Interspeech 2024
1-5 September 2024, Kos, Greece

Prosodic marking of syntactic boundaries in Khoekhoe

Kira Tulchynska[1], *Sylvanus Job*[2], *Alena Witzlack-Makarevich*[1], *Margaret Zellers*[3]

[1]Hebrew University of Jerusalem, Israel
[2]University of Namibia, Namibia
[3]Kiel University, Germany

kira.tulchynska@mail.huji.ac.il, sylvanusjob@gmail.com, awitzlack@mail.huji.ac.il, mzellers@isfas.uni-kiel.de

Abstract

Khoekhoe (ISO 639-3: naq) is an understudied low-resource Khoe-Kwadi language spoken in Namibia. Khoekhoe is an SOV language with a strict syntactic clause-medial position: the first constituent of the clause is separated from the rest of the clause by a group of auxiliaries. We investigate the locations of candidate intonation phrase boundaries in relation to syntactic boundaries, observing the phonetic-prosodic characteristics of our candidate intonation phrases in order to confirm the initial phrasing analysis and compare boundary strength associated with different syntactic boundary types. We find evidence of differential prosodic marking strategies employed in clause-final versus the clause-medial position following auxiliaries, with, for example, stronger effects of final segmental lengthening following auxiliaries compared to rather attenuated final lengthening in clause-final position.

Index Terms: intonation phrase, prosodic boundary, prosody-syntax interface, Khoekhoe, Khoisan, Khoe-Kwadi

1. Introduction

1.1. The language and its speakers

Khoekhoe (*Khoekhoegowab* in the official use) is a Khoe-Kwadi language primarily spoken in Namibia, the Republic of South Africa and Botswana. With about 200,000 speakers Khoekhoe is the largest among the so-called "Khoisan" languages [1]. Only about a dozen of these languages belonging to three different genealogical units are still spoken today.

Among the Khoisan languages, Khoekhoe is relatively well described. It is recognized as a language of instruction in the Namibian educational system, though even this by far largest Khoisan language is occasionally claimed to be endangered [2].

1.2. Khoekhoe phonology

Like all other Khoisan languages, Khoekhoe is characterized by an extensive use of click consonants in their phonology [3, 4]. Furthermore, all Khoisan languages share a root phonotactic template with a skewed distribution of consonants and vowels. Each slot in the template has its own inventory and own set of distinctive features. The onset of the first syllable of the root (abbreviated as O) can be a simplex consonant or a cluster of two obstruents. The root-medial consonant (Cm) is always simplex. The coda (N) is always a non-click nasal consonant.

Roots regularly combine into compounds and various suffixes (in the shape of CV or C) attach to stems, so that phonological words have more variable shapes and can end in closed syllables.

Khoekhoe has both short and long vowels as well as oral and nasal vowels. All nasal vowels are phonologically long.

Table 1: *Khoisan root phonotactic templates*

Template	Example
OV_1CmV_2	*gowa* 'language'
OV_1V_2	*ǂgai* 'to call'
OV_1N	*ǂhon* 'to wobble'

Vowels are also characterized by an asymmetric distribution within the root template [3].

1.3. Khoekhoe morphology and syntax

Khoekhoe has a substantial amount of bound inflectional markers, including an elaborate system of person-gender-number suffixes on nouns and other elements of a clause.

The basic word order is SOV, though other word order are possible and common, e.g. the object can be fronted to the clause-initial position leading to the OSV order in case the subject is present as well. Also the verb including some auxiliaries can be fronted and yield clauses with the VSO order. Arguments are uncommon after the final verb and these structures are best analyzed as extra-clausal. This study includes clauses with any order and is not limited to SOV.

The clause structure has a few fixed slots: The first unit of a clause (e.g. an NP, a prepositional phrase, a conjunction or a fronted verb) is obligatorily followed by the subject marker. The subject marker is suffixed to the head noun if the first unit is the subject (*-s* in (2)). In all other cases, the subject agreement marker cliticizes to whichever element occupies the clause-initial position, as e.g. *=ta* for the first person singular subject ('I') in (1).

(1) Xawe =ta **ge** nēsisa ra ǂnâ!uā.
but =1SG.SBJ **DECL** AUX now IPFV understand CL
'But I understand now.'

(2) ti ǂnao-s **ge** **ge** ari-kha ūhâ i
my granny-3F.SG **DECL** **PST** AUX dog-2M.DU.OBL have STAT CL
'My granny had two dogs.'

The subject marker is immediately followed by a group of grammatical particles, which we label *auxiliaries* in this paper. They include one of the clause type particles (e.g. *ge* 'declarative') present in the majority of main declarative clauses, as in (1) and (2). The clause type particle is optionally followed by one or two tense-aspect auxiliaries. They include *ge* 'past', as in (2), *go* 'recent past', *nî* 'future/irrealis', *ra* 'imperfective' [5]. The frequent prosodic boundary after these auxiliaries (called *auxiliary-final* or AUX) and the clause-final boundary (CL) in

10.21437/Interspeech.2024-833

(1) and (2) are the focus of this paper. The tense-aspect auxiliaries can also occur immediately before the verb, e.g. the imperfective *ra* in (1). In this position they are not accompanied by a special prosodic boundary. Further tense-aspect particles can appear at the end of the clause, e.g. the stative non-present *i* ('STAT') in (2), and thus correspond to clause boundaries.

Most of the auxiliaries have the CV shape, only the future/irrealis *nî* has a CVV shape and the modal auxiliary *mūt* [mu:t] (borrowed from Afrikaans *moet* 'must') has a CVVC shape.

1.4. Intonation boundaries and syntactic units

Intonation phrases are commonly (though not exclusively) found to correspond with syntactic boundaries, see e.g. [6] on English. However, the relationship between intonation phrases and syntactic phrases can be flexible, especially in spontaneous speech, with intonation phrases able to be assigned to syntactic units of varying size and scope [7]. Gradient variation in intonational boundary marking has been found to correlate with the size of the syntactic units with which an intonational boundary coincides [8]; intonational boundaries correlating with larger syntactic units can be realized with, for example, an increased degree of segmental lengthening [9] or a tendency for lower fundamental frequency (F0) in phrase-final position [10].

Intonational phrasing has thus far not been studied in Khoekhoe. In the only full-length account of the grammar of Khoekhoe, [11] mentions that a common position for intonation boundaries is after the auxiliary slot and before the rest of the clause (the exact phrasing is vague and explicitly mentions only the first declarative particle *ge*). [12] has indirectly addressed some aspects of prosodic phrasing, noting that tone sandhi is constrained by prosodic position, and that superhigh tones cannot appear in non-initial position in prosodic constituents, a constraint that appears to be linked to phrase-internal downtrends in F0 (cf. [13] for downtrends in other African languages).

Since Khoekhoe, like other Khoe-Kwadi languages, is a tone language, it is fruitful to investigate non-pitch-related cues to phrase boundary marking. Thus the current study also investigates phrase-final segmental lengthening and changes in phonation quality. Final lengthening is often regarded as a universal phrase boundary cue [14, 15], making it a good reference point for evaluating other possible phrase boundary phenomena. Similarly, a number of studies have identified non-modal phonation, especially creaky voice, associated with prosodic phrase boundaries, although the majority of these studies have focused on European languages [16, 17].

1.5. Research questions

The aim of the current study is to provide a first description of intonational boundary marking in Khoekhoe. In particular, we investigate acoustic correlates of prosodic boundaries arising in two specific positions: auxiliary-final-clause-medial (hereafter: auxiliary-final, AUX) and clause-final (CL). We also assess the degree to which the acoustic features may be useful in classifying boundary type in Khoekhoe.

2. Data and methodology

2.1. Corpus

The analysed corpus includes speech from eight native speakers of Khoekhoe (4 female, 4 male, between 18 and 45 years old). Speakers in each conversation (three dialogues and one conversation among three colleagues) were well acquainted with each others: three conversations took place among friends, one conversation was between a married couple. The speakers come from a range of professional and socio-economic backgrounds and include teachers, farm workers, housewives, and students. All participants provided written informed consent for the recordings to be used in linguistic research. A public release of the corpus is in preparation.

The conversations were recorded in 2022 and 2023 in three locations in Namibia (Windhoek, Gibeon, Otjiwarongo). For two of the recordings, headset microphones (Shure WH20) and a video camera (Zoom Q8n-4K) were used. The other two conversations were recorded with a Zoom H8 handy recorder with the Zoom XYH-6 microphone capsule. For the latter two conversations, since complete channel separation was not possible, speech produced in overlap was excluded from the analysis.

2.2. Annotation

From the corpus, which we manually supplemented with interlinear morpho-syntactic glosses in ELAN [18, 19], we extracted 617 candidate intonation phrases that were surrounded by non-hesitation pauses lasting at least 120 ms. The selection of the phrases was carried out by the first author, with ambiguous cases decided by consensus agreement among the other authors. On the basis of the interlinear glosses and preceding morpho-syntactic analysis, these phrases were then manually annotated with one of three labels based on the right boundary: CL for clause-final boundaries, AUX for boundaries following an auxiliary, and OTHER for all other cases.

Each intonation phrase was segmented into syllables and phones using the BAS WebService [20] via the CHUNKPREP → G2P → MAUS → PHO2SYL pipeline for Language Independent, with attachment of an Imap mapping file for Khoekhoe, and with specification of the input tier name for G2P and CHUNKPREP (other options were default). Out of the 617 intonation phrases, 116 could not be processed by MAUS, possibly because BAS WebService pipelines are not trained on Khoekhoe. These phrases were excluded from further analysis.

Next, for each of the remaining 501 intonation phrases, the last four syllables were annotated for the presence or absence of a coda as well as for phonological vowel length (short vs. long, where the category of long vowels includes all nasal vowels and diphthongs). 78 intonation phrases had fewer than four syllables and were not included in the analyses, resulting in a final dataset comprising a total of 423 phrases. Then, the duration of the syllable rhyme, the mean F0 of the vowel, and local jitter (our operationalization of phonation quality) within the vowel were extracted using Praat [21]. The F0 (in semitones re: 1 Hz) and jitter measurements were calculated in Praat's editor, zooming into five seconds around the middle of the vowel and extracting the values using Praat's default settings, with minimum and maximum F0 adjusted for male and female speakers.

3. Results

3.1. Descriptive results

Of the 423 intonation phrases that were annotated, 234 (55.32%) fell at clause boundaries (CL), 48 (11.35%) were in the post-auxiliary position (AUX), and 141 (33.33%) were in other positions (OTHER). Variability in duration of the vowel in different positions is visualized in Figure 1. We implemented linear mixed models with package `lme4` [22] in R [23] to test

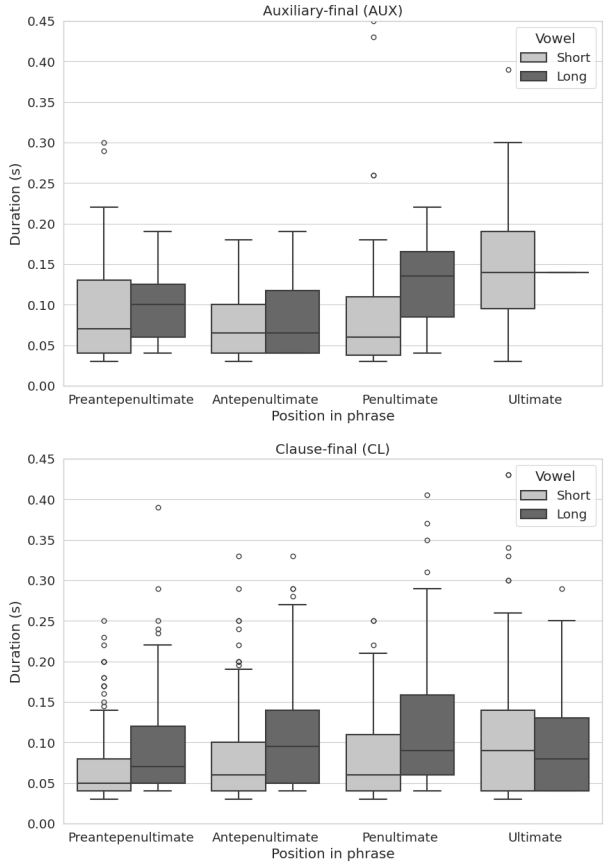

Figure 1: *Vowel duration at AUX (above) and CL (below) boundaries in the four phrase-final syllables.*

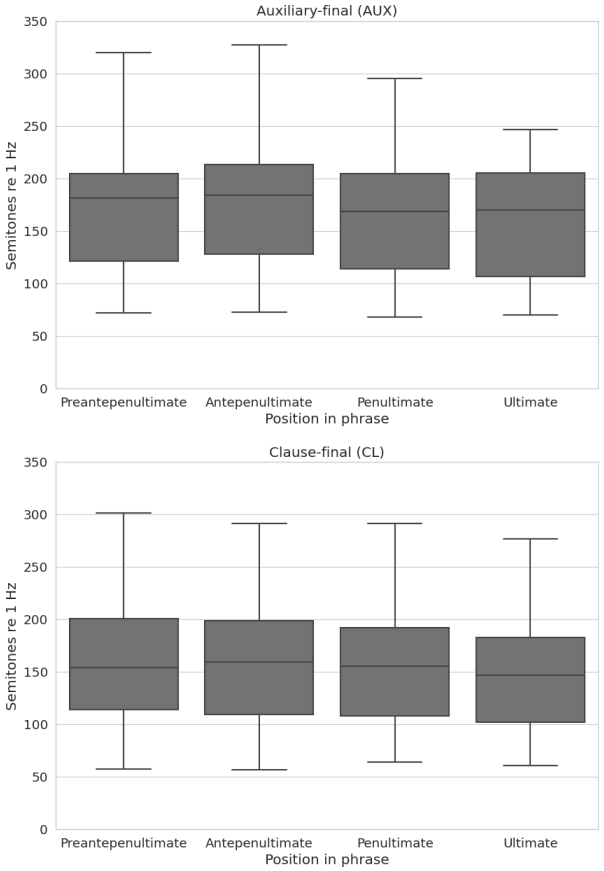

Figure 2: *F0 at AUX (above) and CL (below) boundaries in the four phrase-final syllables.*

whether the phrase boundary type as well as the distance of the syllable from the boundary had an influence on the vowel length, F0, and jitter measured in that syllable. Significance testing is carried out with $\alpha = 0.05$.

For purposes of the models, the order of the four final syllables is represented as a numeric variable *position*, with values being -3 (preantepenultimate), -2 (antepenultimate), -1 (penultimate), and 0 (ultimate).

Results for a linear mixed model investigating variation in length of the vowel are shown in Table 2. In addition to the fixed effects, the model contains a random intercept for speaker; adding a random slope did not improve model fit, so we adopted the simpler model. A significant main effect for final lengthening is found (*position*, where the intercept value represents the phrase-final syllable); there is also a significant interaction with the clause type, where final lengthening is mitigated in CL compared to AUX. Note that while the terms for vowel length (*long*) did not achieve significance in this model, the model including this term performed better than a model without it (model comparison using likelihood ratio testing: $\chi^2(2) = 32.2$, $p < 0.001$).

Results for a linear mixed model investigating variation in F0 are shown in Table 3. In addition to the fixed effects, the model contained a random intercept by speaker, in parallel with the length model. Overall, F0 is lower at CL and OTHER boundaries than at AUX boundaries; this difference is not mediated by nearness to the boundary in our data, although it is important to note that the estimates shown in the first two model coefficients are for the phrase-final syllable. Although the interaction between position and boundary type did not individually achieve statistical significance, we include the interaction in the F0 model to enable more direct comparability with the duration model.

The linear mixed model for jitter did not identify any significant differences in jitter related to nearness to the boundary or the type of boundary.

3.2. Classification

We explored the possibilities to differentiate between clause-final (CL) and auxiliary-final (AUX) intonation phrases by employing various classification models. These models utilized the acoustic features for which the dataset was annotated.

Out of the 48 phrases labeled AUX, all except one ended in a short final syllable, with the sole exception being the intonation phrase ending in the modal auxiliary *mūt* [mu:t]. Consequently, to ensure homogeneity in the dataset, only phrases with a final syllable lacking a coda and featuring a short vowel were included for the classification. This subset included 132 intonation phrases labeled CL and 47 labeled AUX.

Given the skewed nature of the data, with only 26.25% of phrases belonging to the AUX class, and the relatively small dataset size, we employed random resampling to under-sample the data for each of the models described below. Additionally, the data was scaled using the `StandardScaler()` function from the scikit-learn [24] library for Python. We then calculated

Table 2: *Results of the linear mixed model for syllable length. Syllables with short vowels in phrase position AUX are in the intercept. The variable position has its 0 at the final position in the phrase (thus earlier syllables are represented with negative values for position). Conditional R^2 = 0.070; marginal R^2 = 0.050.*

	Est	SE	t (p)
(Intercept)	0.13	0.01	13.22 (0.000)
long	0.01	0.01	1.68 (0.092)
CL	-0.03	0.01	-3.28 (0.001)
OTHER	-0.01	0.01	-1.12 (0.262)
position	0.02	0.00	4.40 (0.000)
long:position	-0.01	0.00	-1.87 (0.062)
CL:position	-0.01	0.01	-2.32 (0.020)
OTHER:position	-0.01	0.01	-1.30 (0.195)

Table 3: *Results of the linear mixed model for F0. Syllables in phrase position AUX are in the intercept. The variable position has its 0 at the final position in the phrase (thus earlier syllables are represented with negative values for position). Conditional R^2 = 0.648; marginal R^2 = 0.016.*

	Est	SE	t (p)
(Intercept)	174.30	23.00	7.58 (0.004)
CL	-19.09	4.60	-4.15 (0.000)
OTHER	-11.37	4.83	-2.35 (0.019)
position	-3.61	2.22	-1.62 (0.105)
CL:position	-1.29	2.45	-0.52 (0.600)
OTHER:position	0.28	2.58	0.11 (0.912)

the mean of the 5-fold cross-validation scores. This process was repeated 50 times with random seeds ranging from zero to 50, and the overall estimated accuracy score of the model was computed as the mean of the cross-validation scores across these 50 down-samplings.

The classification features comprised only the rhyme duration, presence or absence of a coda, and the distinction between short and long vowels for the last four syllables, with the final syllable having only the information on duration, since the dataset included only phrases with a short final syllable.We employed four classifiers from scikit-learn: k-Nearest Neighbors (KNN), Linear Support Vector Classifier (LinearSVC), Support Vector Classifier (SVC), and Random Forest Classifier.

For the KNN classifier, we calculated the mean cross-validation scores separately for each value of k ranging from 1 to 30 to determine the optimal k value for the dataset. The best estimated accuracy score was achieved with k=10, yielding 69.16%. Since the number of samples exceeded the number of features, LinearSVC was trained with `dual=False`, resulting in a mean estimated accuracy score of 64.9%. The SVC model utilized default settings, leading to mean estimated accuracy score of 70.42%. Lastly, employing a Random Forest Classifier model with `random_state=0`, we obtained a mean estimated accuracy score of 68.79%. The SVC model demonstrated the best performance among the four models.

Afterward, we employed the SVC model for classification using various subsets of features, such as the F0 and jitter of the last four syllables, and the slope of F0 between final and penultimate syllables. However, in all cases, the resulting estimated accuracy score was lower then 70.42% that was obtained with the SVC model set of features described above.

4. Discussion

We investigated acoustic features of intonation phrase boundaries in Khoekhoe, finding that the segment length as well as the F0 vary systematically at intonation phrase boundaries of different types, while phonation quality (modelled as jitter) did not show any consistent patterns related to intonation phrase boundary type. Using a classifier, we found that segmental length was able to predict the type of intonation phrase boundary with better than chance performance, while including F0 did not improve the performance of the classifier.

Our classifier results suggest that the difference in phrase boundary marking is in the degree of final lengthening. Our F0 results were likely affected by the fact that our analysis did not account for tonal structure. Nevertheless, we still found an overall effect for F0 declination over the final four syllables of phrases. Future analyses accounting for the specific tonal structure may be able to identify stronger effects of F0 declination at intonation phrase boundaries.

One important feature of our descriptive findings is that AUX boundaries and CL boundaries appear to be marked in differential ways. From a syntactic embedding point of view, AUX-final should be a lower-level boundary than CL-final. Thus, if prosodic boundary strength is linked to syntactic boundary strength, as argued by [8], acoustic marking of AUX-final boundaries should be less strong than acoustic marking of CL-final boundaries. However, in our data, we find conflicting results. AUX-final boundaries showed stronger marking via segmental lengthening, but an apparently lesser degree of F0 declination. CL-final boundaries, on the other hand, had more attenuated final lengthening, but a steeper slope for the F0 declination, which could be interpreted as a stronger boundary cue.

We interpret the results of this study in the following way: It is possible that the differential acoustic marking of AUX-final and CL-final boundaries represents the emergence of grammaticalized prosodic strategies to differ between the two types of syntactic phrase boundaries. These different types of phrase boundaries could then be described as "prosody-dependent constructions" (cf. [7]), where a specific prosodic organization is a necessary feature of the syntactic construction.

A weakness of this study is the somewhat unbalanced data set, which arose from our initial decision to use only intonation phrase boundaries which were adjacent to silence. This decision was motivated by the desire to avoid circular argumentation in the identification of intonation phrase boundaries. However, since we have identified differential marking in the AUX and CL positions, a follow-up study could expand the dataset by locating AUX positions not followed by silence but on the basis of their syntactic position exclusively.

5. Conclusions

Our results, showing differential acoustic marking of AUX-final and CL-final intonation phrase boundaries in Khoekhoe, highlight the importance of investigating prosodic phrase marking in a wider variety of languages than have previously been investigated, since different languages and language families may adopt dramatically different strategies for the alignment of syntactic and prosodic marking. Our work contributes to expanding understanding of the ways in which this alignment may be implemented or varied in different languages.

6. Acknowledgements

Tulchynska and Witzlack-Makarevich were supported by the project *Peripheral Khoekhoe varieties: A comprehensive documentation and description* funded by Israel Science Foundation (personal research grant no. 2892/20). Zellers was supported by a research grant from Kiel University.

7. References

[1] M. Brenzinger, "The twelve modern Khoisan languages," in *Khoisan languages and linguistics. Proceedings of the 3rd international symposium, July 6-10, 2008, Riezlern/Kleinwalsertal*, A. Witzlack-Makarevich and M. Ernszt, Eds. Cologn: Rüdiger Köppe, 2013, pp. 1–31.

[2] W. H. G. Haacke and E. Eiseb, *A Khoekhoegowab dictionary with an English-Khoekhoegowab index*. Windhoek: Gamsberg Macmillan, 2002.

[3] A. Witzlack-Makarevich and H. Nakagawa, "Linguistic features and typologies in languages commonly referred to as 'Khoisan'," in *The Cambridge Handbook of African Linguistics*, H. E. Wolff, Ed. Cambridge University Press, 2019, pp. 382–416.

[4] H. Nakagawa, A. Witzlack-Makarevich, D. Auer, A.-M. Fehn, L. A. Gerlach, T. Güldemann, S. Job, F. Lionnet, C. Naumann, H. Ono, and L. J. Pratchett, "Towards a phonological typology of the Kalahari Basin Area languages," *Linguistic Typology*, vol. 27, no. 2, pp. 509–535, 2023.

[5] W. H. G. Haacke, "Namibian Khoekhoe (Nama/Damara)," in *The Khoesan languages*, R. Vossen, Ed. London: Routledge, 2013, pp. 141–151, 325–340.

[6] W. Croft, "Intonation units and grammatical structure," *Linguistics*, vol. 33, no. 5, pp. 839–882, 1995.

[7] N. P. Himmelmann, "Prosodic phrasing and the emergence of phrase structure," *Linguistics*, vol. 60, no. 3, pp. 715–743, 2022.

[8] D. Watson and E. Gibson, "The relationship between intonational phrasing and syntactic structure in language production," *Language and Cognitive Processes*, vol. 19, no. 6, pp. 713–755, 2004.

[9] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1707–1717, 1992.

[10] M. Swerts, "Prosodic features at discourse boundaries of different strength," *The Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 514–521, 1997.

[11] R. S. Hagman, *Nama Hottentot grammar*. Bloomington: Indiana University Press, 1977.

[12] J. Brugman, "Segments, tones and distribution in Khoekhoe prosody," Ph.D. dissertation, Cornell University, 2009.

[13] I. Maddieson, "Phonetics and African languages," in *The Languages and Linguistics of Africa*, T. Güldemann, Ed. Berlin: De Gruyter Mouton, 2018.

[14] J. Fletcher, "The prosody of speech: Timing and rhythm," *The Handbook of Phonetic Sciences*, pp. 521–602, 2010.

[15] L. Paschen, S. Fuchs, and F. Seifart, "Final lengthening and vowel length in 25 languages," *Journal of Phonetics*, vol. 94, p. 101179, 2022.

[16] K. Surana and J. Slifka, "Is irregular phonation a reliable cue towards the segmentation of continuous speech in American English," in *Proceedings of Speech Prosody 2006*, 2006.

[17] T. Drugman, J. Kane, T. Raitio, and C. Gobl, "Prediction of creaky voice from contextual factors," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7967–7971.

[18] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "ELAN: a professional framework for multimodality research," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, and D. Tapias, Eds. Genoa, Italy: European Language Resources Association (ELRA), May 2006. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2006/pdf/153_pdf.pdf

[19] "Elan," Computer program, Nijmegen, 2024. [Online]. Available: https://archive.mpi.nl/tla/elan

[20] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, Sep. 2017.

[21] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," Computer program, 2023. [Online]. Available: http://www.praat.org/

[22] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[23] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2023. [Online]. Available: https://www.R-project.org/

[24] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.