill-Manning, I. H. Witten, Identifying hierarchical structure in sequences: A linear-time algorithm, Journal of Artificial Intelligence Research 7 (1997) 67–82.

[15] M. Johnson, Unsupervised word segmentation for sesotho using adaptor grammars, in: Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology, 2008, pp. 20–27.

[16] Z. S. Harris, From phoneme to morpheme, Language 31 (1955) 190–222. URL: http://www.jstor.org/stable/411036.

[17] P. Cohen, B. Heeringa, N. M. Adams, An unsupervised algorithm for segmenting categorical timeseries into episodes, in: Proceedings of Pattern Detection and Discovery: ESF Exploratory Workshop London, 2002, pp. 49–62.

[18] A. Sorokin, A. Kravtsova, Deep convolutional networks for supervised morpheme segmentation of russian language, in: Proceedings of 7th International Conference in Artificial Intelligence and Natural Language (AINL 2018), 2018, pp. 3–10.

[19] M. Creutz, K. Lagus, Unsupervised models for morpheme segmentation and morphology learning, ACM Transactions on Speech and Language Processing (TSLP) 4 (2007) 1–34.

[20] H. Poon, C. Cherry, K. Toutanova, Unsupervised morphological segmentation with log-linear models, in: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2009, pp. 209–217.

[21] K. Sirts, S. Goldwater, Minimally-supervised morphological segmentation using adaptor grammars, Transactions of the Association for Computational Linguistics 1 (2013) 255–266.

[22] Z. S. Harris, Morpheme Boundaries within Words: Report on a Computer Test, Springer Netherlands, 1970, pp. 68–77.

[23] T. Ruokolainen, O. Kohonen, S. Virpioja, M. Kurimo, Supervised morphological segmentation in a low-resource learning setting using conditional random fields, in: Proceedings of the Seventeenth Conference on Computational Natural Language Learning, 2013, pp. 29–37.

[24] J. Goldsmith, Unsupervised learning of the morphology of a natural language, Computational linguistics 27 (2001) 153–198.

[25] J. Goldsmith, An algorithm for the unsupervised learning of morphology, Natural language engineering 12 (2006) 353–371.

[26] M. Creutz, K. Lagus, Unsupervised discovery of morphemes, in: Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning, 2002, pp. 21–30.

[27] M. J. P. Creutz, K. H. Lagus, Morfessor in the morpho challenge, in: Proceedings of the PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes, 2006, pp. 12–17.

[28] Ö. Kılıç, C. Bozsahin, Semi-supervised morpheme segmentation without morphological analysis, in: Proceedings of the workshop on language resources and technologies for Turkic languages, LREC, 2012, pp. 52–56.

[29] T. Ruokolainen, O. Kohonen, S. Virpioja, M. Kurimo, Painless semi-supervised morphological segmentation using conditional random fields, in: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers, 2014, pp. 84–89.

[30] J. Lafferty, A. McCallum, F. Pereira, et al., Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: International Conference on Machine Learning, 2001, pp. 282—-289.

[31] S.-A. Grönroos, S. Virpioja, P. Smit, M. Kurimo, Morfessor flatcat: An hmm-based method for unsupervised and semi-supervised learning of morphology, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics, 2014, pp. 1177–1185.

[32] X. Zhu, A. B. Goldberg, Introduction to semi-supervised learning, Springer Nature, 2022.

[33] A. Sorokin, Convolutional neural networks for low-resource morpheme segmentation: baseline or state-of-the-art?, in: Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology, 2019, pp. 154–159. URL: https://aclanthology.org/W19-4218. doi:10.18653/v1/W19-4218.

[34] L. Wang, Z. Cao, Y. Xia, G. De Melo, Morphological segmentation with window lstm neural networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2016, pp. 2842–2848.

[35] R. Cotterell, T. Mueller, A. Fraser, H. Schütze, Labeled morphological segmentation with semi-markov models, in: Proceedings of the Nineteenth Conference on Computational Natural Language Learning, 2015, pp. 164–174.

[36] M. Battista, V. Pirrelli, Una piattaforma di morfologia computazionale per l'analisi e la generazione delle parole italiane, Technical Report, ILC-CNR, 1999.

[37] E. Zanchetta, M. Baroni, Morph-it! a free corpus-based morphological resource for the italian language, in: Proceedings of corpus linguistics conference series 2005 (ISSN 1747-9398), volume 1, 2005, pp. 1–12.

[38] R. Delmonte, et al., Computational Linguistic Text Processing–Lexicon, Grammar, Parsing and Anaphora Resolution, Nova Science Publishers, 2008.

[39] E. Pianta, C. Girardi, R. Zanoli, The textpro tool suite., in: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), 2008, p. 2603–2607.

[40] F. Tamburini, M. Melandri, Anita: a powerful morphological analyser for italian., in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2012, pp. 941–947.

[41] L. Talamo, C. Celata, P. M. Bertinetto, Derivatario: An annotated lexicon of italian derivatives, Word Structure 9 (2016) 72–102.

[42] M. Shardlow, F. Alva-Manchego, R. Batista-Navarro, S. Bott, S. Calderon Ramirez, R. Cardon, T. François, A. Hayakawa, A. Horbach, A. Hülsing, Y. Ide, J. M. Imperial, A. Nohejl, K. North, L. Occhipinti, N. Peréz Rojas, N. Raihan, T. Ranasinghe, M. Solis Salazar, M. Zampieri, H. Saggion, An extensible massively multilingual lexical simplification pipeline dataset using the MultiLS framework, in: R. Wilkens, R. Cardon, A. Todirascu, N. Gala (Eds.), Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 38–46. URL: https://aclanthology.org/2024.readi-1.4.

[43] M. Shardlow, F. Alva-Manchego, R. Batista-Navarro, S. Bott, S. Calderon Ramirez, R. Cardon, T. François, A. Hayakawa, A. Horbach, A. Hülsing, Y. Ide, J. M. Imperial, A. Nohejl, K. North, L. Occhipinti, N. P. Rojas, N. Raihan, T. Ranasinghe, M. S. Salazar, S. Štajner, M. Zampieri, H. Saggion, The BEA 2024 shared task on the multilingual lexical simplification pipeline, in: E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, Z. Yuan (Eds.), Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 571–589. URL: https://aclanthology.org/2024.bea-1.51.

[44] M. Aronoff, A decade of morphology and word formation, Annual review of anthropology (1983) 355–375.

[45] A. Sorokin, Improving morpheme segmentation using bert embeddings, in: International Conference on Analysis of Images, Social Networks and Texts, Springer, 2021, pp. 148–161.

[46] K. North, M. Zampieri, M. Shardlow, Lexical com-

plexity prediction: An overview, ACM Computing Surveys 55 (2023) 1–42.
[47] M. Baroni, S. Bernardini, A. Ferraresi, E. Zanchetta, The wacky wide web: a collection of very large linguistically processed web-crawled corpora, Language resources and evaluation 43 (2009) 209–226.
[48] A. Roventini, A. Alonge, N. Calzolari, B. Magnini, F. Bertagna, Italwordnet: a large semantic database for italian., in: In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000), 2000, pp. 783–790.
[49] P. Colé, J. Segui, M. Taft, Words and morphemes as units for lexical access, Journal of Memory and Language 37 (1997) 312–330.
[50] L. Occhipinti, Complex word identification for italian language: a dictionary-based approach, in: Proceedings of Clib24, Sixth International Conference on Computational Linguistics in Bulgaria, 2024, pp. 119–129.