Sensitivity of Syllable-Based ASR Predictions to Token Frequency and Lexical Stress

Alessandro Vietti[1], Domenico De Cristofaro[1] and Sara Picciau[1]

[1]Free University of Bozen-Bolzano, Libera Università di Bolzano

Abstract
Automatic Speech Recognition systems (ASR) based on neural networks achieve great results, but it remains unclear which are the linguistic features and representations that the models leverage to perform the recognition. In our study, we used phonological syllables as tokens to fine-tune an end-to-end ASR model due to their relevance as linguistic units. Furthermore, this strategy allowed us to keep track of different types of linguistic features characterizing the tokens. The analysis of the transcriptions generated by the model reveals that factors such as token frequency and lexical stress have a variable impact on the prediction strategies adopted by the ASR system.

Keywords
Automatic Speech Recognition, Syllable, Phonology.

1. Introduction

The syllable is crucial in the process of spoken word recognition. It serves as an integral component within the prosodic system because it encompasses both traditional segmental and suprasegmental levels, facilitating the extraction of lexical and syntactic structures from acoustic information [1, 2]. Specifically, the syllable serves as the linguistic unit where crucial information for speech segmentation, rhythmic patterns, and lexical access is encoded [3]. In the field of Automatic Speech Recognition (ASR), graphemic segment has traditionally been the primary unit of processing. However, recent studies endorse the use of syllables or phonetic units of similar duration as an alternative strategy [4, 5, 6]. In latest ASR research employing Transformer-based neural models, the role of syllables is investigated both as tokens for word recognition and as components influencing internal speech representations within neural networks [7, 8, 9]. In our study, a neural ASR model was trained to process and recognize phonological syllables, integrating them into word structures. Our goal is to conduct a linguistic analysis on the output of syllabic processing by the speech recognition system. Through fine-tuning a large acoustic model, the study mapped speech signals onto phonological transcriptions segmented into syllables and words. The primary objective of our linguistic analysis is to test the effect of syllable token frequency and lexical stress on the accuracy of output neural representation. To understand how the ASR processes syllables and words differently, we developed a fine-grained linguistic annotation system. This approach was essential to move beyond the limitations of purely numerical metrics like Word-Error-Rate or, in our context, Token-Error-Rate. By employing this system, we could accurately categorize prediction types and link them with specific linguistic aspects of speech. We utilized Multiple Correspondence Analysis and Multinomial Logistic Regression to explore and uncover patterns that relate the neural network's output behavior to the linguistic factors.

2. Methodology

2.1. Data preparation and experimental setup

The preparation of the experiment started with the collection of the data to fine-tune the pre-trained Microsoft model WavLM-large [10]. Our dataset consists of approximately 30 hours of Italian data from the crowd-sourced corpus Common Voice [11], using 6,500 samples (5,000 for training, 500 for testing, and 1,000 for validation). The total Italian subset in Common Voice 13.0 comprises 6,881 speakers and spans approximately 343 hours of recorded speech. Since we are interested in observing the role that some phonological aspects might play in the recognition process, we used WebMAUS [12] to obtain X-SAMPA transcriptions of the corpus. In addition, we forced the model to recognize phonological syllables as tokens, instead of automatically generated subwords based on probability, frequency and likelihood [13]. We designed a custom tokenizer that relies on the Maximal Onset Principle [14] and the Sonority Sequencing Principle [15] and considers exceptionally /s/+stop clusters and geminates as part of the syllable onset [16, 17]. In

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
†These authors contributed equally.
Alessandro.Vietti@unibz.it (A. Vietti); dodecristofaro@unibz.it (D. D. Cristofaro); sapicciau@unibz.it (S. Picciau)
0000-0002-4166-540X (A. Vietti)

order to observe the placement of the recognized tokens and word boundaries in detail, we set the output format of the model so that tokens are separated by blank spaces and words are separated by pipes, as it can be seen in example (1)

(1) il | vwO to | a sso lu to |

2.2. Creation of the database

Once we tested the model and obtained the predictions, we extracted a sample of 300 pairs of reference and predicted sentences (*Rs* and *Ps*, respectively). The detailed observation of the pairs allowed us to define a set of prediction types. Word-level prediction types are those that affect canonical word boundaries and consist of three categories: merged words, meaning two reference words recognized as one; divided words, consisting of a single reference word recognized in two or more words; and token movement, namely the change of a reference token position within adjacent word boundaries. At a token level, prediction types represent deviances in terms of token insertion, substitution and deletion, as well as correctly recognized tokens. We then designed a set of labels (prediction tags *PT* - see Appendix A.1) representing the prediction types to annotate the tokens of our dataset. The labels consist of a sequence of affixes indicating the detected recognition events. Word-level affixes are *mer, div, mv* and, in case of token movement, *forw* or *back* to mark the direction of the shift; token level affixes are *ins, sub, del, eq*. Lastly, the suffix *syl* or *word* indicates if the phenomenon regards an individual token or the whole word. An example of our annotation can be seen below.

(2)

<table>
<tr><td>S</td><td colspan="3">essere</td><td colspan="4">umano</td></tr>
<tr><td>R</td><td>E</td><td>sse</td><td>re</td><td colspan="2">u</td><td>ma</td><td>no</td></tr>
<tr><td>P</td><td>E</td><td colspan="2">stre</td><td>ro</td><td>u</td><td>na</td><td>no</td></tr>
<tr><td>PT</td><td>eq_syl</td><td colspan="2">sub_syl</td><td>mv_forw_sub_syl</td><td>eq_syl</td><td>sub_syl</td><td>eq_syl</td></tr>
</table>

Given our dataset size of approximately 5900 tokens, a manual annotation of each entry would have been extremely time-consuming. Therefore, we designed an algorithm to operate a comparison of reference and predicted tokens (*Rt* and *Pt*, respectively) with the aim to obtain a semi-automated *PT* labeling. The algorithm works as follows: first, it attempts to identify the correspondences between reference and predicted words (*Rw, Pw*) despite potential mismatches given by prediction types affecting word boundaries. Each pair of sentences is split into words, and a function to calculate similarity based on Levenshtein distance is used to confirm or dismiss word matches. If the similarity score is lower than the established threshold, it indicates a mismatch. When this occurs, similarity is calculated between *Rw* and adjacent *Pw*s and viceversa. If a (partial) match is found, the word-level *PT* is appended to the corresponding tokens; otherwise, unmatched words are labelled as inserted (when not found in *Rs*) or deleted (when not found in *Pt*). Once word-level matches are identified, the algorithm proceeds with the comparison of each *Rt* and *Pt* within *Rw* and *Pw* respectively, and it then assigns the corresponding *PT* at a token level. The mechanism to find token matches within words and assign token-level *PT* is analogous to the one described above. The implementation of this algorithm allowed us to automatically annotate most part of the dataset. However, many entries required manual intervention, as in the cases of assimilation or predictions characterized by a very low quality, which resulted in significant mismatches. Lastly, we added to our dataset some phonological information about each token in order to conduct our linguistic analysis. We included relative frequency of *Rt* in the whole dataset used for the training and lexical stress, as well as presence of the token in the training vocabulary, POS of *Rw*, and *Rs* speech rate. However, only the first two variables were taken into consideration for the statistic analysis in this work.

3. Results

3.1. Explorative analysis

To analyze our prediction database, we first looked at the distribution of prediction types. Next, we used Multiple Correspondence Analysis (MCA) to explore the relationships between prediction types, token frequency, presence in the training vocabulary, and lexical stress. The syllable-based fine-tuned ASR model showed a high degree of accuracy in prediction, with only 28% of tokens having notable recognition errors, making *eq_syl* the most frequent category.

The following figures show the detailed distribution of marked prediction types. Our structured labeling system allows us to separately examine token-level phenomena and those affecting sentence structure due to word boundary errors. Figure 1 highlights that substitution is the most common token-level operation, followed by deletion and insertion. This means that most incorrectly recognized tokens still appear in the model's hypothesized transcription. However, token deletions and insertions (including entire words like prepositions, determiners, or auxiliary verbs) lead to more significant recognition discrepancies. It should be noted that the use of automatically generated phonological transcriptions as references increases the number of substitutions due to speech variability in the corpus.

Figure 2 shows the distribution of operation/equality tags affecting canonical word boundaries. Merging is the

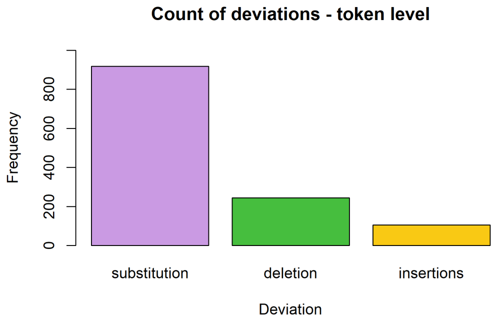

Figure 1: Count of deviations at a token level

most frequent process, involving 401 tokens, followed by divided words with 206 occurrences, and movement of single tokens with 48 instances. The movement label applies to single tokens, unlike other categories. Tokens in merged and divided words were mostly recognized correctly, with substitution being the second most common operation. Token deletion occurs more often in merged words, while token insertion is higher in divided words. For moved tokens, the distribution of equal and substituted tokens is nearly identical. Deletions and insertions do not apply to moved tokens since they can't be missing or added in the prediction.

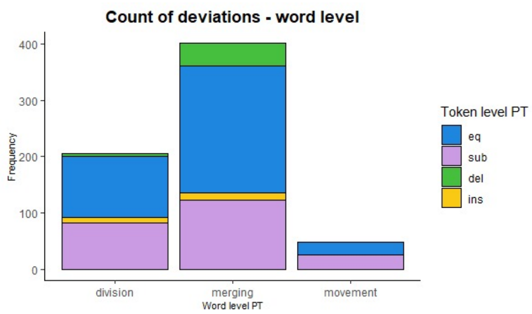

Figure 2: Count of deviations at a word level

Figure 3 shows the Multiple Correspondence Analysis (MCA) results using the *FactoMinerR* R package. This analysis reveals patterns between prediction types (event_syllable), token frequency (freq_tok_R_cat), presence in the training vocabulary (in_vocab_R), and lexical stress (stress_R). The relative frequency of tokens in the dataset was discretized into three levels using quantiles to obtain a uniform distribution of tokens across the three categories: from zero to one-third of tokens is "low frequency" (0-0.5%), from one-third to two-thirds is "mid frequency" (0.5-2.23%), and from two-thirds to one is "high frequency" (2.23-6.87%). Part of speech (POS) and syllable type (tok_type_R) were added later as supplementary variables to guide linguistic interpretation of the analysis. Insertion, being the least frequent operation, and complex syllable types (like CCVCC) were excluded due to their low frequency.

MCA is a dimensionality reduction technique for categorical variables, so the significance of the dimensions is derived from the distribution of the levels of the variables projected onto the plane. Interestingly, the top section shows that unstressed high-frequency tokens (over 2.23%), mainly subordinating conjunctions and determiners, are associated with deletion. The bottom-left section includes mid-frequency items (0.5% - 2.23%) with simple syllabic structures (CV) that are typically recognized correctly. Tokens with low frequency or which are absent from the training vocabulary are on the right side of the MCA chart. These less frequent, complex syllable tokens, often occurring in proper nouns and numerals, are typically handled with substitution.

3.2. Multinomial analysis

To statistically validate the findings from the MCA (figure 3), we conducted a multinomial logistic regression analysis using the *nnet* R library. The model examines the interaction between token frequency and lexical stress and, in this analysis, expresses the regression coefficients in odds (instead of logits) (see Appendix A.2). By looking at the plots of the model predictions and jointly evaluating the pairwise comparisons from the two tables (see Appendix A.4 and A.3), we can get a clearer interpretation of the results of the regression analysis. In Figure 4, we notice that when the prediction is equal to the reference, token frequency has a significant effect in the case of stressed syllables, whereas it appears to be less statistically relevant for unstressed syllables. Additionally, the difference in the presence or absence of lexical accent becomes significant as the frequency increases from low to mid to high. Regarding substitution, the patterns seem complementary to those observed in the matching of reference and prediction (i.e., in the *equal* plot). When syllables have a low frequency in the dataset, the probability that they are replaced with other syllabic tokens significantly increases. Although we have not explored which syllabic tokens or types they are replaced with and based on what criteria, it is safe to assume that it may be due to phonetic similarity. Specifically, there is a significant difference only between low frequency and the combined mid and high frequencies for both stressed and unstressed syllables. As for deletion, the regression coefficients reveal that the probability of deletion of unstressed syllables increases with frequency, but

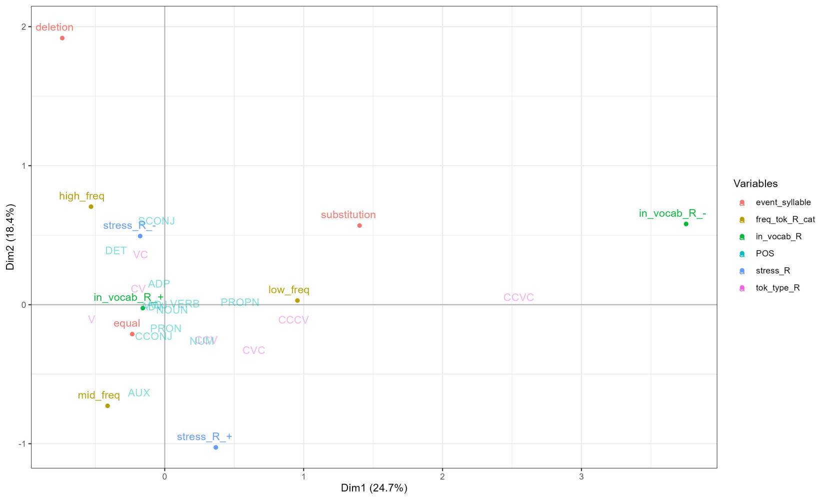

Figure 3: Multiple Correspondence Analysis (MCA) (A.5)

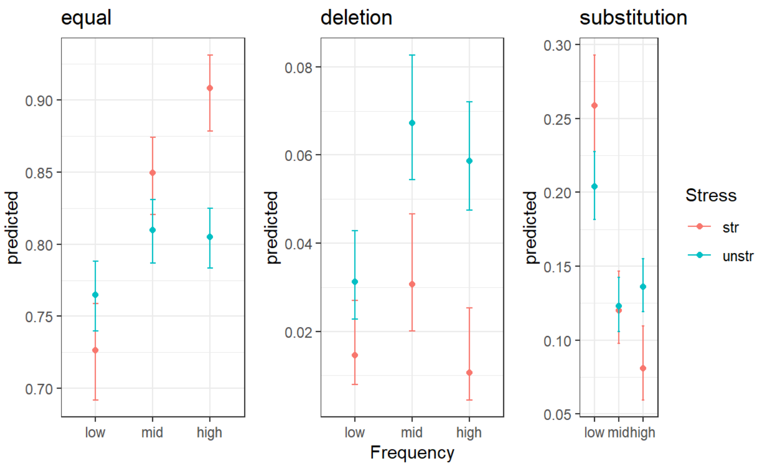

Figure 4: Interaction between token frequency and stress

only in the transition from low to medium frequency, with no further increase from medium to high frequency. For stressed syllables, the neutralization of a frequency effect is confirmed from the analysis of the coefficient. A quick exploration of the most deleted mid-frequency syllables shows that the preposition 'a' or V syllables in word-initial position are more likely deleted.

4. Conclusions and future work

This study provides insights into the role of syllables in ASR performance, particularly when integrating phonological information into the recognition process. By fine-tuning a neural ASR model to process and recognize phonological syllables, we were able to conduct a detailed linguistic analysis of its output. Our findings indicate that syllable frequency and lexical stress significantly impact ASR accuracy. Specifically, stressed syllables are more accurately recognized than unstressed ones, especially as frequency increases. Contrary to our expectation, among the low-frequency syllables, stressed tokens are more prone to substitution, whereas mid-frequency unstressed ones are more susceptible to deletion. This demonstrates the neural model's sensitivity to both distributional information in the dataset and phonological information and highlights the model's ability to detect varying syllabic prominence at the lexical level within the signal. As future work, we plan to include other linguistic factors as independent variables to refine our analysis. An interesting approach is to evaluate the impact of unstressed syllables and specific parts of speech by conducting an analysis exclusively on content words. Furthermore, we aim to investigate in detail syllable substitution in relation to token frequency and phonetic similarity to compare the weight of each factor whenever this strategy is adopted to deal with low-frequency tokens. In conclusion, our study showed the influence of token frequency and prominence in ASR predictions while demonstrating that complex computational tools, like modern neural networks, can be effectively utilized by linguists to simulate and test linguistically relevant hypotheses.

References

[1] M. E. Beckman, The parsing of prosody, Language and Cognitive Processes 11 (1996) 17–68. URL: https://doi.org/10.1080/016909696387213. doi:10.1080/016909696387213.

[2] S. Hawkins, R. Smith, Polysp: A polysystemic, phonetically-rich approach to speech understanding, Italian Journal of Linguistics 13 (2001) 99–189.

[3] J. M. McQueen, L. Dilley, Prosody and spoken-word recognition, in: C. Gussenhoven, A. Chen (Eds.), The Oxford Handbook of Language Prosody, 2021, pp. 508–521.

[4] S. Greenberg, Speaking in shorthand—a syllable-centric perspective for understanding pronunciation variation, Speech Communication 29 (1999) 159–176.

[5] N. Morgan, H. Bourlard, H. Hermansky, Automatic speech recognition: An auditory perspective, in: S. Greenberg, W. A. Ainsworth, A. N. Popper, R. R. Fay (Eds.), Speech Processing in the Auditory System, Springer, New York, 2004, pp. 309–338.

[6] G. Coro, F. V. Massoli, A. Origlia, F. Cutugno, Psycho-acoustics inspired automatic speech recognition, Computers & Electrical Engineering 93 (2021) 107238. URL: https://doi.org/10.1016/j.compeleceng.2021.107238. doi:10.1016/j.compeleceng.2021.107238.

[7] C. S. Anoop, A. G. Ramakrishnan, Suitability of syllable-based modeling units for end-to-end speech recognition in sanskrit and other indian languages, Expert Systems with Applications 220 (2023) 119722. URL: https://doi.org/10.1016/j.eswa.2023.119722. doi:10.1016/j.eswa.2023.119722.

[8] C. J. Cho, A. Mohamed, S.-W. Li, A. W. Black, G. K. Anumanchipalli, Sd-hubert: Sentence-level self-distillation induces syllabic organization in hubert, arXiv (2024). URL: http://arxiv.org/abs/2310.10803.

[9] V. N. Vitale, F. Cutugno, A. Origlia, G. Coro, Exploring emergent syllables in end-to-end automatic speech recognizers through model explainability technique, Neural Computing and Applications 36 (2024) 6875–6901. URL: https://doi.org/10.1007/s00521-024-09435-1. doi:10.1007/s00521-024-09435-1.

[10] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, M. Zeng, X. Yu, F. Wei, Wavlm: Large-scale self-supervised pre-training for full stack speech processing, IEEE Journal of Selected Topics in Signal Processing 16 (2022) 1–14. doi:10.1109/JSTSP.2022.3188113.

[11] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, G. Weber, Common voice: A massively-multilingual speech corpus, arXiv (2020). URL: https://doi.org/10.48550/arXiv.1912.06670. doi:10.48550/arXiv.1912.06670.

[12] F. Schiel, A statistical model for predicting pronunciation, in: Proceedings of the ICPhS 2015, Glasgow, UK, 2015, p. paper 195.

[13] T. Kudo, 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), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 66–75. URL: http://arxiv.org/abs/1804.10959.

[14] D. Kahn, Syllable-based generalizations in English phonology, Ph.D. thesis, Massachusetts Institute of Technology, 1976. URL: https://dspace.mit.edu/handle/1721.1/16397.

[15] G. N. Clements, The role of the sonority cycle in core syllabification, in: J. Kingston, M. E. Beckman (Eds.), Papers in Laboratory Phonology: Volume 1: Between the Grammar and Physics of Speech, volume 1, Cambridge University Press, 1990, pp. 283–333. URL: https://doi.org/10.1017/CBO9780511627736.017. doi:10.1017/CBO9780511627736.017.

[16] G. Marotta, L. Vanelli, Fonologia e prosodia dell'italiano, Carocci editore, 2021.

[17] M. Krämer, The Phonology of Italian, Oxford University Press, Oxford, New York, 2009.

A. Appendix

A.1. Prediction types (PT)

Label	Prediction	Reference
eq_syl	do po | al ku ni |	do po | al ku ni |
sub_syl	mO do | ve tSo |	mO do | de tSo |
ins_syl	i | lo ro | a bi ta tta |	i | lo ro | a bi tat |
del_syl	kom ple ta men te | sO - |	kom ple ta men te | so lo |
sub_syl_word	kon | E | di ven ta to |	non | E | di ven ta to |
ins_syl_word	te | i |	ti |
del_syl_word	so pra ttu tto | - | ma ssa ka tSe ts |	so pra ttu tto | in | ma ssa tSu se tts |
mv_eq_forw_syl	o ri dZi | ni mi ti ke |	o ri dZi ni | mi ti ke |
mv_sub_forw_syl	E stre | ro u ma no |	E sse re | u ma no |
mv_eq_back_syl	da ve | tra te |	da | ve tra te |
mv_sub_back_syl	tu tta vi a no |	tu tta vi a | non |
div_eq_syl	a | pu ddZa | da	a ppo ddZa ta |
div_sub_syl	a | pu ddZa | da	a ppo ddZa ta |
div_ins_syl	fra | zi i |	fra zi |
mer_eq_syl	kwa ttro po sti |	kwa ttro | po sti |
mer_sub_syl	sE | la u re a to |	si | E | la u re a to |
mer_ins_syl	pu kwe stE ro no | kO lle	kwe stEr mo | ko lle |
mer_del_syl	fi nO - tto |	fi no | ad | O tto |

A.2. Summary of the model

y.level	term	estimate	std.error	statistic	p.value	conf.low	conf.high
deletion	(Intercept)	0.0201225	0.3193815	-12.2296295	0.0000000	0.0107603	0.0376305
deletion	freq_tok_R_catmid	1.7960895	0.3890354	1.5052919	0.1322490	0.8378774	3.8501310
deletion	freq_tok_R_cathigh	0.5827861	0.5518310	-0.9784428	0.3278554	0.1976013	1.7188128
deletion	stress_Runstr	2.0315288	0.3607487	1.9647709	0.0494408	1.0017356	4.1199589
deletion	freq_tok_R_catmid:stress_Runstr	1.1304773	0.4389646	0.2793846	0.7799497	0.4782054	2.6724478
deletion	freq_tok_R_cathigh:stress_Runstr	3.0560086	0.5878588	1.9003027	0.0573934	0.9655355	9.6725487
substitution	(Intercept)	0.3561515	0.0875308	-11.7946878	0.0000000	0.3000050	0.4228061
substitution	freq_tok_R_catmid	0.3962947	0.1468929	-6.3011683	0.0000000	0.2971548	0.5285107
substitution	freq_tok_R_cathigh	0.2504159	0.1906013	-7.2645468	0.0000000	0.1723541	0.3638329
substitution	stress_Runstr	0.7477364	0.1136480	-2.5579395	0.0105294	0.5984269	0.9342990

A.3. Pairwise comparison by stress

freq_tok_R_cat	pred_type	term	3	estimate	std.error	df	statistic	p.value
low	equal	stress_R	str - unstr	-0.04	0.02	12	-1.83	0.09
mid	equal	stress_R	str - unstr	0.04	0.02	12	2.24	0.05
high	equal	stress_R	str - unstr	0.10	0.02	12	6.08	0.00
low	deletion	stress_R	str - unstr	-0.02	0.01	12	-2.44	0.03
mid	deletion	stress_R	str - unstr	-0.04	0.01	12	-3.75	0.00
high	deletion	stress_R	str - unstr	-0.05	0.01	12	-6.12	0.00
low	substitution	stress_R	str - unstr	0.06	0.02	12	2.69	0.02
mid	substitution	stress_R	str - unstr	0.00	0.02	12	-0.20	0.85
high	substitution	stress_R	str - unstr	-0.06	0.02	12	-3.55	0.00

A.4. Pairwise comparison by frequency

stress_R	pred_type	term	3	estimate	std.error	df	statistic	adj.p.value
str	equal	freq_tok_R_cat	low - mid	-0.1228141	0.0218502	12	-5.6207337	0.0003371
str	equal	freq_tok_R_cat	low - high	-0.1817374	0.0216323	12	-8.4012049	6.8e-06
str	equal	freq_tok_R_cat	mid - high	-0.0589233	0.0190927	12	-3.0861663	0.0282878
unstr	equal	freq_tok_R_cat	low - mid	-0.044829	0.0166793	12	-2.6877091	0.0592601
unstr	equal	freq_tok_R_cat	low - high	-0.0400907	0.0162106	12	-2.4731219	0.0879759
unstr	equal	freq_tok_R_cat	mid - high	0.0047383	0.0153965	12	0.3077519	1.0
str	deletion	freq_tok_R_cat	low - mid	-0.0160783	0.0080354	12	-2.0009421	0.2056249
str	deletion	freq_tok_R_cat	low - high	0.0039688	0.006598	12	0.6015186	1.0
str	deletion	freq_tok_R_cat	mid - high	0.0200472	0.0081225	12	2.4681071	0.0887877
unstr	deletion	freq_tok_R_cat	low - mid	-0.0359457	0.0087751	12	-4.096334	0.0044462
unstr	deletion	freq_tok_R_cat	low - high	-0.0273429	0.008036	12	-3.4025497	0.0157348
unstr	deletion	freq_tok_R_cat	mid - high	0.0086028	0.0095059	12	0.9049905	1.0
str	substitution	freq_tok_R_cat	low - mid	0.1388925	0.0208492	12	6.6617705	6.96e-05
str	substitution	freq_tok_R_cat	low - high	0.1777686	0.0209563	12	8.4828288	6.2e-06
str	substitution	freq_tok_R_cat	mid - high	0.0388761	0.0176918	12	2.1974142	0.1450819
unstr	substitution	freq_tok_R_cat	low - mid	0.0807747	0.0150172	12	5.3788191	0.000497
unstr	substitution	freq_tok_R_cat	low - high	0.0674336	0.0148412	12	4.5436876	0.0020205
unstr	substitution	freq_tok_R_cat	mid - high	-0.0133411	0.0130966	12	-1.018664	0.9853835

A.5. Explanatory Legend for MCA Variables

Variable	Category	Description
event_syllable	deletion	Indicates the omission of a syllable
	substitution	Marks the replacement of a syllable with another one
	equal	Suggests no change in syllable token
freq_tok_R_cat	high_freq	Tokens that occur frequently in the dataset
	mid_freq	Tokens that have a moderate frequency of occurrence
	low_freq	Rare tokens with low frequency of occurrence
in_vocab_R	in_vocab_R_+	Tokens that are part of the vocabulary set
	in_vocab_R_-	Tokens not found in the vocabulary
POS (Part of Speech)	DET	Determiner
	NOUN	Noun
	VERB	Verb
	ADP	Adposition or preposition
	PRON	Pronoun
	AUX	Auxiliary verb
	CONJ	Conjunction
	RCONJ	Relative conjunction
stress_R	stress_R_+	Indicates that the token is stressed
	stress_R_-	Indicates that the token is unstressed
tok_type_R	CV	Consonant-Vowel syllable structure
	CVC	Consonant-Vowel-Consonant syllable structure
	CCVC	Consonant-Consonant-Vowel-Consonant syllable structure
	CCCV	Consonant-Consonant-Consonant-Vowel syllable structure