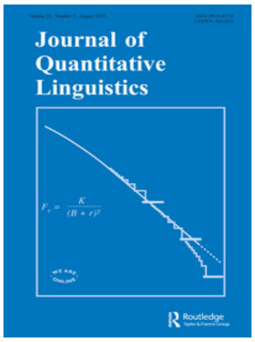

Journal of Quantitative Linguistics

ISSN: 0929-6174 (Print) 1744-5035 (Online) Journal homepage: https://www.tandfonline.com/loi/njql20

Quantitative Comparative Linguistics based on Tiny Corpora: *N*-gram Language Identification of Wordlists of Known and Unknown Languages from Amazonia and Beyond

Frank Seifart & Roger Mundry

To cite this article: Frank Seifart & Roger Mundry (2015) Quantitative Comparative Linguistics based on Tiny Corpora: *N*-gram Language Identification of Wordlists of Known and Unknown Languages from Amazonia and Beyond, Journal of Quantitative Linguistics, 22:3, 202-214, DOI: 10.1080/09296174.2015.1037161

To link to this article: https://doi.org/10.1080/09296174.2015.1037161

© 2015 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 09 Jul 2015.

Submit your article to this journal

Article views: 302

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=njql20

Journal of Quantitative Linguistics, 2015
Vol. 22, No. 3, 202–214, http://dx.doi.org/10.1080/09296174.2015.1037161

Quantitative Comparative Linguistics based on Tiny Corpora: *N*-gram Language Identification of Wordlists of Known and Unknown Languages from Amazonia and Beyond*

Frank Seifart[1,2] and Roger Mundry[3]
[1]Department of Linguistics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany; [2]Amsterdam Center for Language and Communication (ACLC), University of Amsterdam, Amsterdam, The Netherlands; [3]Department of Primatology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

ABSTRACT

Can an unknown Amazonian language be identified by statistical procedures based on *n*-gram frequencies if only a short list of words is available and at the same time, the available data of the potential candidate languages are also limited to relatively short wordlists? In this paper we show that *n*-gram frequencies (specifically 1-grams and 2-grams) allow us to identify languages reliably based on as few as 20 words, as long as these are transcribed consistently, and as long as characteristic monogram and bigram frequencies for these languages have previously been established based on consistently transcribed data. If no such consistently transcribed data are available, as is the case of our Amazonian case study, such procedures clearly fail for wordlists with 50 or fewer words. Our study thus contributes to exploring the limits of such automated detection procedures, both in terms of corpus size and transcription quality.

1. INTRODUCTION

Automated language identification has been shown to work reliably when large amounts of consistently transcribed data are available, such as internet corpora. Such data allow identifying languages not only by characteristic, relative frequencies of phonemes or characters (i.e. monograms) or of sequences of two of these (i.e. bigrams), but also through characteristic, highly frequent words (or parts of words) such as articles. Recently,

*Address correspondence to. Frank Seifart, Linguistics, Universiteit van Amsterdam, Spuistraat 210, 1012 VT Amsterdam, The Netherlands. Tel: +31 20 525 4557. E-mail: f.c.seifart@uva.nl

© 2015 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

methods have also been developed specifically for the identification of short texts, e.g. queries of only a few words length (Řehůřek & Kolkus, 2009; Gottron & Lipka, 2010), but these procedures still require large amounts of training data from the candidate language.

The point of departure for the current study is a case study of the Amazonian language called Carabayo which is of unknown affiliation and spoken by a tribe living in voluntary isolation in the Colombian Amazon region (Franco, 2012; Seifart & Echeverri, 2014). A short list of words in their language was noted down (most of them without translation) from conversations among Carabayos that was overheard in 1969 during a brief encounter with one Carabayo family. During this encounter, it was established that the Carabayo language is mutually unintelligible with any of the living indigenous languages of the area. It may well be, however, that the Carabayo language corresponds (more or less closely) to one of a number of languages of the area that were documented in wordlists in the 19th century but are thought to have become extinct since then. The quality of transcription for all of this material is poor, i.e. we expect inconsistencies and errors in the graphic representation of sounds, as well as in the segmentation of the phonetic material into words.

Here we explore the potential of statistical procedures to identify the Carabayo language as being one of these extinct languages, or at least as being clearly more similar to one of these languages than to the others. To evaluate our approach, we assembled a set of comparable data from phonetic transcriptions of specimens of naturally occurring spoken language and of wordlists from languages of known origin (German, Polish, etc.). Our approach was successful for these test data, providing clear indications of the language identity based on as few as 20 words (unknown language) with as few as 200 words for candidate languages. This finding contributes to our understanding of the minimal amount of data necessary for addressing such questions. However, our approach was unsuccessful in the Amazonian case study, suggesting that such procedures crucially rely on consistent orthographic representation of data.

2. DATA

The entire existing Carabayo data consist of 52 words that a Capuchin monk in 1969 overheard the Carabayo people say and noted down, without knowing what these words mean (see Seifart & Echeverri, 2014). We identified, from the ethnohistorical record, five languages as possible candidates to identify Carabayo. These languages were reportedly spoken in the same area, some of them until the early 20th century: Coeruna, Curetú, Yurí, Passé and Uainuma. All of them are presumed to be extinct now but were documented in wordlists in the 19th century. These wordlists were collected by the German botanists Martius and Spix in the early 19th century and were published in Martius (1867). These lists range in length from around 150 to 250 words. They follow the same scheme, including basic vocabulary such as kinship and body part terms as well as some local flora and fauna.

For the evaluation of our approach, we assembled a comparable set of data from languages of known affiliations. As a set of data comparable to the Carabayo data we chose random samples of 5, 10, 20, 50, or 100 words, respectively, of conversational German, taken from a corpus of 11,500 words of conversational spoken German, specifically from the German subtitles of the movies *007 – Quantum of Solace*, *101 Dalmatians*, and one episode of the television series *24.* As data comparable to that of our candidate languages we chose standard 200-word lists of basic vocabulary (the so-called Swadesh lists) for German, Dutch, Polish, French, Burmese and Arabic. All data were transliterated to ASJP orthography, a simplified phonological transcription system used in quantitative language comparison (see Brown, Holman, Wichmann, & Velupillai, 2008).

3. METHODS

Our approach consists of calculating the likelihood of the unknown language to be a sample from each of the candidate languages. A comparison of the obtained likelihoods then reveals the strength of evidence in favour of any of the candidate languages. This approach is comparable to a naïve Bayesian approach which Gottron and Lipka (2010) have shown to work best on small corpora, when compared to approaches based on, e.g. Markov processes or frequency ranks.

More specifically, to determine the most likely language of origin among the candidate languages we compared the probabilities of finding the observed frequency distribution of patterns (mono- or bigrams; see below) in Carabayo given the frequency distributions of the respective patterns in the candidate languages. This comparison was based on likelihoods. In a first step we thus determined for each unique pattern in the unknown and the candidate languages its frequency of occurrence (absolute number) in each of the language samples (Figures 3–5). For the candidate languages we considered 800 randomly selected patterns each, which correspond to the amount available in the shortest of the wordlists of candidate languages, to ensure similar precision of word frequencies for all candidate languages.

Then we determined for each of the patterns found in Carabayo its frequency of occurrence in each of the candidate languages. When a pattern of Carabayo was not found in a given candidate language we set its frequency of occurrence to one (assuming that it actually occurs in the candidate language, but very rarely such that it was missing in the sample). Consequently, we increased the frequencies of occurrence of those patterns of Carabayo found in the respective candidate language by one, to account for them being presumably more common than the patterns not found (column 'freq. known + 1' in Table 1). The reason for assigning those patterns of Carabayo that were not found in the candidate language a frequency of occurrence of one was that in order to be able to compare probabilities using likelihoods, the data sets used to derive the probabilities must be identical with regard to the response variable (here, the frequencies of

Table 1. Fictitious example of the determination of the probabilities by which patterns in an unknown language should occur, assuming that is actually a sample from a certain known language. The example shows the frequencies of occurrence of 15 patterns in the unknown and the known language. Note that the values in the column headed "Expected" sum to one. See text for details.

Pattern	Frequency unknown	Frequency known	Frequency known + 1	Expected	Used
p1	7	0	1	0.0125	0.0125
p2	0	5	6	0.0750	
p3	0	3	4	0.0500	
p4	6	8	9	0.1125	0.1125
p5	3	0	1	0.0125	0.0125
p6	0	9	10	0.1250	
p7	2	5	6	0.0750	0.0750
p8	4	0	1	0.0125	0.0125
p9	5	0	1	0.0125	0.0125
p10	3	4	5	0.0625	0.0625
p11	1	9	10	0.1250	0.1250
p12	3	5	6	0.0750	0.0750
p13	0	8	9	0.1125	
p14	5	0	1	0.0125	0.0125
p15	7	9	10	0.1250	0.1250

occurrence of the patterns in Carabayo) which means that the same set of frequencies of patterns in Carabayo must be used with each candidate language. Furthermore, in the subsequent steps probabilities of a given pattern to occur must be from the interval between 0 and 1, but excluding these two values. The derived frequencies of occurrence can be regarded as the expected frequencies of occurrence of the patterns of Carabayo given that it actually is a sample from the candidate language.

In a next step, we turned the expected frequencies of occurrences into probabilities. Specifically, we divided the expected frequencies of occurrences by the total number of patterns in the candidate language plus the number of different patterns in the candidate language plus the number of patterns found in the unknown but not in the candidate language (to account for the addition of 1 to each of the frequencies of the candidate patterns; column "expected" in Table 1). It is worth noting that probabilities of patterns found in Carabayo, but not in a given candidate language, were generally low (column "used" in Table 1).

Finally, we determined the likelihood of the actual frequencies of occurrences as their binomial probabilities given the expected probabilities of the patterns (derived from a given candidate language) and assuming a sample size being equal to the total number of patterns in Carabayo. Note that this approach basically means building one model for each candidate language with the frequencies of occurrence of the patterns in Carabayo as the response and their probabilities of occurrence in the respective candidate language as the predictor. The relative likelihood of a given candidate language then represents the relative strength of evidence in favour of the respective candidate language to be the actually best in the set of candidate languages considered.

The patterns we investigated were monograms, bigrams without word boundaries, and bigrams with word boundaries. For example, *punk* contains the bigrams *pu*, *un*, and *nk* (without word boundaries) and *_p*, *pu*, *un*, *nk*, and *k_* (with word boundaries).

For evaluating the overall performance of the approach and the effect of sample size (in terms of the number of considered patterns from Carabayo) on the accuracy with which the right candidate language would be detected we tested it with samples from spoken German as the unknown language and wordlist data of five languages as the candidate languages, one of them from German (see Section 2). We used samples of 5, 10, 20, 50, and 100 words (each replicated five times) from spoken German and tested each of them with 800 randomly selected patterns of the candidate languages.

In order to avoid undue influence of any particular random selection of 800 patterns out of the candidate language we repeated the random selection 1000 times per sample.

To evaluate how well the approach works with the three different ways of generating patterns (monograms, bigrams with and without word boundaries) and sample sizes of Carabayo, we used evidence ratios derived as the quotient of the likelihoods of two models. These can be interpreted as how much more likely one of the two (i.e. the one being in the numerator of the quotient) is the truly better model of the two. Since we wanted to test how well the approach identifies German, we used the evidence ratio in favour of German, that is, the likelihood derived for German divided by the largest of the likelihoods assigned to any of the other languages. If this is smaller than one it means that a language other than German was estimated to be the most likely candidate language, whereas a value of, for example, two means that German was twice as likely as any other of the candidate languages to be the language of origin of the sample. For Carabayo we proceeded correspondingly, with the exception that we derived evidence ratios for each of the considered languages of origin by dividing their likelihood by the largest likelihood revealed for any of the other four languages.

4. RESULTS

4.1. German Identification

With samples of German tested as the unknown language the approach performed well when the number of words of the unknown language was 20 or more (Figure 1). If 50 or more words were considered, German was invariably indicated as the most likely language of origin and also was usually associated with large evidence ratios. Furthermore, performance (as indicated by evidence ratios) was best when using bigrams with word boundaries and worst with monograms. However, there was remarkable variation between evidence ratios for the exact same sample of words (range of evidence ratios revealed for the exact same sample of the unknown language in Figure 1) indicating that using different samples of 800 patterns from the candidate languages can considerably influence its relative likelihood of being the sample of origin. Similarly, different samples of the unknown language could reveal quite different findings, even when they were of the same size (variation of evidence ratios revealed for different samples of the same size in Figure 1). This was particularly

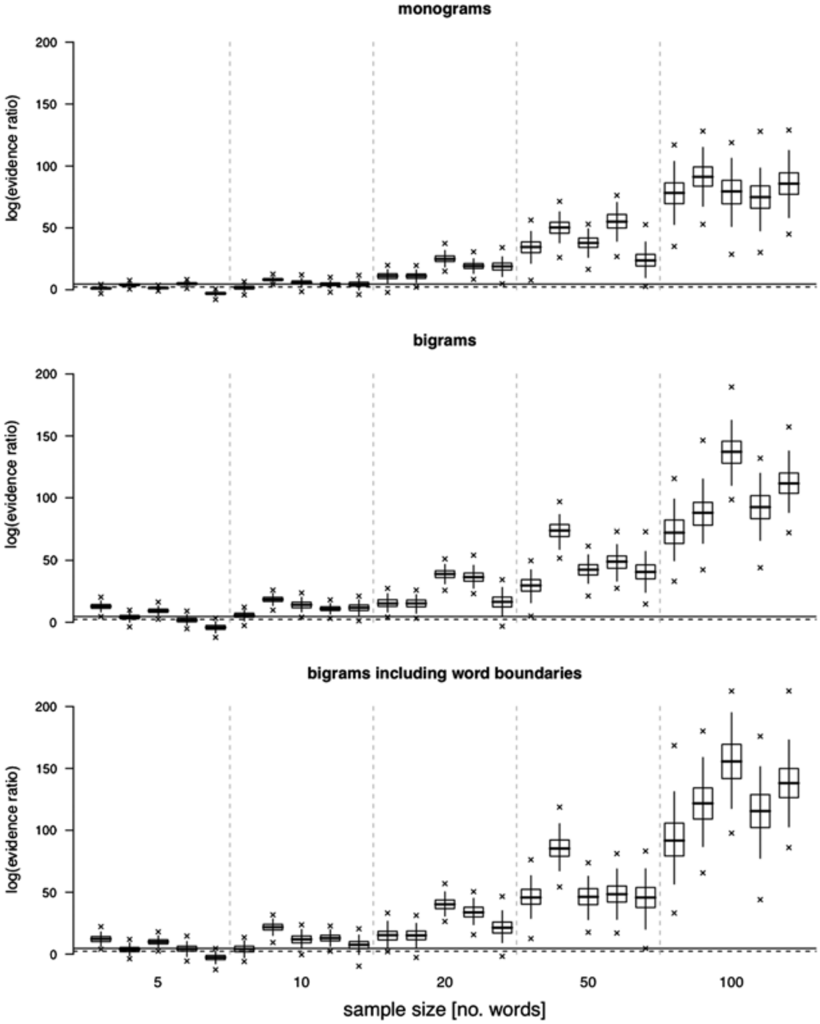

Fig. 1. Results for German. We used 5 different sample sizes (in terms of number of words; x-axis) and for each of them five independently drawn samples. Evidence ratios (log-transformed) in favour of German are shown on the y-axis. Indicated are medians (horizontal lines), quartiles (boxes), percentiles (2.5 and 97.5%; vertical lines), and maximum and minimum (laying crosses). Each box is based on 1000 random samples of 800 patterns from the five candidate languages. Evidence ratios between 10 (dashed vertical line) and 100 (solid vertical line) indicate ‘moderate evidence to support’ and values above the solid line indicate ‘moderately strong evidence to support’ in favour of German being the most likely origin of the sample. Note that whenever the sample contained at least 20 words German was invariably indicated to be the most likely language of origin.

evident when sample sizes were smaller (five or 10 words) in which case some samples revealed clear support for German to be the most likely language of origin (e.g. leftmost sample of five words using bigrams with word boundary) but others did not at all (e.g. rightmost sample of five words using bigrams with word boundary).

4.2. Carabayo Identification

When using bigrams with word boundaries none of the five candidate languages we considered revealed clearly more support as being identifiable with Carabayo than any of the others (Figure 2). In fact, two languages (Passé and Yurí) revealed similar support for being the language of origin, but even for those two languages around half of the samples of 800 patterns out of them revealed evidence ratios smaller than zero indicating that another language was indicated to be the most likely language of origin. Furthermore, when considering monograms rather than bigrams another language (Uainuma) was indicated as the most likely language of origin, but again, for about 50% of the samples some other language was indicated to be the most likely language of origin.

Results for Carabayo were based on bigrams with word boundaries (left), bigrams without word boundaries (middle) and monograms (right) (see also Figures 3–4). Evidence ratios for the individual language were derived by dividing the respective likelihood by the maximum of the likelihoods of the other languages. Indicated are medians (horizontal lines), quartiles (boxes), percentiles (2.5 and 97.5%; vertical lines), as well as maximum and minimum (laying crosses). Each box is based on 1000 random samples of 800 patterns from the candidate languages. Evidence ratios between 10 (dashed vertical line) and 100 (solid vertical line) indicate "moderate evidence" and values

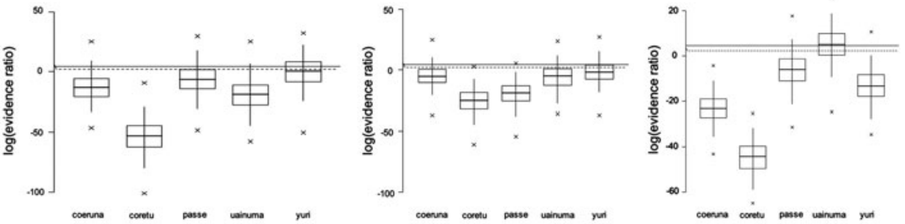

Fig. 2. Results for Carabayo based on bigrams with word boundaries (left). Results for Carabayo based on bigrams without word boundaries (middle). Results for Carabayo based on bigrams with monograms (right).

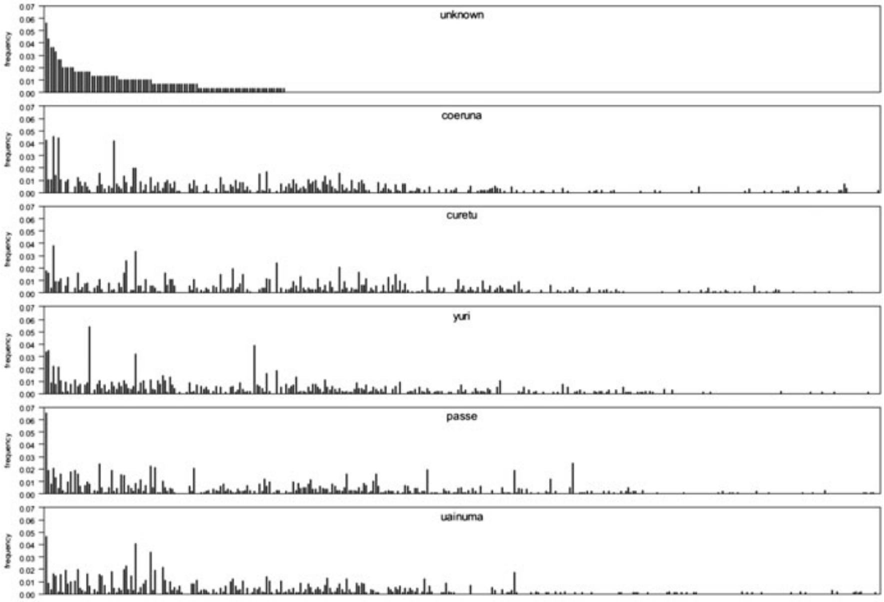

Fig. 3. Frequency of bigrams with word boundaries in Carabayo (top) and the five languages of potential origin considered. Bigrams are displayed in the same sequence for all languages and ordered by decreasing frequency of occurrence in Carabayo.

above the solid line indicate “moderately strong evidence” that the respective language is the most likely origin of the sample. Note that when using bigrams with word boundaries no language stuck out strongly from the others and that two languages (Passé and Yurí) received considerable support in roughly a quarter of the samples from the candidate language. When using bigrams without word boundaries it is even less clear that any language would stick out. When using monograms results differed considerably from those obtained when using bigrams, with or without word boundaries, in the sense that this time Uainuma somewhat stuck out as the most likely language of origin (although this was only the case in roughly 50% of the samples from Uainuma).

5. DISCUSSION

The application of *n*-gram frequency analyses for the identification of known languages, such as German, showed that, even if the information available

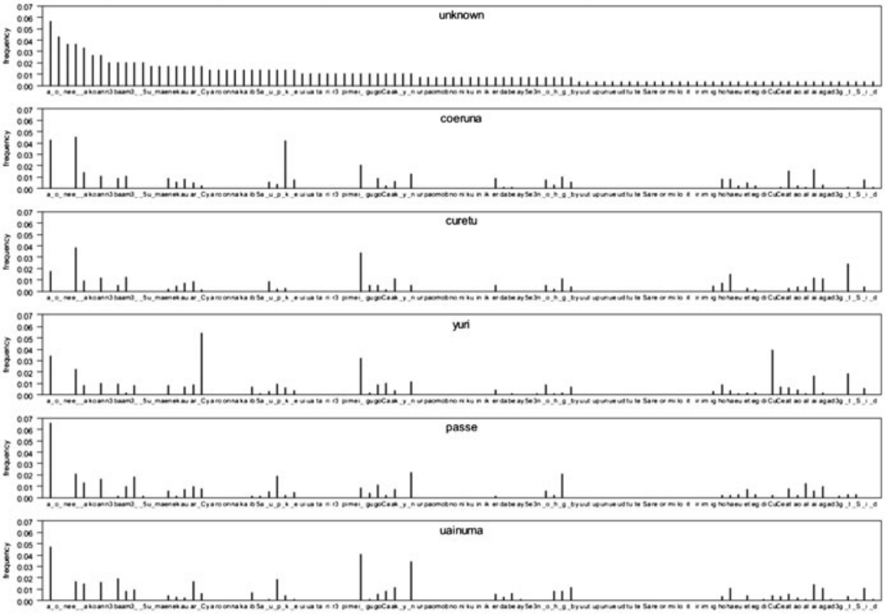

Fig. 4. Frequency of bigrams with word boundaries in Carabayo (top) and the five languages of potential origin considered. Bigrams are displayed in the same sequence for all languages and ordered by decreasing frequency of occurrence in Carabayo. Only bigrams that occur in Carabayo are depicted.

on candidate languages is limited to only 800 patterns, samples of only 20 words can usually be identified as being from one of these languages. However, our results also indicated the limitations of using only 800 patterns from a candidate language, since this lead to a large variation in evidence ratios revealed for the exact same sample of patterns from the unknown language, depending on the particular sample from the candidate language (Figure 1). Furthermore, it revealed uncertainties in identification even if samples of 100 words from an unknown language were used, as shown by remarkable variation between samples of the same size in Figure 1.

The application of the same analysis to identify the Carabayo language clearly failed, even though 50 words were available; that is, an amount that yielded good results in the German test case. This suggests that at least in part this failure is due to the nature of the data of the candidate languages here. Indeed, the huge variation between the different results indicates that using 800 patterns from these candidate languages might not give a very

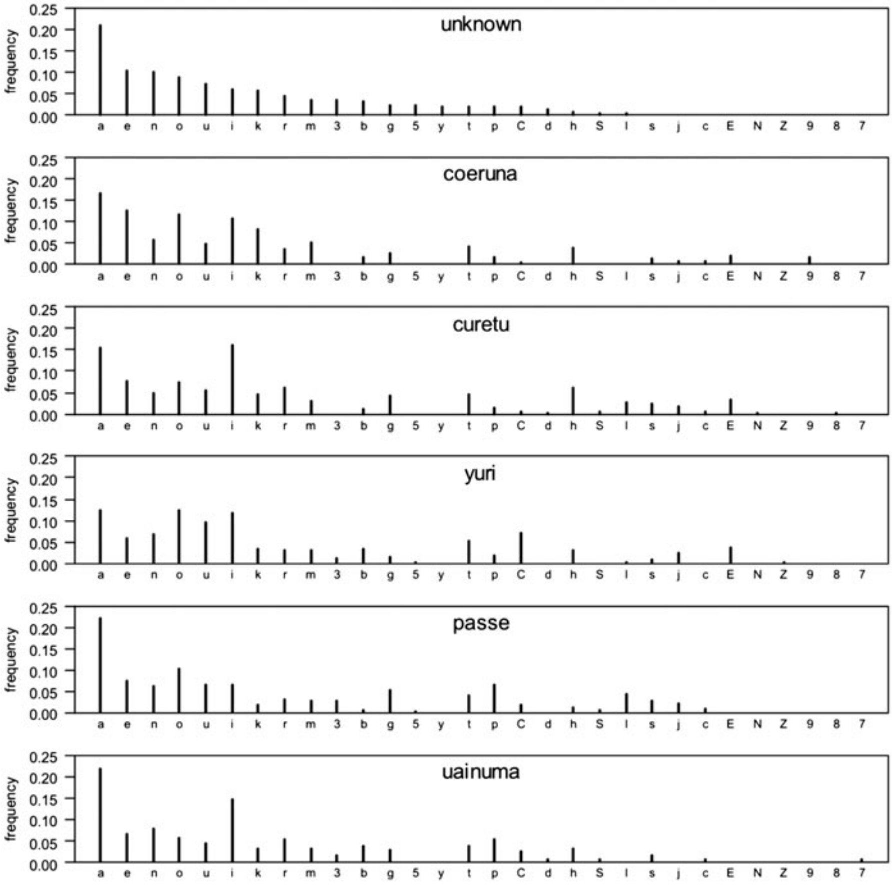

Fig. 5. Frequency of monograms in Carabayo (top) and the five candidate languages (see text). Monograms are displayed in the same sequence for all languages and ordered by decreasing frequency of occurrence in Carabayo.

accurate description of their properties with respect to the relative frequencies of bigrams. We suggest that this is due to the inconsistencies and inaccuracy of the available transcriptions of these candidate languages. These word lists were written down from the mouth of Amazonian Indians by German botanists in the early 19th century, at a time when there was no standardized transcription system or training in phonetics or phonetic transcription. The lack of differentiation between the Amazonian candidate languages revealed in our analysis thus probably reflects a failure to grasp and graphically represent the distinctive sounds of these languages by those who collected the wordlists. This is in stark contrast with the careful transcription by professionals in the case of German and the languages compared with it, which yields highly consistent representation of the sound system of these languages, probably comparable in consistency to orthographic representation (although this might be less true to the phonetics of a language). Inconsistencies in word boundary segmentation in the Amazonian data might additionally have made identification difficult, although taking into account word boundaries did not improve accuracy greatly, even in our German test case.

6. CONCLUSION

With the limited data available, an approach based on *n*-gram frequencies could not identify Carabayo as being more similar to one of the candidate languages considered, but it performed considerably better for identifying German, using a comparable set of German data and data from five candidate languages for German. We conclude that languages can be identified relatively well using *n*-gram frequencies even if the amount of data available is severely limited, both for the sample to be tested (20 words) and for the potential candidate languages (800 patterns), as long as the orthographic representation of these data are consistent.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the authors.

REFERENCES

Brown, C. H., Holman, E. W., Wichmann, S., & Velupillai, V. (2008). Automated classification of the world's languages: A description of the method and preliminary results. *STUF - Language Typology and Universals*, *61*, 285–308.

Franco, R. (2012). *Cariba malo: episodios de resistencia de un pueblo indígena aislado del Amazonas* [Cariba malo: Episodes of resistance of of an isolated indigenous tribe of the Amazon]. *Documentos históricos Imani 2*. Leticia: Universidad Nacional de Colombia. www.bdigital.unal.edu.co/6140/3/9789587611618.pdf

Gottron, T., & Lipka, N. (2010). A comparison of language identification approaches on short, Query-style texts. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. Rüger & K. van Rijsbergen (Eds.), *Advances in Information Retrieval* (pp. 611–614). Lecture Notes in Computer Science 5993. Berlin, Heidelberg: Springer. http://link.springer.com/chapter/10.1007/978-3-642-12275-0_59

Martius, C. F. P. von. (1867). *Beiträge zur Ethnographie und Sprachenkunde Amerika's, zumal Brasiliens. Vol. II: Zur Sprachenkunde. Wörtersammlung Brasilianischer Sprachen. Glossaria linguarum Brasiliensium. Glossarios de diversas lingoas e dialectos, que fallao os Indios no imperio de Brazil* [Contributions to the ethnography and linguistics of the Americas, especially of Brazil. Vol. II: On languages. Word lists of various languages and dialects, spoken by the indians in the Brazilian Empire]. Leipzig: Friedrich Fleischer.

Řehůřek, R., & Kolkus, M. (2009). Language identification on the web: Extending the dictionary method. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 357–368). Lecture Notes in Computer Science 5449. Berlin, Heidelberg: Springer. http://link.springer.com/chapter/10.1007/978-3-642-00382-0_29

Seifart, F., & Echeverri, J. A. (2014). Evidence for the identification of Carabayo, the language of an uncontacted people of the Colombian Amazon, as belonging to the Tikuna-Yurí linguistic family. *PLoS ONE* 9, e94814. doi: 10.1371/journal.pone.0094814