Evaluating cluster analysis on f0 contours: an information theoretic approach on three languages

Constantijn Kaland, T. Mark Ellison

Institute of Linguistics, University of Cologne, Germany

ckaland@uni-koeln.de, tellison@uni-koeln.de

Abstract

This paper proposes a method for evaluating cluster analyses of f0 contours. Contour clustering was recently introduced as a theory-neutral approach to grouping f0 contours. This approach is applicable to any type of speech data (spontaneous to scripted) from any language. It lets the user explore how f0 shape differences relate to differences in linguistic content. Many cluster analysis methods leave the number of clusters as a parameter. The current paper proposes a method for identifying the ideal value for this parameter for a given dataset. In particular, we use a the Bayes-equivalent method of finding the minimum description length (MDL). This method is illustrated here using f0 data from three typologically different languages, obtained using different elicitation methods. The results show that the MDL measure selects cluster counts corresponding natural classes identified by humans.

Index Terms: cluster analysis, f0 contour, evaluation metric, information theory.

1. Introduction

A well known question in prosody research is how form and function are related in languages (e.g. [1]). For example, how does the shape of an f0 contour relate to its communicative function in a certain language? Although research has provided different approaches to this issue, much work still focusses on a small number of well-known languages [2]. In addition, autosegmental metrical approaches are frequently applied to identify an inventory of pitch accents and boundary tones [3], [4]. Often, the resulting description of intonation in a given language is limited to a small number of fixed functions, commonly based on scripted speech from a small number of speakers. These descriptions rarely represent natural, spontaneously produced speech, which is in part the result of methodological limitations. In order to lower the workload threshold for including less-known languages in prosody research, and in order to incorporate more diverse speech data, new methods need to be developed. Most importantly, these new methods need to work with all languages and improve the analysis of their prosody and intonation. This paper proposes a novel way of evaluating cluster analyses of f0 contours.

Cluster analysis (CA) has recently been proposed as a means of identifying prototypical f0 contours [5]. This method starts with N f0 contours (individual observations) each in a singleton cluster. Larger clusters are formed by merging smaller clusters (agglomerative hierarchical clustering). Each contour is numerically represented by time-series measurements (vectors) of the fundamental frequency (either Hz, ST or a speaker-normalized conversion). Merging clusters into larger ones halts only when all observations form a single cluster. Clusters are merged deterministically based on 1) the distance between all possible pairs of observations as expressed in a distance matrix, here calculated using Euclidean distances [6] and 2) the linkage criterion used when clusters with multiple observations need to be merged, here *complete linkage*. Complete linkage "iteratively merges two clusters from the current CA that have the smallest diameter when merged into a single cluster" ([7], p.1131). This linkage criterion maintains maximal dissimilarity between the clusters, in order to obtain the largest differences between contours from unmerged clusters. The output of the hierarchical clustering is a dendrogram; a tree structure showing the entire merging process from separate clusters for each individual observation (bottom) to all observations in a single cluster (top). The height at which one analyses the dendrogram corresponds to the number of clusters in the analysis.

Hierarchical clustering, unlike K-means clustering, performs the analysis and allows for the resulting dendrogram to be inspected *before* deciding on the number of clusters. However, determining the number of clusters often remains a key issue in CA. There are numerous statistical methods available to determine the number of clusters based on several aspects of the CA [8]. However, many of these evaluation methods are generic, i.e. not tailored to the type of data under analysis. A coarse division has been made between evaluation of a 'goodness of fit' given a within and between cluster variation (global measures) and the evaluation of whether or not a cluster should have been merged with another one (i.e. not clustered) or subdivided into more clusters (local measures) [9]. A common way to find the number of clusters is by finding the point at which the variance explained by the CA no longer increases such that it is worth assuming more clusters (so called 'elbow' or 'knee' method). Such a stopping-criterion is entirely based on the fact that with more clusters, more variance is explained. In the approach outlined in this paper, we account for multiple factors that play a role in evaluation the CA, specifically when applied to f0 contours. The choice of the number of clusters is particularly crucial when the analysis seeks to reveal an inventory of phonologically distinct contours ([3] [4]) from data in which there is (potentially) a high degree of 'allophonic' contour variation. Thus, the two main competing factors to take into account here are explaining contour variation (seeks a higher number of clusters) whilst finding a compact set of contour prototypes (seeks a low number of clusters). Allowing many clusters may reveal smaller and smaller (potentially uninformative) yet real differences between the clusters (overfitting). At the same time, assuming that there are contour prototypes (as done in most theories of intonational phonology, e.g. [1]) implies to look for contours that can be subsumed under the same type. The approach outlined in this paper assumes that the ideal number of contour clusters defines an optimum between those competing factors. It is important to note that choosing the right number of clusters remains a challenge, particularly when exploring the prosody of an understudied language. There might be lit-

tle to no background knowledge about which f0 movements are meaningful, a dendrogram might not provide sufficient directions, and it is important not to rely on distinctions perceived by the researcher which will reflect their language background. Thus, it is crucial to evaluate the CA based on the f0 variation in the dataset separate from interpreting the outcome based on other knowledge about the language (e.g. previous literature).

This paper proposes to evaluate the number of clusters using the Minimum Description Length (MDL) principle of [10]. MDL is based Shannon's (e.g. [11]) theory of information. Shannon related the level of information in a random variable X, with the probabilities of its $(x_i)_i$ possible outcomes (see 1).

$$\mathrm{H}(X) = -\sum_{i=1}^{\mathrm{n}} \mathrm{p}(x_i)\log(\mathrm{p}(x_i)) \quad (1)$$

The Shannon entropy (or information cost) for rolling a dice (6 possible outcomes) is -(6*1/6)*log(1/6) = 1.79, where entropy is expressed in nats (natural bits, each around 1.44 binary bits). In contrast, the entropy of flipping a coin is much smaller: $-2 * \frac{1}{2} * \log(\frac{1}{2})$ = 0.69. Events with low probability contain more information. The entropy measure can be used to find the least information needed to describe a certain dataset. By Ockham's Razor (e.g. [12]), this will be the best account. Hence we propose the model-selection criterion of Minimum Description Length (MDL, [10]).

In order to find the number of clusters that best fits a speech dataset according to MDL, all information needed to describe the data given the clustering model must be evaluated. This includes 1) the cost of specifying a mean contour for each cluster. This cost is expected to increase with more clusters. The measure also must take into account 2) the cost of classifying each contour in the data into the corresponding cluster. This quantity is likely to increase with more clusters. Finally, 3) the cost of specifying each contour, given that we know the mean contour for that cluster. This quantity will decrease with more clusters, if the CAs are sensible, as in smaller clusters the variance from the mean will be less. Given that (1) and (3) most clearly differ in direction with higher number of clusters, the sum of the three information cost measures is expected to approximate a U-shape, with the lowest point indicating the least information cost, corresponding to the MDL of the dataset. In short, the total information cost (IC) per round of CA is the sum of:

1. IC of expressing the mean contour per cluster relative to the contour average across the whole dataset,
2. IC of expressing for each contour what cluster it is in,
3. IC of expressing each contour relative to the mean contour of the cluster it belongs to.

In the remainder of the paper we illustrate the application of the above evaluation metric using data from three typologically different languages. We have chosen datasets with elicited data in order to obtain a high level of control over the number of contours produced. Thus the accuracy of the evaluation metric can be assessed by comparing the MDL with the knowledge of the context in which the contours were elicited. The evaluation method is coded in R [13], built-in and shared along with the tool available for contour clustering ([5], ver. 2022-04).

2. Datasets

The following subsections describe the three datasets to which we have applied contour clustering and the proposed evaluation metric. The datasets come from typologically different languages (German, Papuan Malay and Zhagawa) and from different constituents (intonation phrase, noun phrase, syllable). The investigation in this paper compares the number of contours as hypothesized in the respective studies with the number of contours suggested by the evaluation metric *within* each language. The following subsections list relevant aspects of the experiments used to collect the speech data. Additional methodological details (e.g., speakers, number of items) are listed in the respective papers and not repeated here for space reasons.

2.1. German

The German data came from the standard variety spoken in Germany (ISO: deu). The elicitation took the form of a discourse completion task in which participants responded to a pre-recorded speaker by reading out loud written stimuli. The stimuli were presented so that the final sentence (target utterance) occurred in different speech acts (exclamative [E] or question [Q]) and in different focus conditions (given [G] or contrast [C]). For example, the pre-recorded speaker would say: "Anna hat sich in ihrer Dissertation jetzt auf Germanen spezialisiert" (*Anna is doing a PhD in history and has specialized in Germanic peoples*). The participant would then respond with:

[E-G] "Ja, das hat sie mir neulich erzählt. Sie ist irre viel unterwegs, um an Originalquellen von Germanen heranzukommen. Wo die schon überall Germanen erforscht hat!"
Yes, I heard that, too. Anna has been traveling a lot for this. The places where she's done research on the Germanic peoples!

[Q-G] "Wirklich? Da ist sie bestimmt viel unterwegs, um an Originalquellen von Germanen heranzukommen. Weißt du zufällig, wo die schon überall Germanen erforscht hat?"
Really? I'm sure Anna has been traveling a lot for this. Do you know more? Where has she done research on the Germanic peoples?

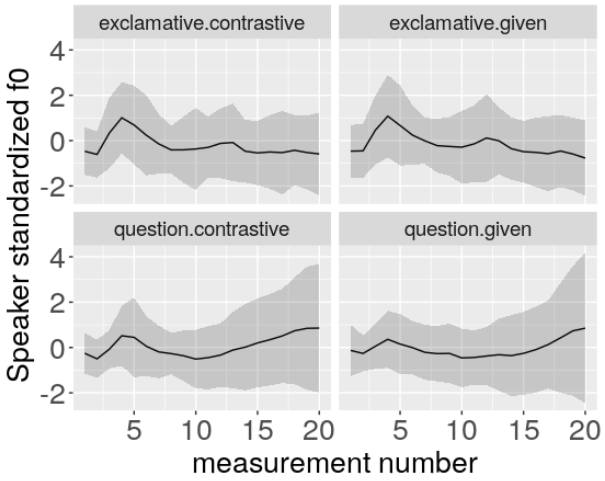

Figure 1: *DEU: mean f0 contours in each condition.*

In the contrastive context, the pre-recorded sentence has *Paul* as the subject of the first sentence and *Anna* is introduced as a contrastive referent in the pre-final sentence read by the participant. See detailed descriptions of the stimulus materials and procedure in [14] and [15]. The f0 contours of the recorded target utterances were measured using 20 points per contour, which were equally distributed over the contour. The mean f0 in each experimental condition shows that exclamatives end low,

whereas questions end high (Figure 1). Furthermore, there appears to be more subtle differences between the focus conditions in that the height and the width of the initial f0 peaks are larger for contrastive contexts than for given contexts. Thus, at least four different clusters are expected when all the (subtle) differences can be successfully clustered.

2.2. Papuan Malay

Papuan Malay (ISO: pmy) is spoken in the Eastern-Indonesian provinces Papua and West-Papua [16]. The data comes from elicitations of contrastively focused noun phrases (noun-adjective) that refer to pictures differing in shape and color (details in [17]). The descriptions all have the same structure (matrix sentences), referring to a picture displayed on the left side of a screen (antecedent A) and a picture displayed on the right side of the screen (target T). The noun phrases occur in either medial or final phrase position, for example:

"Saya liat [A] di sebla kiri, dang saya liat [T] di sebla kanang."
I see [A] on the left side, but I see [T] on the right side.

"Di sebla kiri saya liat [A], dang di sebla kanang saya liat [T]."
On the left side I see [A], but on the right side I see [T].

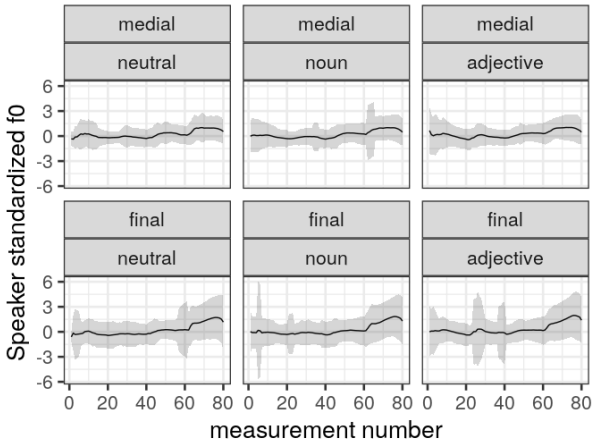

Figure 2: *PMY: mean antecedent f0 contours in each condition.*

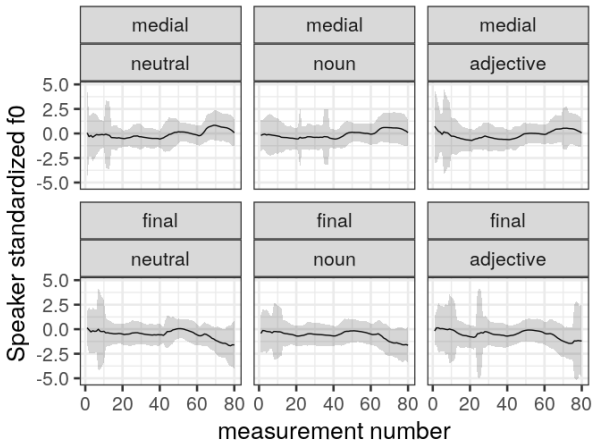

Figure 3: *PMY: mean target f0 contours in each condition.*

The noun phrases referring to the pictures expressed a shape contrast; eliciting focused nouns (e.g., A: babi hitam, T: pisang hitam, *black banana - black pig*), a color contrast; eliciting focused adjectives (e.g., babi hitam - babi mera, *black pig - red pig*), or no contrast at all (e.g. A: babi hitam, pisang mera, *black pig, red banana*). All nouns and adjectives were bisyllabic. The f0 contour of each of the four syllables in the target noun phrase was measured using 20 points per syllable. The mean contours of the entire noun phrase (80 points) in each of the experimental conditions (phrase position: medial, final; focus: neutral, noun, adjective) are shown in Figure 2 (A) and Figure 3 (T). The contours do not appear to differ between the focus conditions. For both A and T there appears to be a difference between the phrase positions: in medial position, the final syllable shows a mid level contour (invariable between A and T), whereas in final position antecedents end high and targets end low. Thus, three clusters is hypothesized to be a sufficient number to describe the differences between the Papuan Malay contours in this dataset.

2.3. Zhagawa

Zhagawa (ISO: zag) is a tone language spoken in Darfur, Africa. The data was obtained from one speaker (no f0 speaker normalisation) and consists of elicitations of body parts, kinship terms, animals and colours for the purpose of investigating number marking by tone ([18], see [5] for details). F0 contours (20 measurements) were taken from the final syllable of 212 elicited words. The final syllable is the reported location for tonal number marking ([19],[20],[21]). These analyses differ in the number of tones are used to mark number. A preliminary CA indicated the highest success rate of distinguishing singulars from plurals with six clusters [5], corroborating the analysis in [19], in which three register tones precede a modulated tone (LH, LM, MH, ML, HM, HL). The state of the research for Zhagawa represents an ideal test case for the evaluation metric; there are hypotheses about the use of tone for number marking, but the exact inventory of tones still remains uncertain.

3. Cluster analyses evaluation

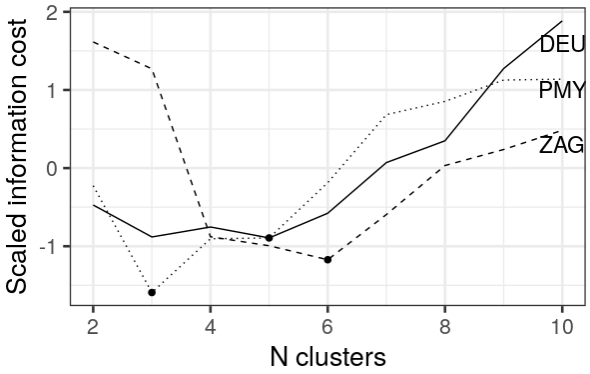

Figure 4: *Scaled information cost per number of clusters for each of the languages. Dots indicate MDL.*

The f0 measures from the DEU, PMY and ZAG datasets were run through several rounds of cluster analyses (2-10 clusters). The output CAs were subsequently measured for their information costs and then scaled to compare them in a line graph (Figure 4). The MDLs indicated 5, 3 and 6 as optimal cluster numbers for the respective data sets. In what follows, the mean contours per cluster are plotted and discussed per language.

For German, analysis with 5 clusters (Figure 5) indicates that the majority of contours in cluster 2 (60/68) and 5 (21/22) are produced when the elicited speech act concerned exclamatives. In these clusters a final rise can be observed. The other clusters also show a majority of exclamatives, although with relatively lower percentages; cluster 1: 67/78, cluster 3: 53/94, cluster 4: 25/35. No (contrast/given) cluster shows a

clear majority of subject focus for either category. Cluster 2 (37/68) and cluster 4 (16/35) show the highest percentage of contours elicited in the contrastive focus condition. Accent ratings in a perception task revealed that some participants placed a pitch-accent on the object *Germanen* when the subject was contrastively focused, which led to an increased f0 towards the end of the phrase. This is exactly what can be observed for the contours in cluster 2 and 4. In this way, the object-accented contours have their unaccented object counterparts in cluster 1 (exclamatives) and 2 (questions) respectively. The contours in cluster 3 show smaller excursions overall, and most often occur in exclamative, given contexts.

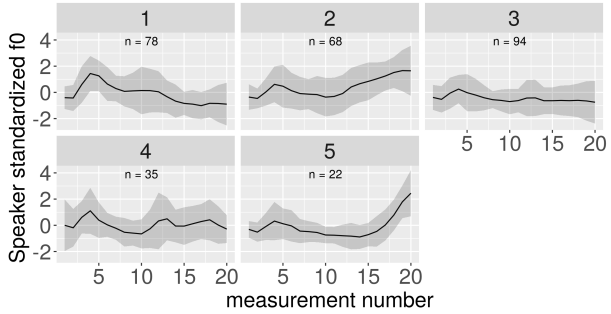

Figure 5: *DEU: mean f0 contour in each cluster (N = 5).*

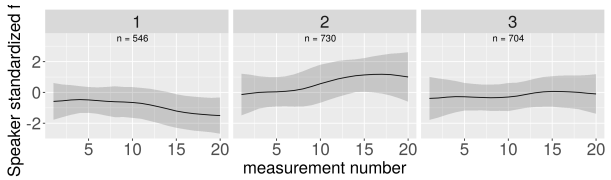

Figure 6: *PMY: mean f0 contour in each cluster (N = 3).*

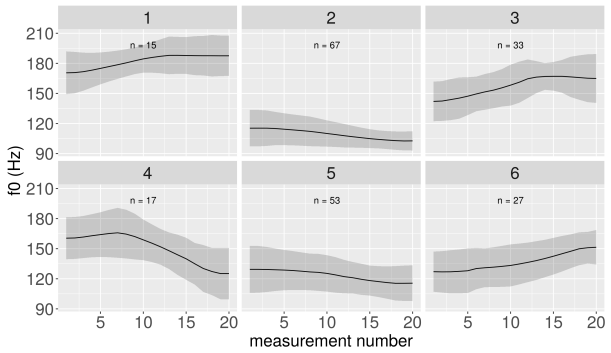

Figure 7: *ZAG: mean f0 contour in each cluster (N = 6).*

The Papuan Malay analysis with 3 clusters reveals overall shallower contours than German. This may be a result of the contour length: an entire phrase is used in German and a noun phrase in Papuan Malay. The differences between the contours can best be described by looking at the final syllable in the noun phrase (measurement points 60-80). This part shows a final fall (cluster 1), final rise (cluster 2) or a sustained mid-level f0 (cluster 3). Note that these three prototypes match the variation observed in Figure 2 and 3, with falls mainly produced in target final position, rises in antecedent final position and sustained f0 in medial positions of both antecedent and target.

The analysis with 6 clusters for Zhagawa provides a close match with one of the analyses in the literature [19]. That is, cluster 1: MH, cluster 2: ML, cluster 3: LH, cluster 4: HL, cluster 5: HM, cluster 6: LM. It should be observed that the absolute levels do not match the registers perfectly (c.f. cluster 2 and 5). In addition, the CA performed in [5] indicated the most accurate distinction between singulars and plurals in this dataset using 6 clusters (83.33%). This accuracy level did not improve when more clusters were permitted.

4. Discussion

The evaluation metric applied to these three data sets can be used widely to identify the number of clusters, and corresponding prototypes, in a set of f0 contours. Note that in addition to the typological differences, the datasets concerned different (prosodic) constituents (Sec. 2) and different prosodic phenomena (DEU: speech act/focus, PMY: phrase position/focus, ZAG: number marking). Still, MDL selected a number of categories for each dataset that could be interpreted as a plausible outcome for each of the investigated languages. The evaluation metric was able to discover relevant contour differences that were not manipulated in the experimental setup (DEU) as well as to reduce the contour differences to the most essential ones, even lower than the number of experimental conditions (PMY).

It should be noted that the number of clusters chosen by MDL remains sensitive to a number of factors related to the dataset. Most importantly, fine-grained differences that are potentially linguistically relevant do not necessarily show if there are larger-scale differences. For example, the f0 range of boundary tones generally comes out in different clusters before scaling or alignment differences of pitch accents are revealed (see also discussion in [5]). This is exemplified by the German data in this paper. Although some unexpected differences were found, the potential differences between contrastive and non-contrastive pitch accents on the subject were not entirely revealed with 5 clusters. It is likely that the unexpectedly accented object masked some of these differences, however, more research is needed to confirm this hypothesis.

Evaluating hierarchical cluster analyses using entropy-based metrics of information cost are a crucial step towards a principled choice of the number of clusters for a particular dataset. More general critical notes should be made, however. The outcome of the analyses presented here do not constitute a basic set of phonologically relevant f0 movements that compare to an autosegmental-metrical analysis. This is in part due to the fact that none of the datasets used here consisted of all possible contours of a language. In this respect it is important to note that in the typological overviews ([3],[4]) the inventories of phonologically relevant f0 movements had a weaker empirical basis. Contour clustering is able to bridge this gap by obtaining reproducible and entirely data-driven results showing which contours are (linguistically) relevant. It should also be noted that the outcome in terms of prototypical contours obtained by CA does heavily rely on the type of data under investigation. None of the datasets used in this paper are fully representative of *the prosody* of a certain language. This is a natural result of the data being collected in specific contexts. It is therefore crucial that multiple carefully-obtained datasets from one language are combined in order to provide a representative sample. In addition, perception research should be incorporated to further verify hypotheses created on the basis of any contour clustering.

5. Acknowledgements

Research for this paper was funded by the German Research Foundation (DFG) Project-ID 281511265 SFB 1252. The authors thank Heiko Seeliger for providing the German data.

6. References

[1] D. R. Ladd, *Intonational phonology*, 2nd ed., ser. Cambridge studies in linguistics. Cambridge ; New York: Cambridge University Press, 2008.

[2] N. P. Himmelmann and D. R. Ladd, "Prosodic Description: An Introduction for Fieldworkers," *Language Documentation & Conservation*, vol. 2, no. 2, pp. 244–274, Dec. 2008. [Online]. Available: http://scholarspace.manoa.hawaii.edu/handle/10125/4345

[3] S.-A. Jun, Ed., *Prosodic typology: the phonology of intonation and phrasing*, ser. Oxford linguistics. Oxford ; New York: Oxford University Press, 2005.

[4] ——, *Prosodic typology II: the phonology of intonation and phrasing*, ser. Oxford linguistics. Oxford: Oxford University Press, 2014.

[5] C. Kaland, "Contour clustering: A field-data-driven approach for documenting and analysing prototypical f0 contours," *Journal of the International Phonetic Association*, pp. 1–30, 2021. [Online]. Available: https://doi.org/10.1017/S0025100321000049

[6] S. Aghabozorgi, A. Seyed Shirkhorshidi, and T. Ying Wah, "Time-series clustering A decade review," *Information Systems*, vol. 53, pp. 16–38, Oct. 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0306437915000733

[7] A. Growendt and H. Rglin, "Improved Analysis of Complete-Linkage Clustering," *Algorithmica*, vol. 78, no. 4, pp. 1131–1150, Aug. 2017. [Online]. Available: http://link.springer.com/10.1007/s00453-017-0284-6

[8] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs, "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set," *Journal of Statistical Software*, vol. 61, no. 6, 2014. [Online]. Available: http://www.jstatsoft.org/v61/i06/

[9] A. D. Gordon, *Classification*. Boca Raton: Chapman & Hall/CRC, 1999, oCLC: 152674146. [Online]. Available: http://www.crcnetbase.com/isbn/9781584880134

[10] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, Sep. 1978. [Online]. Available: http://doi.org/10.1016/0005-1098(78)90005-5

[11] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, Jul. 1948. [Online]. Available: http://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[12] E. Sober, *Ockham's razors: a user's manual*. Cambridge: Cambridge University Press, 2015.

[13] R Core Team, "R: the R project for statistical computing," 2020. [Online]. Available: https://www.r-project.org/

[14] S. Repp, "The Prosody of *Wh* -exclamatives and *Wh* -questions in German: Speech Act Differences, Information Structure, and Sex of Speaker," *Language and Speech*, vol. 63, no. 2, pp. 306–361, Jun. 2020. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0023830919846147

[15] H. Seeliger and C. Kaland, "Boundary tones in German wh-questions and wh-exclamatives - a cluster-based approach," in *Proceedings of the 11th International Conference on Speech Prosody 2022*, S. Frota and M. Vigrio, Eds., Lisbon, Portugal, 2022, pp. 27–31.

[16] A. Kluge, *A grammar of Papuan Malay*. Berlin, Germany: Language Science Press, 2017. [Online]. Available: http://langsci-press.org/catalog/book/78

[17] C. Kaland and N. P. Himmelmann, "Time-series analysis of F0 in Papuan Malay contrastive focus," in *Proceedings of the 10th International Conference on Speech Prosody 2020*, 2020, pp. 230–234. [Online]. Available: http://dx.doi.org/10.21437/SpeechProsody.2020-47

[18] Language Archive Cologne, "Beria," 2018, publisher: re3data.org - Registry of Research Data Repositories. [Online]. Available: http://doi.org/10.17616/R3JV4W

[19] Tourneux, H., "Inventaires phonologiques et formation du pluriel en zaghawa (Tchad)," *Afrika und bersee*, vol. 75, no. 2, pp. p.267–277, 1992, place: Hamburg. [Online]. Available: https://db.degruyter.com/view/IABO/iab19933512

[20] A. M. Wolfe, "Towards a Generative Phonology and Morphology of the Dialects of Beria," Master's thesis, Harvard University, 2001, pages: 110.

[21] S. N. Osman, "Phonology of the Zaghawa Language in Sudan," in *Proceedings of the 9th Nilo-Saharan Linguistics Colloquium*, ser. Nilo-Saharan, A.-A. Abu-Manga, L. G. Gilley, and A. Storch, Eds. University of Khartoum: Kln: Kppe, 2006, meeting Name: Nilo-Saharan Linguistics Colloquium OCLC: ocm77077755.