Toward a Khipu Transcription "Insistence": a Corpus-Based Study of the Textos Andinos

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37364622">https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37364622</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Toward a Khipu Transcription “Insistence”:
A Corpus-Based Study of the Textos Andinos

A thesis presented

by

Manuel Medrano

to

Applied Mathematics
in partial fulfillment of the honors requirements
for the degree of Bachelor of Arts
Harvard College
Cambridge, Massachusetts

March 29, 2019
Abstract

This thesis analyzes the *Textos Andinos*, a compilation of sixteenth-century Spanish transcriptions of indigenous narrations of khipus—knotted-string recording devices used in the Inka Empire for recording information. I compile the largest digitized and syntactically-annotated corpus of khipu transcriptions to date from the *Textos*. Textual interpretation is employed to suggest an exegetical typology of khipu transcriptions. I apply Ascher and Ascher’s (1997) concept of “insistence” to illuminate the idiosyncrasies of the texts. The output of the close reading—a primordial division of 72 khipu transcriptions—is subjected to exploratory multivariate analysis, based in corpus linguistics, to suggest a statistical typology of the corpus. Chronology and the recording of currency emerge as the most significantly distinguishable typological categories for describing khipu narration in the early colonial Andes. A significant differentiation is found in the essential narrative structures of pre- and postconquest khipu transcriptions. Novel statistical support is offered for Urton’s (1998) hypothesis: postconquest khipu narrations were characterized by attenuated clauses and enumerated lists, constituting a flattening of the expressive capacity of khipus following the Spanish conquest. I offer formal principles for a Khipu Transcription Corpus (KTC)—a novel online repository of early colonial khipu transcriptions. Following these principles, it is argued that aggregate analysis of the texts in a corpus framework establishes the enabling infrastructure for a statistically-informed khipu transcription insistence.
Acknowledgments

I am indebted to Professor Gary Urton both for his guidance during this project and for his supervision of my khipu research since 2016. I am humbled by the amount of support that he has provided during every stage of this project and during my time at Harvard. I would like to thank Professor Sabine Hyland and Dr. Jon Clindaniel, who provided valuable input regarding the project methodology. I thank my roommates for their constant on-campus encouragement. Finally, I would also like to thank my parents and my brother for their continual support and inspiration.
Contents

Abstract ..................................................................................................................................................... ii
Acknowledgments .................................................................................................................................... iii
List of Figures .......................................................................................................................................... vi
List of Tables .......................................................................................................................................... vii

1 Introduction and Motivation .................................................................................................................. 1
1.1 “By Means of Knots in Some Strings” ............................................................................................... 1
1.2 The Problem of Khipu “Decipherment” ............................................................................................. 5
1.3 The Problem of Narrative-Historical Khipus ................................................................................... 10
1.4 Historical Corpus Linguistics for Khipu Transcription Studies ....................................................... 14

2 An Exegetical Study of the Textos Andinos ........................................................................................ 19
2.1 Introduction to the Textos Andinos ................................................................................................... 19
2.2 An Exegetical Khipu Transcription Typology .................................................................................. 22
2.3 On the Importance of Typological Thinking ..................................................................................... 34

3 Methods: Formation of a Digital Khipu Transcription Corpus ........................................................ 37
3.1 Treating the Textos Andinos as a Corpus ......................................................................................... 37
3.2 Data Preparation and Transcription Cleaning ................................................................................ 43
3.3 Spelling Standardization and Modernization ................................................................................... 50
3.4 POS-Tagging and Lemmatization ..................................................................................................... 52
3.5 XML Conversion ............................................................................................................................... 56

4 A Multivariate Study of the Textos Andinos ...................................................................................... 59
4.1 Multivariate Motivation .................................................................................................................... 59
4.2 Linguistic Feature Choice ................................................................................................................ 61
4.3 Supplemental Khipu Transcription Variables .................................................................................. 63
4.4 Introduction to Principal Component Analysis ................................................................................ 69
4.5 Principal Component Analysis Implementation ............................................................................... 71
4.5.1 Lemma Implementation ............................................................................................................ 72
4.5.2 UPOS Implementation .............................................................................................................. 88
4.6 Hierarchical Agglomerative Clustering Implementation ............................................................... 102
4.7 “From Knots to Narratives”: Statistical Support for Urton (1998) ............................................... 113

5 Considerations and Challenges ........................................................................................................... 121
5.1 Challenges to Corpus Linguistics ................................................................................................... 121
5.2 Challenges to Aggregate Interpretation of Khipu Transcriptions .................................................. 124

6 Summary of Results and Future Directions ...................................................................................... 128
  6.1 Toward a Statistical Khipu Transcription Typology ................................................................. 128
  6.2 Implications for the Future: Outlining a Formal Khipu Transcription Corpus ..................... 132

7 Conclusion: Toward a Khipu Transcription “Insistence” ............................................................... 135
  7.1 Final Remarks ................................................................................................................................. 135

Bibliography ............................................................................................................................................. 137

Appendix A Supplementary Data and Figures .................................................................................... 151
  A.1 Principal Component Analysis ....................................................................................................... 151
    A.1.1 Lemma Implementation ........................................................................................................... 151
      A.1.1.1 Factor Loadings and Contributions .................................................................................. 151
      A.1.1.2 Parallel Analysis Plot ...................................................................................................... 159
    A.1.2 UPOS Implementation ........................................................................................................... 159
      A.1.2.1 Factor Loadings and Contributions .................................................................................. 159
      A.1.2.2 Parallel Analysis Plot ...................................................................................................... 166
  A.2 Hierarchical Agglomerative Clustering ......................................................................................... 167
    A.2.1 Baker’s Gamma for Dendrogram Comparison ....................................................................... 167
    A.2.2 Critical Rejection Value for k = 2 in Fowlkes-Mallow ......................................................... 168

Appendix B Transcription Metadata .................................................................................................... 170

Appendix C Supplementary Code ......................................................................................................... 181
  C.1 Chapter 3 Supplementary Code ..................................................................................................... 181
    C.1.1 Tokenization, Lemmatization and UPOS-tagging ................................................................. 181
  C.2 Chapter 4 Supplementary Code ..................................................................................................... 182
    C.2.1 XML Corpus Import ............................................................................................................... 182
    C.2.2 Principal Component Analysis ............................................................................................... 183
      C.2.2.1 Lemma Implementation .................................................................................................. 183
      C.2.2.2 UPOS Implementation .................................................................................................... 186
    C.2.3 Hierarchical Agglomerative Clustering .................................................................................. 189
      C.2.3.1 Lemma Implementation .................................................................................................. 189
      C.2.3.2 UPOS Implementation .................................................................................................... 190
    C.2.4 Hierarchical Clustering Comparison ...................................................................................... 191
  C.3 Metadata Appendix, Summary Statistics and Transcription Map Supplementary Code ............... 193
    C.3.1 Transcription Metadata Appendix .......................................................................................... 193
    C.3.2 Summary Statistics .................................................................................................................. 193
    C.3.3 Transcription Map .................................................................................................................. 194
## List of Figures

1.1 Illustrative Khipu Diagram ................................................................. 4
2.1 Map of Transcription Locations ........................................................ 20
2.2 Exegetical Khipu Transcription Typology ......................................... 33
3.1 Illustrative *Textos Andinos* Page Excerpt ..................................... 44
3.2 Excerpt of Annotated XML Transcription Corpus ............................. 58
4.1 Scree Plot (Lemma Implementation) .................................................. 74
4.2 Variable Factor Map for Dimensions 1 and 2 (Lemma Implementation) ........................................................................ 76
4.3 Variable Factor Map for Dimensions 3 and 4 (Lemma Implementation) ........................................................................ 78
4.4 Contributions of Individuals (Lemma Implementation) ...................... 80
4.5 Categorical Supplementary Variables with 99% Confidence Ellipses (Lemma Implementation) ........................................ 81
4.6 Biplot (Lemma Implementation) ........................................................ 84
4.7 Pre/Postconquest Transcription Segmentation (Lemma Implementation) ........................................................................ 87
4.8 Currency Transcription Segmentation (Lemma Implementation) ......... 87
4.9 Scree Plot (UPOS Implementation) .................................................... 90
4.10 Variable Factor Map for Dimensions 1 and 2 (UPOS Implementation) ........................................................................ 91
4.11 Contributions of Individuals (UPOS Implementation) ....................... 94
4.12 Categorical Supplementary Variables with 95% Confidence Ellipses (UPOS Implementation) ........................................ 95
4.13 Biplot (UPOS Implementation) ........................................................ 98
4.14 Pre/Postconquest Transcription Segmentation (UPOS Implementation) ........................................................................ 101
4.15 Currency Transcription Segmentation (UPOS Implementation) ........ 101
4.16 Statistical Khipu Transcription Typology ......................................... 102
4.17 Lemma Implementation of Pre/Postconquest Dendrogram .................. 105
4.18 UPOS Implementation of Pre/Postconquest Dendrogram ................... 106
4.19 Tanglegram ...................................................................................... 108
4.20 \((k, B_k)\) Plot ................................................................................... 112
4.21 UPOS Contributions Versus Expectation Under Uniform Contributions ................................................................. 115
6.1 Comparative Khipu Transcription Typologies ................................... 130
A.1 Variable Correlation Significance Plot (Lemma Implementation) ........ 155
A.2 Parallel Analysis Plot (Lemma Implementation) .................................. 159
A.3 Variable Correlation Significance Plot (UPOS Implementation) ........... 162
A.4 Parallel Analysis Plot (UPOS Implementation) .................................... 166
A.5 Baker’s Gamma Null Distribution Density under Permutation Test ........ 167
A.6 Baker’s Gamma Bootstrap Distribution ............................................ 168
List of Tables

3.1 Transcription Corpus Summary Statistics ................................................................. 40
4.1 First Five Eigenvalues and Explained Variance (Lemma PCA Implementation) .......... 73
4.2 First Five Eigenvalues and Explained Variance (UPOS PCA Implementation) ........... 89
A.1 Significant Factor Loadings (Lemma PCA Implementation) ....................................... 153
A.2 Active Variable Contributions (Lemma PCA Implementation) ..................................... 153
A.3 Variable Correlation Matrix (Lemma PCA Implementation) ....................................... 154
A.4 Individual Contributions (Lemma PCA Implementation) .......................................... 158
A.5 Significant Factor Loadings (UPOS Implementation) ............................................. 160
A.6 Active Variable Contributions (UPOS PCA Implementation) .................................... 161
A.7 Variable Correlation Matrix (UPOS PCA Implementation) ..................................... 161
A.8 Individual Contributions (UPOS PCA Implementation) ........................................... 165
Chapter 1

Introduction and Motivation

The mystery has been dispelled and we now know the quipu for just what it was in prehistoric times, and what it is, in its limited use today; simply an instrument for recording numbers.

—Charles Mead, Foreword to The Ancient Quipu or Peruvian Knot Record (1923)

1.1 “By Means of Knots in Some Strings”

In November 1533 AD, a Spanish conquistador named Hernando Pizarro led a military expedition from Cajamarca, a town in the Peruvian highlands, to the coastal religious center of Pachacamac. Our protagonist’s surname may sound familiar: Hernando’s brother Francisco, just one year earlier, had led the small militia that conquered the Inka Empire. This formidable nation state was the most powerful of precolombian South America, stretching over 4,000 kilometers along the spine of the Andes mountains, from modern Ecuador and Columbia down to central Chile. Hernando, who took copious notes during his journey, noted a curious local practice as his garrison pillaged a former Inka storehouse along Capac Ñan, the royal Inka road:

“all of the neighboring chieftains come to these storehouses along the road… the [storehouse keepers] maintain a deposit of lumber, maize and more, and they record, by means of knots in some strings, what each chieftain has brought.”

Lacking the words to describe what he saw, Pizarro offered the first written record of numerical recording in khipus—the knotted string devices used for state administration in the former Inka Empire. Yet, Pizarro’s periphrastic account remains a fitting representative of the Spanish colonial writings on khipus—mostly references in passing to strings and knots—which, by and large, suffered from confusion, contradiction and dismissal. These contradictory

---

1 Hernando Pizarro, “A los Señores Oydores de la Audiencia Real de Su Magestad,” in Informaciones sobre el antiguo Perú, Colección de Libros y Documentos Referentes a la Historia del Peru, vol. 3 (2nd series), ed. Horacio H. Urteaga (Lima: Sanmartí, 1920 [1533]), 175; my translation, emphasis added. Original: “A estos pueblos del camino vienen á servir todos los caçiques comarcanos… tienen depósitos de leña é mahiz é de todo lo demás, é cuentan por unos nudos en unas cuerdas de lo que cada caçique ha traydo.”

accounts weigh on the present: today, after over a century of academic study, the khipu script remains largely undeciphered.

Khipus (meaning “knot” in Quechua, the indigenous Andean language) were the primary information transmission and storage devices of the Inka Empire (1400-1532 AD). The Inkas, who lacked a formal system of graphical writing, relied on a class of cord keepers called khipukamayuqs to fill twisted cords with data pertaining to the administration of their great empire. Numerical khipus contain varied counts and measures, conveying snapshots of an empire in motion—the receipt of goods at military storehouses, enumeration of taxpayers, recording of calendrical events and counts of camellid flocks—devices which facilitated extended political and social control by a small ruling elite in Cuzco (the Inka capital) over a population often estimated at over 10 million. Indeed, the most senior khipukamayuqs likely crafted their knotted signs within a broadly conventionalized system of meaning. Murúa notes that khipukamayuqs developed their skills in Cuzco for two years of a four-year course: “in the third [year] they were shown, by certain khipus, important things for governance and authority; and in the fourth [year] they learned, in the cords as well, many histories.” Enriquéz tells us that selected Inka khipukamayuqs received continual training in interpreting cords from early childhood. Through the mid-sixteenth century, khipus, narrated by their khipukamayuqs, were integral to the formation of the colonial Spanish viceroyalty: the knotted accounts were referenced to (1) write a Spanish history of the Inkas, (2) obtain broad population statistics, (3) establish Spanish tax levels in line with Inkaic tribute and (4) apportion lands and people in the first encomiendas.

---


4 Quoted in Carlos Radicati di Primeglio, Introducción al Estudio de los Quipus (Lima: Biblioteca de la Sociedad Peruana de la Historia, 1951), 136-137. Original: “En el tercero se les enseñaba, en ciertos quipus, cosas importantes para el gobierno y autoridad y en el cuarto aprendían también en los cordeles muchas historias…”


This extensive use of khipus makes it more intriguing that the Spanish relied so heavily on a means of communication that they themselves never learned to read or interpret. The foreignness of the khipus’ three-dimensional script may be, in-part, to blame for this Spanish indifference. However, more insidious was the conflict between Spanish graphical writing and indigenous khipu recording. Spalding states strongly: “the Spanish mercenaries who invaded the Andes carved out their dominion as much with letters as with weapons.”

Curatola Petrocchi and Puente Luna state that the penetration of alphabetic writing into indigenous communities constituted a “process of ‘colonization’ of indigenous systems of information.” Indeed, the imposition of graphical writing, coupled with Catholic proselytization and the Spanish encomienda system, constituted a multi-front attack on Andean ways of life and social organization dating to long before the short-lived Inka Empire.

Before the next section, which further motivates the study of colonial khipu accounts, a brief overview of the physical structure of these knotted-string devices is offered for those who may be unfamiliar. Figure 1.1 is an illustrative diagram of a khipu, which is composed of cotton and/or camelid fibers. The spun and plied threads are either naturally colored or are dyed using organic pigments. The basic structure of a khipu consists of a thick, horizontal “primary cord,” from which hang some number (between one and over a thousand) of thinner “pendant cords.” Vertical “top cords” may emanate from the primary cord, which can recount the sum total of the knotted pendants through which they pass. Khipu pendant cords often have knots tied into them, which compose a hierarchical, base-10 system of enumeration (at least in most cases; see below). In this way, knots tied at three different horizontal levels on a khipu cord signify three orders of magnitude: ones, tens and hundreds. Pendant cords may themselves also host attached strings, called “subsidiary cords,” which may have their own, nested subsidiaries.

---

7 Spalding, “Quipu versus escritura,” 65; my translation.
8 Marco Curatola Petrocchi and José Carlos de la Puente Luna, “Contar concertando: quipus, piedritas y escritura en los Andes coloniales,” in El quipu colonial: estudios y materiales, eds. Marco Curatola Petrocchi and José Carlos de la Puente Luna (Lima: Fondo Editorial, 2013), 194; my translation.
The schematic offers an idealized view of khipus, which often rest faded, broken or tangled beyond recognition in museum archives around the world. For this reason, an active effort continues to excavate and catalog the remaining archaeological khipus, led by the Khipu Database Project (KDB) in the Department of Anthropology at Harvard University. In full, khipus are hierarchically-organized, branching structures with multiple nodes for recording meaning. These include binary loci (e.g. the direction of the attachment knot with which pendant cords are attached to the horizontal primary cord) categorical variables (e.g. cord color) and numerical features (the values recorded in the knots themselves). The Inkas utilized these different loci of meaning to spin countless pendant cords, each with an expansive information capacity aided by spatial and tactile relationships with its surrounding cords. Exactly how these structural elements were related to each other to produce narrative speech is a field of continuing study.

---

This background information is offered to contextualize the forthcoming discussion of colonial writing about Inka writing. Isolating the expressive capacity of this technology in practice is one of many aims of this study, especially in comparing the precious few transcribed narrations of pre- and postconquest khipu cords. I probe whether a discernible shift in khipu grammar emerges in the Spanish chronicles, once khipus were summoned only for bureaucratic recitations to Spanish conquistadors and no longer transmitted information horizontally between Inka khipukamayuqs.\footnote{Topic, “De ‘audiencias’ a archivos,” 48, 59.} Despite the focus on colonial authors, sources and this question, I aim to maintain the khipus, and the text of their narrations, as the central protagonists of this essay. After a century of academic study thus far, it is hoped that we might someday offer a response to Spalding’s pessimistic prescription: “With the disappearance of the khipu technique at the level of the Andean state, we lost the opportunity to understand one of the most amazing and original solutions which human societies have designed for the problem of social organization at a grand scale.”\footnote{Spalding, “Quipu versus escritura,” 74; my translation.}

1.2 The Problem of Khipu “Decipherment”

Why study a system of brittle cords, gathering dust in private collections and display cases in over 88 museum institutions spread throughout the world?\footnote{Gary Urton, \textit{Inka History in Knots: Reading Khipus as Primary Sources} (Austin: University of Texas Press, 2017a), 261-264. This number grows continually, following archaeological excavation, discovery of new khipus in museum archives and donations of previously unpublished khipus to museum institutions.} The question invites a broad set of motivations, with applications in Inka and colonial-period contexts. The early colonial period (c. 1532-1600 AD) offers a glimpse into khipus as a means of navigating the Spanish ruling order: during this time, cotton and animal fiber cords assumed newfound roles including, but not limited to, rosaries, revolt-planning epistles and mass attendance logs.\footnote{Assadourian, “String Registries,” 137; Charles 2007; Hyland 2017.} The ubiquity of khipus in pre- and postconquest Andean life promises broad interdisciplinary intrigue, inviting creative approaches to interpreting what is only a small subset of the hundreds of thousands of khipus in use at the height of the Inka Empire.\footnote{Andrés Chirinos Rivera, \textit{Quipus del Tahuantinsuyo: Curacas, Incas y su Saber Matemático en el Siglo XVI} (Lima: Comentarios, 2010), 92.}

Additional motivation is drawn from the approximate third of the 1,045 surviving khipus with non-numerical, “narrative” properties—many believed to record royal dynasties, religious accounts and biographical information. Even today, we have yet to confirm the name of a single
indigenous individual in preconquest South America from a primary source.\textsuperscript{16} This fact is especially troubling, given evidence that khipus encoded the names of provincial residents of the Inka Empire.\textsuperscript{17} Historical events on the continent also remain opaque before 1532; surviving oral histories represent a subset of the subset of historical events that were logged on khipus. De la Vega (1966 [1609]) describes several categories of information:

> These men recorded on their knots all the tribute brought annually to the Inka… They recorded the number of men who went to the wars, how many died in them, and how many were born and died every year… everything that could be counted, even mentioning battles and fights, all the embassies that had come to visit the Inka, and all the speeches and arguments the king had uttered.\textsuperscript{18}

Given the purported flexibility of khipu recording, one can only imagine the breadth of histories resting silently in string across the world.

Even among the two-thirds of surviving khipus labeled numerical khipus, there remains a question of interpretation: a seven-turn long knot can confidently be stated as recording seven, but seven what? Seven fanegas (Spanish bushels) of wheat? Or, perhaps seven taxpaying tributaries—a single household within a regional tax administration? Robert and Marcia Ascher cast this problem as a binary classification of numbers as either magnitudes or labels.\textsuperscript{19} This distinction is captured by comparing a Harvard undergraduate’s height and postal code. To say that the author is 170cm tall offers a magnitude. However, the nonnumerical label “90012” is no more than an arrow, pointing toward Los Angeles, California—my own hometown. Magnitudes retain their meaning when summed, unlike labels: for example, the approximate weights of passengers and luggage on airplanes are often combined to determine optimal flight balance. While nonnumerical labels are not meaningfully summed, these are amicably grouped. Take the following set, hypothetically recorded on an illustrative khipu cord: \{13, 13, 20977973, 90012, (626) 755-9991, 03/29/2019\}. Upon first appearance, the collection of items lacks intrinsic meaning. However, in combination, a narrative emerges: a Harvard student with ID number


\textsuperscript{17} Martti Pärssinen and Jukka Kiviharju, \textit{Textos andinos: Corpus de textos khipu incaicos y coloniales}, vol. 1 (Madrid: Instituto Iberoamericano de Finlandia and Universidad Complutense de Madrid, 2004), 50.


20977973 and initials MM (the 13th letter of the alphabet) hails from Los Angeles, California, can be reached by cell phone at (626) 755-9991 and submitted an undergraduate thesis on March 29, 2019. In their combination, these “structural indicators” form “units of meaning constructed with numbers.”

It is apparent that such collections of labels, encoded in the twists, colors and knots of khipus, fueled a “mnemotechnic” system: particular arrangements of khipu elements served as “point[s] of departure for local memory.” These observations invite us to dissect the confident epigraphic statement of Charles Mead, in his foreword to Locke’s (1923) seminal work on numerical khipu recording: “exaggerated notions of the possibilities of quipu… have been generally obtained.”

A broader goal of this writing is to counter Mead’s conservatism, using mathematical techniques—applied to colonial-era khipu transcriptions—to highlight the communicative capacity of khipu recording and to confront the more provocative question: “What category of number [are we] looking at?”

Study of khipus invites broad sensory engagement. Indeed, there is a curious fascination in western educational models with “growing out” of early techniques. Counting serves as a telling case study. Young students learn numbers tactually—arranging blocks or figurines into piles of 5s, 10s, etc. Early speech development accompanies counting exercises, as the student pairs piles with verbal announcements (“three blocks,” “four action figures,” etc.). Upon graduation from these early methods, students learn abstract mathematical symbols, replacing the pile of seven blocks with the cipher “7”—a symbol without intrinsic tactile meaning—in stark contrast to the tactile intuition of a khipu long-knot, wrapped around itself seven times. Equally curious is the belief that progress implies minimizing the number of senses used at any given time. Modern graphical writing—a close cousin to counting—echoes this trend. Early spatial learning occurs in three dimensions: students learn to build towers of blocks before building sentences. Early grade school brings another graduation: students are taught reading and writing, as information intake itself becomes mute. Graphical writing flattens sensory experience—quite

---


22 Charles Mead, foreword to The Ancient Quipu or Peruvian Knot Record (New York: American Museum of Natural History, 1923).


literally—on to sheets of paper and computer keyboards. Consider the oddity of reading and writing in silence: modern libraries hold rows of solitary students, huddled quietly in adjacent cubicles, interrupted by the staccato of keyboard strokes.

Khipu narration collapsed language, writing and mathematics into an immersive sensory experience, removing the “do not touch” signs so eminently common in modern museum contexts.\textsuperscript{25} Indeed, the act of khipu recitation was in its essence performative—a skill conveying seemingly superhuman abilities to bureaucratic khipukamayuqs and cementing ideological dominance through “political arithmetic”\textsuperscript{26} and “statistical control.”\textsuperscript{27} In contrast to modern information intake, khipu narration maintained touch, sound and vision as unified elements. The rarity of khipus mentioned in writing about writing further motivates the present study.\textsuperscript{28}

Indigenous narrative ability inspired awe in early colonial authors: de Gamboa noted that “it is a thing of wonder to see the minute details preserved in [khipu] cords, of which there are experts as among us there are authors.”\textsuperscript{29} Khipu recitation emerges as a “writing event”—an organic exchange, often filled with negotiation over tax assignments and community obligations—that merged countability with the internal organization of local communities.\textsuperscript{30} Preconquest Andean government administration required mathematical literacy for those in power, coupled with performance to maintain the trust of provincial subjects. At the moment of conquest, the mathematical and narrative ability of Inka administrators—aided by the calculative yupana (counting board)—rivaled the quantitative acumen of the conquering Spanish.\textsuperscript{31}

A notable trend among Andeanists is to use khipus—including colonial-era transcriptions of khipus—as motivation to declare a being or essence of the Inka Empire. The Aschers (1997) prefer “insistence”: a concept isolating the idiosyncrasies and defining cultural coherencies of

\begin{itemize}
  \item Quilter, preface to \textit{Narrative Threads}, xv.
  \item Ascher and Ascher, \textit{Mathematics of the Incas}, 165.
  \item Assadourian, “String Registries,” 135.
  \item Pedro Sarmiento de Gamboa, \textit{Historia de los Incas} (Madrid: Ediciones Atlas, 1960 [1572]): 212; my translation, emphasis added. Original: “Es cosa de admiración ver las menudencias que conservan en estos cordelejos de los cuales hay maestros como entre nosotros del escribir.”
  \item Curatola Petrocchi and Puente Luna, “Contar concertando,” 197. See also Salomon 2002 for applications of “writing events” to internal community organization in a modern ethnohistorical context.
  \item Rivera, \textit{Quipus del Tahuantinsuyo}, 175.
\end{itemize}
different civilizations. Chase isolates insistence as a “meaningful repetition of important themes or principles.” These attempts appear to be widespread. Following a description of quantitative khipu elements, the Aschers declare summation and cross-categorization to be key elements of “Inka insistence.” Chirinos Rivera states confidently that the “ability to count and calculate was intimately associated with the being of the Inkas,” confirming that “just and equitable partitions [of goods with khipus]… appears a substantial part of Inka essence.”

Curatola Petrocchi and de la Puente Luna describe postconquest uses of khipus and yupanas as “a shared procedure of ongoing countability evident in the Andes until the moment of Spanish arrival.” Commentaries on khipus often expand to broader statements about cloth or weaving: Conklin adds that “cotton had vast ancient… associations” and that “the great symbolic importance of fabric… [speaks] broadly of the ritual importance of cords for the Inka.” These commentaries are unified by worthy aims: to contrast Andean understandings of literacy and administration to those of the conquering Spanish, capturing the visceral emotion of Guaman Poma’s warning that for many indigenous Andean peoples, “to write [graphically] is to cry.”

Assigning expansively interpretive essences to the Inkas—especially those based primarily in postconquest written sources—risks muddling the chronology of the colonial moment. Early Spanish writings about the Inkas are perhaps best-described as “history transpiring within a house of mirrors,” whereby the mutual unintelligibility of indigenous and European cultural practices planted contradictions and confusions into the historical record. Interpretive inconsistencies remain remarkably consistent in the sixteenth-century chronicles. An explicit recognition of the genealogy of khipu commentaries is needed to maintain situational context—especially when attempting to infer operational characteristics of the Inka Empire from

35 Rivera, Quipus del Tahuantinsuyo, 179; my translation, emphasis in original.
36 Ibid., 139; my translation, emphasis added.
37 Curatola Petrocchi and de la Puente Luna, “Contar concertando,” 193; my translation, emphasis added.
39 Quoted in Harrison, “Pérez Bocanegra’s Ritual formulario,” 287.
postconquest cords. Adding to the interpretive challenges of khipu studies are fundamental differences between European and Andean “modes of thought and representation”—often such that the definition of writing itself is left unquestioned. Brokaw states concisely: “sufficient ambiguity, contradiction, and general ignorance exists in the colonial chronicles [about khipus] to reject any simple interpretation of statements made in such documents.” In more extreme cases, khipus inspire romantic reflection, serving as a “springboard for the Western imagination” entirely divorced from indigenous context. One only need consider Walpole’s eighteenth-century musings, describing a purported khipu color poetry as “[holding] a dialogue with a rainbow… [verses] that hang like ropes of onions… though at first I tangled the poem and spoiled the rhymes.”

Nevertheless, these warnings invite exploratory questions: With so much effort seemingly placed on assigning an overarching essence or being to the Inkas by way of khipus, why not turn the direction of interpolation back on the conquering Spanish? By identifying broad patterns in colonial khipu transcriptions, might we assign an essence—strongly, I claim, an insistence—to sixteenth-century khipu transcriptions? What idiosyncrasies and coherencies in Castilian diction and vocabulary emerge in khipu transcriptions that might yield insights into the khipu grammar practiced in the colonial moment? Two volumes of compiled colonial khipu narrations, soon introduced, will help to probe the translative moment of khipu transcription.

### 1.3 The Problem of Narrative-Historical Khipus

Before introducing the data set, we revisit the de la Vega quotation that begins this chapter to probe the issue of “narrative” or “historical” khipus—the third of existing khipus which are believed to encode non-numerical, linguistic data:

… the purpose of the embassies or the contents of the speeches, or any other descriptive matter could not be recorded on the knots, consisting as it did of continuous spoken or written prose, which cannot be expressed by means of

---


43 Ibid., 112.


knots, since these can give only numbers and not words. To supply this want they
used signs that indicated historical events or facts or the existence of any
embassy, speech or discussion…  

The Inka Garcilaso’s statement centers the problem of khipu narration. If the content of speeches
could not be read from khipus, what type of khipu sign might differentiate a speech from a war?
More broadly, was narrative encoding confined to numbers as labels, or might a combination of
khipu elements (knot direction, color, spin/ply, etc.) have formed a synergistic system of
meaning that inspired narrative recitation?

Mid- and late-colonial khipus offer explicit associations between khipu elements and
language. Analyzing six seventeenth-century khipus from the Santa Valley, Medrano and Urton
found that attachment knot direction encoded the moiety affiliations of 132 indigenous
tributaries, whilst pendant cord color segmented villagers by first name. The study built off of
work by Hyland (2014) in the central Andean highlands, who associated knot direction (Z/S)
with village moiety groupings on a nineteenth-century hybrid khipu-alphabetic board from
Mangas. Following a 2015 discovery of two multicolored khipus from San Juan de Collata,
Huarochirí province, Hyland (2017) labeled the artifacts “the first khipus ever reliably identified
as narrative epistles by the descendants of their creators” and evidence of a shared, logosyllabic
writing system.

Studies of mid- and late-colonial khipus require acknowledgment of profound cultural,
and in turn contextual, changes in systems of meaning endured differentially throughout the
former Empire—even above, the three cited examples span three centuries. The Mangas board
lacks a primary cord; the Collata khipus lack pendant cord knots; even the Santa Valley khipus,
which most closely resemble Inka-period khipus, encoded tribute values—units of monetary

48 Manuel Medrano and Gary Urton, “Toward the Decipherment of a Set of Mid-Colonial Khipus from the Santa
Valley, Coastal Peru,” *Ethnohistory* 65, no. 1 (2018a): 18-19. It is worth noting that color segmentation remains at
the level of groupings. That is, while the number of distinct colors/color combinations closely aligns with the
number of distinct tributaries, efforts continue to map specific colors to specific sets of names. It is hypothesized
that monetary contribution formed a secondary category to group the taxpayers, explaining the incongruity of
individual color-name counts. The Santa Valley khipus are detailed further in Urton 2017a.
49 Sabine Hyland, Gene A. Ware and Madison Clark, “Knot Direction in a Khipu/Alphabetic Text from the Central
Andes,” *Latin American Antiquity* 25, no. 2 (2014). The board remains the best extant example of “Inka-style” khipu
cords in direct association with graphical writing (189).
50 Sabine Hyland, “Writing with Twisted Cords: The Inscriptive Capacity of Andean Khipus,” *Current
Anthropology* 58, no. 3 (2017): 416.
measure unknown to pre columbian South America. Nevertheless, the progression of examples highlights that while khipus remained the primary “administrative scaffolding” of the colonial republic, their forms changed organically, amassing regional variation over time.\footnote{Rivera, \textit{Quipus del Tahuantinsuyo}, 56.}

In fact, the early chroniclers, and extant artifacts, contrast several of the khipu sign associations proposed above. Take, for example, a hypothetical mapping between cord color and logosyllabic narration. However, there exist khipus which are numerical copies of each other—duplicate accounts—which do not share the same colors, or even relative ordering of colors.\footnote{Ibid., 233. See also Urton 2005 for further examples of matching khipus.} Further, de la Vega and de Gamboa describe khipu recitation as the outcome of lifelong learning and practice:

\ldots the quipucamayus\ldots were bound by their office to \textit{study constantly} the signs and cyphers on the knots so as to preserve in their memories traditions about famous events of the past. Thus, like historians, they were supposed to narrate such matters when requested\ldots\footnote{de la Vega, \textit{Royal Commentaries}, 332; emphasis added.}

\ldots from fathers to sons, [khipukamayuqs] communicated ancient events until the present, \textit{repeating them many times}, as teachings are received from a department, making the listeners say these historical accounts repeatedly, \textit{until they remained fixed in memory}. And in this way each one of their descendants communicated their annals\ldots\footnote{Sarmiento de Gamboa, \textit{Historia de los Incas}, 212; my translation, emphasis added. Original: “… [desde] padres á hijos, se iban refiriendo las cosas antiguas pasadas hasta sus tiempos, repitiéndolas muchas veces como quien lee lección en cátedra, haciéndoles repetir las tales lecciones historiales á las oyentes, hasta que se les quedase en la memoria fijas, y así cada uno á sus descendientes iba comunicando sus anales…”}

Put simply: if Inka-period historical khipus constituted a shared system of self-contained meaning, would constant practice have been necessary for successful recitation? While the Inka Garcilaso likely only learned to read accounting-style numerical khipus, his observation implies that interpretation of historical khipus required a “technical memory”—a system of interpretation constantly refined by the khipu’s creator.\footnote{Assadourian, “String Registries,” 123.} However, de Zárate’s account appears to be directly responsive to de la Vega’s statement:

\ldots in each province\ldots there can be found public houses full of [khipus], which are easily interpreted by the person in charge of them, even though they are many ages older than he is.”\footnote{Quoted in Ibid., 120.}
The khipus referenced in de Zárate’s account likely encoded numerical accounting measures—a category of khipu which exhibited notable standardization during the Inka Empire. Nevertheless, the question arises: Do inconsistencies found in Spanish accounts reflect Spanish misunderstandings of indigenous recording systems, or instead significant regional variation between khipus?

Important to acknowledge is that khipu construction—especially by the mid-sixteenth century—saw widespread adoption, at levels of complexity appropriate to the particular context. Khipu literacy, at the level of complexity of the colonial transcriptions analyzed herein, was not widespread outside of Inka bureaucratic khipukamayuqs. However, khipu proficiency, if thought of as an ability to interpret and manipulate khipus, leads Chirinos Rivera to hypothesize that by the mid-sixteenth century, hundreds of thousands of khipu operators, and over 10,000 expert khipukamayuqs, remained in the former territories of the Empire. Further, farmers throughout the former Empire are thought to have maintained simple numerical khipus for recording camelids—consistent with a more broadly defined khipu literacy.

The information capacity of khipus similarly merits mention. Conklin hypothesized, following correspondence with mathematician Adam Price, that each Inka-style pendant cord can assume eight million potential states (by varying knot direction, color, spin, ply, etc.)—comparable to the number of bits in a single word of ASCII alphanumerical computer code. However, Conklin’s interpretation, which multiplied khipu loci as binary on/off states, is a significant underestimation of the data-storage capacity of pendant cords. Inka-period historical khipus likely did not encode complete narrative script. Instead, these khipus likely recorded “anchor words,” iconographic signs that inspired broader elaboration from the khipukamayuq, who employed a grounded memory. Assadourian captures this “broader elaboration” in his

---

58 Quilter, preface to Narrative Threads, xvii.
60 Chirinos Rivera, Quipus del Tahuantinsuyo, 92.
distinction between signs and extrapolated reasonings.\textsuperscript{63} As such, a khipu containing a sequence of “anchor words” would require a mere fraction of the data required by ASCII for recording narrative language. Conklin’s conservative estimate underplays the role of branching in narrative accounts. Rather than a series of independent on-off “switches” (e.g. subsidiary cord yes/no), each iconographic element facilitated a sequential path of subsequent construction choices—vastly multiplying the information-storage capacity of the cords. In other words, attachment knot direction, color, etc. were likely dependent choices, defining a relational system of meaning with expansive capacity.\textsuperscript{64} As a result, it was hypothesized that a historical khipu could encode several-sentence epistles.\textsuperscript{65}

The above hypotheses provide an informed context with which to approach khipu recording, and its product—transcribed khipu narration—employing the capabilities of statistical aggregation.

1.4 Historical Corpus Linguistics for Khipu Transcription Studies

This thesis proposes a quantitative approach to historical texts, motivated by the realization that the realities we wish to study are “irrevocably lost” to the historical record.\textsuperscript{66} Indeed, just as the paleontologist comes to terms with the miniscule fraction of life which has fossilized, so too does the historical researcher—the linguist in particular—confront the sparsity of the historical record:

…[historical linguists] have no control over their data. Texts are produced by a series of historical accidents; amateurs may complain about this predicament, but the sophisticated historian is grateful that anything has survived at all. The great art of the historical linguist is to make the best of this \textit{bad data}—“bad” in the sense that it may be fragmentary, corrupted, or many times removed from the actual productions of native speakers.\textsuperscript{67}

An earnest review of the Spanish chronicles about khipus confirms many of these gaps. Many colonial references to khipus are products of historical accidents: Hernando Pizarro’s description of items logged “by means of knots in some strings” exists because of his decision to pillage an

\textsuperscript{63} Assadourian, “String Registries,” 128, 132, 144n60 introduces the principle of anchor words, including a technical explanation of “inspired elicitation” in the context of colonial-era khipu transcriptions.

\textsuperscript{64} Ascher and Ascher, \textit{Mathematics of the Incas}, 111.

\textsuperscript{65} Conklin, “A Khipu Information String Theory,” 83.


Inka storehouse. More extensive descriptions of khipus in legal proceedings reflect a pragmatism required by the situation, where khipus were more enablers of oral testimony than valued cultural artifacts.\textsuperscript{68} The colonial writings are fragmentary: even a cursory reading of Iberian trial testimonies demonstrates the extent to which the order of Spanish questions directed—even overpowered—the synthesis of khipu grammar.\textsuperscript{69} Those accounts that do survive often contradict each other and themselves. The Spanish chronicles are corrupted: in aggregate, the more thorough written descriptions of khipus were penned by conquering military generals and Catholic evangelizers. In practice, many encountering khipus either had a direct interest in discrediting their veracity (e.g. the plaintiffs in trial proceedings who otherwise lacked contradicting evidence), diminishing their complexity (e.g. encomenderos establishing a sum total of tribute owed) or lacked a basic understanding of their function.\textsuperscript{70} That references to khipus in the Spanish chronicles—including khipu transcriptions—are removed from the original situations of their narration follows directly from these concerns.

Quantitative historical linguistics offers an increasingly popular inroads to making the most of “bad data.” Principally among these methods is the use of natural language corpora—precisely-compiled, machine-readable collections of textual data—for the aggregate analysis of historical texts.\textsuperscript{71} The guidelines by which corpora are compiled are discussed later. However, a primary motivating argument for statistical study of natural language—and thus khipu transcriptions—is that oration is “inherently multivariate,” and that the choice of particular words in historical texts is attributable to one (or more) of any number of possible variables.\textsuperscript{72} Quantitative historical linguistics offers a vocabulary to determine which of these possible explanatory variables are particularly probable. In practice, the central assumption of the discipline is that formal differences in the frequency of particular speech elements reflect underlying functional differences.\textsuperscript{73} It is for this reason that historical corpus studies—including this one—are concerned with the relative frequencies of constructions in historical speech, by

\textsuperscript{68} Brokaw, “La recepción del quipu,” 122.
\textsuperscript{69} Monica Medelius, “El descargo en las cuentas de quipucamayos en un pleito entre encomenderos,” in El quipu colonial: estudios y materiales, eds. Marco Curatola Petrocchi and José Carlos de la Puente Luna (Lima: Fondo Editorial, 2013), 247n7.
\textsuperscript{70} Brokaw, “La recepción del quipu,” 139.
\textsuperscript{71} Jenset and McGillivray, Quantitative Historical Linguistics, 37, 41, 51.
\textsuperscript{72} Ibid., 49-52.
which it is argued that we can detect (and ideally reconstruct) distinct pragmatic regularities in
the infrastructure of khipu narration. Put more strongly, such a research aim is inseparable from
statistics because its unit of analysis is frequency—of words, parts of speech or different
grammatical constructions.

Multivariate literary computing is also distinguished from an example-based approach,
which characterizes the state of the art in study of historical khipu texts. Studies employing an
example-based approach are essentially grounded in exegesis—analysis of texts based in close
reading, interpretation and comparison of linguistic constructions within and/or between
historical documents. The approach is a powerful one for the qualitative study of text: exegetical
techniques can show whether a chosen linguistic form is present in historical documents. Two
lacunae emerge. The first is the question of replicability: studies based in close reading choose
specific examples or textual excerpts to support their arguments, which may be different than the
texts chosen by another researcher reading the same text. The other researcher may arrive at
different conclusions than the first, due to using different selection criteria, or even from
interpreting the same set of examples as the first in a different way.74 The second issue raised
involves questions of scale: consider, for example, the question not of whether a particular
phrase or construction is reflected in texts, but to what extent its (non-)use exceeds (falls short
of) the overall average in a particular genre or time period.75 These questions require the use of
larger data sets that adhere to formal selection criteria. García García’s observation further
motivates the usage of quantitative tools in historical linguistics: “An exhaustive analysis of any
linguistic issue in a corpus language should be based, ideally, on a study of all the available texts
in that language or that period of language.”76

This analytical ideal makes khipu transcriptions notable as a standalone body of study,
especially against the backdrop of the century-long quest to decipher khipus and their structuring
grammar. These documents—often containing line-by-line recordings of cord-by-cord khipu
recitations, are deservedly differentiable from, and offer several benefits compared to the study
of archaeological khipus. To begin with, khipu transcriptions often contain their own

75 Ibid., 8-9.
76 Luisa García García, “A Case Study in Historical Linguistic Research,” in *Perspectives on the Genitive in
English: Synchronic, Diachronic, Contrastive and Research*, eds. Anthony Bruton, Luisa García García and Joaquín
José Fernández Domínguez, Linguística 13 (Seville: Universidad de Sevilla, 2000), 121.
provenance: the year and location in which a reading took place, the name of the khipukamayuqs who were summoned to narrate their cords and the Spanish witnesses who attested to listening. These documents offer a wealth of metadata—information that was often robbed from archaeological khipus when removed illegally from grave contexts in the nineteenth and early twentieth centuries. It is likely that many archaeological khipus record their own metadata. However, the details offered in transcriptions represent the most immediately complete accounts we have of the content of specific khipus, and the conditions of their narrations. A second strength of khipu transcriptions is their recording of relative chronology. Namely, the documents provide a deterministic account of whether the narrated cords were knotted before or after the Spanish conquest. In the cases of postconquest khipus and trial proceedings, this window can often be narrowed even further to an interval of two to three years. And, even when an exact date of cord creation is omitted, high-level designations are still possible: a khipu describing payments in pesos (the Spanish currency) was certainly knotted after 1532. Relative chronology is notoriously difficult in archaeological khipus, given the relative confidence intervals of radiocarbon dating methods: setting aside questions of sample contamination, even narrow confidence intervals are probabilistic measures, and can fall evenly—say 20 years before and 20 years after—the moment of conquest. The physical structure of khipus does not establish deterministic relative chronology. With the possible exception of Wari (pre-Inka) and late colonial khipus, the basic structure of khipus in-between is comparatively consistent. In contrast, khipu transcriptions allow researchers to determine a timeline of khipu readings—to study diachronic changes in khipu narration not only before or after 1532, but even within the first 50-70 years of the colonial Spanish viceroyalty. This was a unique time period in Peru, as many of the khipukamayuqs would have produced khipu cords under both Inka and Spanish rule.

The study of khipu grammar is rooted in the composition of its narrations, which offers the opportunity to synthesize multiple fields in quantitative historical linguistics. Tools from

---


diachronic studies, document feature selection and authorship attribution will serve useful in our attempt to wrestle with the “bad data” of khipu transcriptions. It is anticipated that this hybrid, corpus-based approach might enact a “conceptual change of pace” from an example-based methodology rooted in exegesis to a multivariate statistical framework, where frequency reflects the functional input of a range of possible variables. These methods are applied to the Textos Andinos: the richest collection of khipu transcriptions available today.

---

Chapter 2

An Exegetical Study of the Textos Andinos

2.1 Introduction to the Textos Andinos

I analyzed two volumes of colonial-era Spanish khipu transcriptions, published in 2004 (vol. 1) and 2010 (vol. 2) through the joint efforts of the Ibero-American Institute of Finland and the Complutense University of Madrid. The transcription compilation project began in 1982 at the Archivo General de Indias (AGI) in Seville, Spain. The two texts comprise 880 pages, including 43 document groupings of varied length that Pärssinen and Kiviharju argue include transcriptions of, or commentary on, khipus. The documents range in original recording date between 1535 and 1653. However, a majority of the documents predate the Toledan viceroyalty (beginning in 1569). The data set includes 72 memorias (khipu transcriptions), detailing khipus related to military conquest, indigenous petitions, interrogations during trial proceedings and recounting of tribute paid in the Inka Empire, among other contexts. Figure 2.1 plots the geographic origins of the 72 khipu transcriptions, with numbers corresponding to an enumerated metadata appendix included at the end of the study (see appendix B). The substantial geographic diversity represented in the corpus is immediately apparent.

---

81 It is important to note that the transcriptions seldom mention explicitly the use of khipus as sources of narrated messages. Often instead present is the memoria (“memory/recounting”), which is traditionally associated with texts derived from khipus. To remain faithful to this tradition, I use the terms “transcription” and “memoria” interchangeably throughout this study. See Pärssinen and Kiviharju, Textos andinos, vol. 1, 39, 84.
I use the term “transcription” to describe a category of sixteenth-century Spanish documents: those which recorded data—linguistic or numerical—read out and transcribed from the cords of khipus. Analysis of khipu transcriptions invites a clarification of the challenges involved in interpreting colonial documents. The written document is merely the final resting place of a three-stage process of interpretation: an indigenous khipukamayuq narrated his khipu in Quechua or Aymara; the vocalized narrative was translated in real time into Old Spanish by a

---

82 Map generated using the ggmap package in R. See Kahle and Wickham 2013.
Spanish interpreter; and, a Spanish scribe wrote the translated narration using standardized notary techniques. Immediate questions arise regarding the viability of direct translation. For example, how would the conquering Spanish translate “llama”—an exotic camelid never-before seen by European eyes? Indigenous measurement systems presented especially formidable challenges to Spanish scribes—Castilian conversions of Andean measurement units form the most frequent errors in the corpus of colonial khipu transcriptions.

An extended analogy may aid in understanding the challenge of using colonial khipu transcriptions to study khipu narration. An autonomous player piano operates by means of a perforated metallic cylinder. Each hole represents a note, with an associated length, such that the passing gap breaks a vacuum in the piano and activates the relevant hammer, striking the chord and producing music. Similarly, the Spanish chroniclers tell us that khipu knots—the “holes” of the player piano cylinder—passed through the hands of the khipukamayuq, producing narration. The khipu texts studied herein are best compared to written transcriptions of the player piano music by passive bystanders. The problem is to reconcile the almost 1,000 pages of transcribed sheet music with 1,045 archaeological khipu “player piano cylinders”—without knowing the appearance, pneumatic structure or number of keys on the intermediary piano. In addition, we add that the music transcribers suffer from hearing loss, have never before seen a player piano and have just learned how to transcribe sheet music. This is the challenge of decipherment studies from colonial-era khipu transcriptions.

Further, even a cursory look at the Textos Andinos introduces additional challenges. Pärssinen and Kiviharju included documents in the volumes written both from and about khipus. Even documents which the editors argue contain khipu transcriptions vary in form and structure. These are grouped below into three categories, to clarify subsequent analysis. Namely, I offer an exegetical typology of khipu transcriptions—based in close reading and literary analysis—that I will follow with a statistical typology that makes use of multivariate techniques to test the

---

categories proposed in the first application. The synthesis of these methods—their relative strengths and considerations—is of primary importance in this study.

2.2 An Exegetical Khipu Transcription Typology

1 – Primary Transcription

I label “primary transcriptions” those writings which contain continuous, uninterrupted narrations of khipu accounts. A khipu reading is considered continuous if the transcribed text plausibly follows the original order of recording in the khipu. The transcribed account of a khipukamayuq, vocalizing his khipu from beginning knot to dangling end, qualifies as a direct transcription. Further, an uninterrupted reading here refers to khipu transcriptions which, once begun in writing, continue without record of interjection or questions asked by visiting Spanish officials. These alternative cases are captured by subsequent categorizations.

Already, the above distinctions invite consideration. What incentives did khipukamayuqs have to vocalize their khipus truthfully to conquering Spanish officials? Might “uninterrupted” khipu readings implicitly answer questions from Spanish administrators, omitted from the written record by the scribe? Even if uninterrupted, would a khipukamayuq’s reading proceed uniformly from beginning to end of the primary cord, rather than vocalizing only the cords relevant to the particular Spanish request?

The above questions are dutifully acknowledged, with the following proviso: an attempt to isolate a distinct insistence among colonial khipu transcriptions requires preliminary document groupings, despite the risk of miscategorization. The proposed labels offer a heuristic blueprint for bringing order to the breadth of sixteenth century Spanish writings—including their inconsistencies, biases and internal tensions. Mackey’s question—“What category of number”—is applied to textual analysis and redeployed to ask “What category of transcription?” It is hoped that segmented document “families” will allow for closer intertextual comparison, approaching an understanding of exegetical khipu transcription typologies.85

Two categories of primary transcription are detailed below:

Numerical

Numerical primary transcriptions capture the contents of numerical khipus—common examples include enumerations of crops, soldiers, taxpayers and livestock. My categorization

85 See Urton and Brezine 2013, from which I borrow this terminology.
allows for narrative accompaniment, so long as each idea, person or thing is grounded in a numerical value. An illustrative example: the village contributed 14 fanegas of chuño (freeze-dried potatoes), 22 pairs of sandals and 7 laborers to aid the construction projects of the Inka king. Excerpts from the *Textos* are introduced to aid in understanding each category.

Our first example arises from an administrative visit to the Chupayuchu ethnic group, ordered in 1549 by viceroy Pedro de la Gasca as part of a national tribute assignment, instituting the first large-scale postconquest recording of Andean statistical information. The two visitors, Juan de Mori and Hernando Alonso Malpartida, recorded labor tribute offered by the four guarangas (1,000-person administrative groupings)—about 4,000 adult laborers—to the Sapa Inka before the Spanish arrival. The selected excerpt reinforces the importance of khipus to early colonial governance: Spanish administrators made use of “numeros viejos” [“old numbers”]—the data of preconquest khipu accountancies—to calibrate optimal tax levels. This calibration included region-specific segmentation of labor tribute, projected from past labor obligations in Tawantinsuyu—the name the Inkas gave to their empire. The excerpt quoted below emerges from a convening of caciques (native chieftains) in the regional seat of power, Conchumayo.

[The khipukamayuqs] were asked what service this said province of the Chupaychos performed for the Inka in Cuzco… they said that… they gave 400 Indians to sow the fields in Cuzco, so that the people there could eat, and to make their offerings. And, to serve as continuing yanaconas [displaced servants] for Guayna Capac, 150 Indians. And, for the continuing protection of the body of Topa Inka Yupanki after his death, 150 Indians. And, for protection of his weapons, 10 yanaconas. And, for the guarding of the people of Chachapoyas, 200 Indians. And, for the guarding of Quito, 200 Indians. And, for the protection of the body of Guayna Capac after his death, 20 Indians. And, to make feathers, 120 Indians. And, to extract honey, 60 Indians…

---

87 Chirinos Rivera, *Quipus del Tahuantinsuyo*, 79.
The passage maintains the properties of primary numerical khipu transcriptions: the khipu reading proceeds, uninterrupted, in response to the original question posed by the visiting administrator. We arrive at symmetric repetition, a first principle of insistence in khipu transcriptions. Vocalized cord-by-cord by an indigenous khipukamayuq, the khipu was logged by the Spanish scribe using repetitive, almost biblical, syntactic consistency. The repeated “and” serves as a notational unit, separating distinct clauses detailing tributaries committed to sowing seeds, protecting the weapons and bodily remains of deceased Inka rulers, collecting feathers and extracting honey, among other tasks detailed in the transcription. The Aschers note the rhythmic property of khipus like those from Conchumayo, whereby repeated numerical values produce what is presumed to have been a symmetric narration. Indeed, the transcription confirms this hypothesis in writing. The Aschers’ rhythmic numerical khipu properties extend to the linguistic output of symmetric numerical inputs. As visible, there exists a mapping between identity (e.g. “Indians for sowing seed in Cuzco,” “Indians for protection of Chachapoyas,” etc.) and magnitude (e.g. 400, 200, etc.)—each quantity is mapped to a descriptor, while no described identity is left without a quantity. These descriptor-magnitude pairings are easily conceptualized as key-value pairs, with the conjunctive “and” serving as the linguistic comma. One wonders: were the linguistic labels encoded in the original khipu cords, tied, spun or twisted in association with the enumerations of tribute laborers? Historical khipus offer potential insights into this challenging question.

Narrative

Narrative primary khipu transcriptions offer uninterrupted readings of historical khipus—the mentions of “speeches,” “ambassador visits” and “events of war” detailed in the writings of de la Vega that motivate this chapter. This class of khipus likely operated through tactile memory, as discussed in section 1.3. In contrast to numerical primary transcriptions, narrative recitations remove magnitudes and insert discrete ideas, people or things. An illustrative example: Village A specializes in sandal production; village B specializes in chuño production; and, village C specializes in cumbi cloth production (note the use of symmetric repetition). The chosen passage offers a narrative transcription from a military context.

 Ascher and Ascher, Mathematics of the Incas, 122.
Our second example derives from a probanza—a sworn witness testimony with notarial significance—executed in Cuzco between 1569-1571. Attributed to Capac Ayllu, the narrative text details the fifteenth-century military conquests of brothers Amaro Topa, Topa Inka and Topa Yupanki, constituting one of the oldest and most complete texts transcribed from khipus. The khipu readers, introducing themselves as descendants of the three siblings above, presented the list of military conquests as an attempt to demonstrate royal lineage. Royal ancestry had political consequences at this time: proven descendants of former Inka nobility, including caciques, were considered “hereditary officials” under Spanish colonial law, exempted from paying tribute to and performing personal services for the occupying viceroyalty. A persuasive case is made by the presenters, considering the granularity of the following historical account:

Memoria of the provinces that Topa Inka Yupanki, father of Huayna Capac Inka, conquered with his brothers Amaro Topa Inka and Topa Yupanki in the provinces of Chinchasuyu and Collasuyu; and Antisuyu; and Cuntisuyu; from Quito to Chile.

In the province of the Quichuas the stronghold of Yara Tuuaramar, and the stronghold of Cinamba, and those people under their rule. In the province of the Angaraes, which is Guamanga, the head of the whole province Urcuslla Curoslla Imrarcas; and obliterating him he captured Chuquis Guaman, its king. And then he conquered Tarma, Atabillos and some, out of fear, placed themselves in obedience. And then the province of Guaillas Guanuco was submitted to peace, and in this province he conquered Chuncho and Pillan; and likewise he conquered the provinces of Cajamarca and Guamachuco and Chachapoyas and Guayocondos.

The ancestral appeal fulfills the conditions of narrative primary transcriptions. The continuous reading proceeds logically—conquered regions are segmented by suyu (“part/province”).

---

93 Pärssinen and Kiviharju, *Textos andinos*, vol. 1, 93-94: lines 376-400 (memoria 2); my translation, paragraph break added. Original: “Memoria de las provincias que conquistó Topa Ynga Yupangui padre de Guaina Capac Ynga con sus hermanos Amaro Topa Ynga y Topa Yupangui en la provincia de Chinchaysuyo y Cullasuyo; y Andesuyo; y Condesuyo; hasta Quito Chile. En la provincia de los quichuas la fortaleza de Yara Tuuaramar, y a la fortaleza de Cinamba, y a los a ellos sujetos. En la provincia de los angaraes que es en Guamanga, cauеsa de toda la provincia Vrcuslla Cuosilla Ymrarcas y asolándola prendió a Chuquis Guaman su rey… Y luego conquistó a Tarma, y Atauillos y algunos se salieron a la obediençia de miedo de paz. Y luego la provincia de Guaillas Guanuco le salió de paz y en esta provincia conquistó a Chuncho y Pillan; y asimesmo conquistó, la provincia de Caxamarca y Guamachuco y Chachapoyas y Guayocondos.”
Repetitive elements appear, consistent with the symmetric structure introduced above. In this case, the conjunctive “and” [“más”] is replaced by the synonymous “and then” [“y luego”], often combined with “he [Topa Inka Yupanki] conquered” in a repetitive syntactic structure. Verbs in particular serve as demarcating units in the battle account, including later variants like “[the siblings] obliterated” [“asolaron”] and “[Topa Inka] brought order to” [“puso en orden”] that exhibit substantial linguistic diversity. While the resulting narration is symmetric, it is decidedly less rhythmic than our example from Conchumayo. Names of subdued leaders are sometimes included (e.g. Chuquis Guaman), but often omitted. Verbs and pronouns switch cases within suyus (i.e. singular to plural, or vice-versa), without naming the subjects of sentences: Did two or three sibling “obliterate” a valley? If not three, then which two siblings? Verb placement floats variably between the beginning and end of sentences. Linguistic ambiguity results, potentially signaling an indigenous reluctance to standardize the verb and pronoun cases and, in turn, selectively present the historical account. Several conquests are recorded twice in the broader transcription, inviting the question: Do the repeated regions represent lapses in memory—complicating the notion of a meticulously encoded historical khipu (which, presumably, would have prevented such errors)?

The asymmetric recitation reveals narrative khipu transcription as a dynamic, performative act, facilitated by the honed abilities of the indigenous khipu-reader. As such, I introduce elaboration—a second property of transcription insistence. Recalling the property of “anchor words” and signs/extrapolated meaning introduced above, khipus emerge as generative vehicles, whereby specific elements (whether pendant knots, twist, attachment, etc.) prompted recitation of additional details. We might hypothesize that Chuquis Guaman—the name of our vanquished local leader—was one such point of extrapolated meaning added to an otherwise formulaic recounting of the suppression of Guamanga. Despite these inconsistencies, the memoria is instructive as an uninterrupted, complete khipu recitation.

Unfortunately, not every colonial-period writing event followed the formal hearing structure of the Cuzco probanza or the crisp segmentation of the Conchumayo tribute reading—namely, an uninterrupted recording of a complete khipu account. Instead, we now turn to the two

---

95 Rowe, “Probanza de los incas,” 199.
96 Pärssinen and Kiviharju, *Textos andinos*, vol. 1, 84.
remaining categories of the exegetical khipu transcription typology, the first breaking the assumption of continuity and the second relaxing the condition of non-interruption.

2 – “Elicited” Transcription

The set of documented khipu reading venues expanded rapidly by the mid-sixteenth century following admission of knotted strings as testimony in Spanish legal proceedings. The growing use of khipus as legal documents cemented their threatening status: while tribute accounts of goods paid to Spanish bureaucrats were double-entered in khipus and writing until the 1550s, legal disputes over tribute discrepancies from this period multiplied as the original Peruvian conquistadors/encomenderos died and attempted transfer of land and labor privileges to their heirs. The resulting trial proceedings called for specific data to answer targeted questions: What was the value of tribute paid by village x from valley y to encomendero z between 1548-1551? What ought village x’s proportional share have been? Were some goods overcharged, while others were under-collected? These questions frame our third excerpt, while introducing the capacity of khipus as indigenous SQL databases—mobile repositories of data, queried to provide answers to precise questions.

Elicited khipu transcriptions capture vocalized subsets of khipus, chosen to provide specific points of recorded data. Such transcriptions need not be in response to questions—khipus narrated out of order, by choice of indigenous khipukamayuqs, are also captured by this designation. However, a distinct sense of “call-and-response” is present, whereby a conditioned khipu query (the call) elicits a response from vocalized khipu cords.

Our third excerpt centers quantitative tribute data, read out from khipus in La Plata (modern Bolivia) between 1572-1578, during a tribute restitution trial brought by the Aymara-speaking Sakaka Indians against the heirs of their former encomendero, don Alonso de Montemayor. The indigenous claimants demanded reparative compensation for excessive tribute solicited by Montemayor from the region of Charka, between 1548-1551, before the official implementation of de la Gasca’s 1551 national tax (legislation codified following the same set of 1549 administrative visits referenced in the first excerpt). Reading from their chinu

---

97 Chirinos Rivera, Quipus del Tahuantinsuyo, 168.
(meaning “knot” in Aymara), the indigenous witnesses presented accounts of goods—with corresponding prices—delivered to Montemayor during the years of overpayment. The following excerpt demonstrates the calculative precision of khipu queries:

And then [the khipukamayuqs] were asked to show, by means of the said khipus, what they gave to the said don Alonso and to other people in his name, during the first year of the four when they say they did not have a [official] tax; and they took their khipus in hand and said they had given the following; and by means of some small stones they had placed on the ground they went on making their count. Together with the khipu, they said the following…

They said, by the said khipu, that in the said year they gave to the said don Alonso de Montemayor 500 loads of chuño, placed in the said city of Potosí.

They were asked how much a load of chuño was valued at in Potosí, in the said year. They said six current pesos. Likewise they said by means of the said khipus and count to have given in the said year 605 camelids…

They were asked how many arrobas of tallow they gave in the said first year to the said don Alonso de Montemayor and taking their khipus in their hands, and by means of the said stones they said 20 arrobas of tallow.

They were asked how much each arroba was valued at in Potosí in that said year and how that which they paid in tax was sold. They said four current pesos [per] arroba.100

A distinct non-continuity is evident in our excerpt—that is, the transcription proceeds in the order of Spanish questions and not necessarily in the order of the vocalized khipu cords. The particular arrangement of Spanish inquiries elicits a matching order from the indigenous account—regardless of that order’s similarity to the original recording. Further, while symmetric repetition appears throughout the trial document (“they were questioned” [“fueron

100 Pärssinen and Kiviharju, Textos andinos, vol. 1, 304-305: lines 76-87, 112-125, 135-149 (memoria 17); my translation, emphasis added. Original: “E luego les fue pedido que muestren por los dichos quipos que dieron al dicho don Alonso y otras personas. en su nombre el primer año. de los quatro que dizen. que no tuuieron tasa e tomaron sus quipos. en las manos. dixerons dado lo siguiente y puestas vnas piedras en el suelo por las quales fueron haziendo su quenta. juntamente con los quipos. dixerons lo siguiente… dixerons. por el dicho quipo. que el dicho año. le dieron. al dicho don Alonso. de Montemayor. qui<<ni>>ientas. cargas. de chuño. puestos. en la dicha villa de Potosý. Fueron preguntados que a cómo valia. la carga de chuño. el dicho año. en Potosý. dixerons. que a seis pesos corrientes. Asimismo dixerons. por los dichos. quipos. y quenta auerles dado. el dicho año. seisçientos y cinco carneros de la tierra… Fueses. preguntado qué tantas arrobas de seuo. dieronle el dicho primero año. al dicho don Alonso de Montemayor e tomando sus quipos. en las manos e por las dichas piedras dixerons. que beinte arrobas de seuo. Fueron preguntados. a como valia. cada arroba en Potosý. aquel dicho año y a cómo se vendía. lo que ellos. pagauan. de tasa dixerons. que a cuatro pesos corrientes. arroba.”
preguntados”]) in separating quantity and price paragraphs, this judicial rhythm more reflects sixteenth-century Spanish trial practices than a grammatical khipu structure.

The use of a yupana—an Andean calculation board—confirms the elicited transcription designation. In response to a targeted Spanish inquiry (e.g. How many arrobas (roughly 20-30 pounds) of tallow were given to Montemayor in the first year?), the presenting khipukamayuq queried the relevant khipu(s), applying algebraic calculations to the values recorded on a subset of the cords to arrive at a response: 20 arrobas. The yupana was a well-documented aid to khipus, which were not themselves amenable to calculation. 101 Within the above except, the vocalized responses which comprise the khipu transcription are themselves mere excerpts from the original khipu. Put simply: if the khipu reading proceeded continuously, from beginning to end, we might expect a statement of each cord. Instead, the reading collapsed cords and cord categories into sums, calculated by the yupana, bypassing the granularity of a complete khipu reading in favor of “the bottom line.” It is worth noting that the vocalized totals (500 llama loads of dried potatoes; 605 llamas; 20 arrobas of tallow, etc.) were likely summed already in top cords on the chinu; in which case, the stones used in court served a mere “confirmatory, authorizing function” to reach an already-memorized result. 102

However, such elicited transcriptions remain instructive. For example, chronicles from Jauja note executive summary khipus, knotted to condense data from fully descriptive khipus. 103 In Charka, the Montemayor account implicitly defines the construction of the chinu presented in trial proceedings: We infer that Sakaka ethnic accounting (knotted in animal fiber chinu) maintained even finer distinctions—later summed for presentation—than necessary for use in a technically complex tribute restitution claim within the Iberian legal system. Further, khipu testimony emerges as a precise and legitimate instrument of indigenous resistance: Sakaka omissions (i.e. omitting tribute not paid) in a subsequent reading of their chinu speak to the strategic maneuverability of tribute calculations, to the benefit of the claimants, 104 even while broader khipu testimony gained numerical legitimacy among the chroniclers, including de Alconchel:

102 Platt, “Without Deceit or Lies,” 249.
103 Chirinos Rivera, Quipus del Tahuantinsuyo, 58-60.
I know that the testimonies offered by way of khipus are true and genuine because these testimonies, often and in many ways, have been verified by way of accounts I have held with the Indians, concerning goods which they have given and which they have owed, and I have found that the khipus held by the said Indians were quite certain.\textsuperscript{105}

A final consideration arises: Why include regional transcriptions—in this case vocalized from postconquest Aymara khipus—in direct comparison with narrations of the Quechuanarrated “numeros viejos” from our first two excerpts? A possibility of homogenizing khipu accounts is acknowledged, although I analyze language of narration explicitly in the subsequent statistical analysis.\textsuperscript{106} Inconsistencies notwithstanding, the broad textual data analyzed herein embody the types of linguistic flexibility for which we aim to account in our search for a khipu transcription insistence. Put simply, researchers ought not be choosy: given such a small subset of khipu narration surviving in writing, each available sample merits analysis. Further, khipu readings from Quechua and Aymara-speaking regions are legitimately compared: while vocalized readings emerged in two indigenous languages, their transcriptions converged on Old Spanish—allowing inclusion in our profile of colonial Spanish documentation. Inclusion of postconquest documents allows us to track khipus chronologically as a device of accommodation—the third property of transcription insistence. Postconquest khipu transcriptions offer glimpses into native incorporation of Spanish social customs. Separate chinu encoded the price of goods in Charkas, demonstrating developing Andean conceptions of monetary value and exchange in Sakaka accountancy.\textsuperscript{107} The Third Provincial Council of Lima (1583), following a declaration of historical khipus as idolatrous objects to be burned just two years earlier, codified use of khipus as mnemonic aids for recording of sins and memorization of prayers—a selective imposition of meaning into which khipus adapted as rosaries in Jesuit

\textsuperscript{105} Quoted in Chirinos Rivera, \textit{Quipus del Tahuantinsuyo}, 68; my translation. Original: “Y sabe este testigo que los dichos sus quipos son muy ciertos y verdaderos porque este testigo muchas y diversas veces ha cotejado algunas cuentas que ha tenido con indios de las cosas que le han dado y le han debido y les ha dado y ha hallado que los quipos que tenían los dichos indios eran muy ciertos.”

\textsuperscript{106} Platt, “Without Deceit or Lies,” 257. Cereceda (1986) argues that khipus ought not be compared in parallel because they were “living beings,” constantly unraveled and raveled again, such that remaining khipus today—and presumably their surviving transcriptions—are best thought of in a living social context with respective, and distinct, states in time.

parishes. Examples abound of khipu flexibility in accommodating colonial systems of meaning and questioning. As such, elicited transcriptions—non-continuous queries of khipu databases—trace a chronological profile of transcription insistence.

3 – “Summary” Transcription

A final category captures khipu transcriptions significantly abstracted from their original context—here deemed “summary” transcriptions. Broadly, I refer to the wholesale summation of numerical khipus across time and space to produce regional totals of population, goods, etc., often enumerated in encomienda titles from the early colonial period. In contrast to the yupana-aided summation of cords within a single khipu or across several simultaneously present khipus, summary transcriptions account for khipu arithmetic at the broadest levels of governance. An illustrative question: How many fanegas of wheat were paid in tribute by indigenous residents of Jauja and Charka to the Spanish crown in 1551? These expansive regional summaries—often combining dozens of other summaries at various levels of governance, emerge in the chosen excerpt.

Our fourth and final excerpt comes from the encomienda title conferred, in Cuzco, by Francisco Pizarro to Hernando de Aldana in 1540. The grant recognized de Aldana for his role in the 1532-1533 siege of Cajamarca, as the messenger sent to convince the ill-fated Atahuallpa to enter the plaza of his eventual capture. The oddity of the encomienda title—in still-unconquered regions of Collasuyu, rather than near Cuzco, which would seem more fitting for de Aldana’s noble deed—conveys particular impressiveness to the specificity of the included population data:

... And I place under your control, in the province of the Aullagas, 823 Indians, with the cacique Acho 260 Indians in the following manner:

---


110 Abercrombie, Pathways of Memory and Power, 138.
One village called Acaluo, with the chieftain Gualca, [with] 38 Indians; and in another named Berenguela 18 Indians; and in another called Millme 53 Indians with the chieftain Colque; and in another called Pisquero 9 Indians; and in another called Yana 29 Indians; and in another called Callapa 17 Indians; and in another called Taparo 37 Indians; and in another called Yanaque 21 Indians; and in another village called Sacina with a chieftain [named] Acho with 14 Indians…

The encomienda document constitutes a high-level regional summary, synthesizing village-by-village population counts to arrive at the 823-Indian total. Symmetric repetition emerges only as a direct product of Spanish legal writing. However, a question of interpretation arises: Do the varied enumerations in the text constitute the sums of individual khipus or the sums of sums of khipus—perhaps the type of “bottom-line” executive summary khipu described in Jauja? The census statistics, while likely derived from khipus, raise questions of variable scale—the fourth principle of transcription insistence. Transcription documents often readily mix various levels of administrative accounting, without clear refinement of context for the outside reader. Further, tabular numerical data can be equivalently recorded using several variations of khipu construction. In other words, while a single khipu structure can inspire production of many possible narrations, varied khipu structures can also produce a single narration—depending on the extent to which readers are securely inserted into a shared system of signification and meaning. As such, we must infer the extent to which calculated totals (often reached with yupanas) sum across khipus and not solely within single khipus.

---

111 Pärssinen and Kiviharju, Textos andinos, vol. 2, 151-152: lines 36-52 (memoria 40); my translation, paragraph break added. Original: “… yo vos deposito en la prouinçia de los Aullagas ochoçientos y veynte e yndios con el cacique Acho duzientos y sesenta yndios en esta manera: vn pueblo que se dize Acaluo con el principal Gualca treinta y ocho yndios y en otro que se llama Berenguela diez y ocho yndios y en otro que se dize Millme çinquenta y tres yndios con el principal Colque y en otro que se dize Pisquero nueve yndios y en otro que se dize Yana yeynte e nueve yndios y en otro que se dize Callapa diez y siete yndios y en otro que se dize Taparo treinta y siete yndios y en otro que se llama Yanaque yeynte e vno y en otro que se dize Pucuro chácara del dicho caçique veynte e yndios y en otro que se dize Yanaqui chácara del dicho caçique veynte e vno yndios y más otro pueblo que se llama Sacina con vn principal Acho con catorze yndios…”

112 Chirinos Rivera, Quipus del Tahuantinsuyo, 61.

113 Pärssinen and Kiviharju, Textos andinos, vol. 2, 149. Further, de la Vega affirms that each region of the former Inka Empire maintained khipus which detailed the historical accounts of the area:

… [each region] has its accounts and knots with their histories and annals and their traditions, and for this reason, it is better remembered what happened in their kingdom than what happened in that of their neighbor.

Quoted in Assadourian, “String Registries,” 123.

114 See Ascher and Ascher 1997. The book includes numerous exercises which are instructive regarding this point.

This reality complicates attempts to match the contained census totals—the nine Indians of Pisquero; the 37 residents of Taparo, etc.—with cords in the KDB, given our uncertainty as to the number of khipus which combined to form the 823-person total. For example, Pucuro, a village of three, was likely enumerated from a single khipu (potentially three cords or three groups of cords on a larger khipu). However, we might hypothesize that larger villages—Quillaca, for example, claims a population of 174—maintained separate khipus which each enumerated different moieties or regions of the village. Further, even if a village population was recorded on a single khipu, we lack context: were individuals registered as single cords, six-cord groups or perhaps some other configuration? Utilizing the above two options: If searching for Quillaca’s census khipu in the KDB, ought we search for khipus with 174 pendants, or 1,044 pendants (= 174·6 in the case of six-cord groups)? Perhaps instead sets of khipus whose sum of pendants equals 174? The challenge is already apparent. Figure 2.2 summarizes the exegetical typology diagrammatically.

![Figure 2.2: Exegetical Khipu Transcription Typology](image)

How might we test the durability of this exegetical typology? Data-based approaches facilitate an efficient means of studying multiple hypotheses. Two such methods—Principal Component Analysis and Hierarchical Agglomerative Clustering—incorporate the tools of quantitative historical linguistics, applied to the compiled corpus of khipu transcriptions.
2.3 On the Importance of Typological Thinking

This study motivates an enabling infrastructure for the large-scale investigation of khipu grammar in khipu transcriptions. What Urton and Brezine call “basic typological questions” in their aggregate archaeological khipu study I call not only basic, but essential and fundamental to the understanding of khipu transcriptions.\textsuperscript{116} That is, it is surprising, given the sheer number of written samples available of khipu narrations (although this number is relative; see below), that a study has not yet compared more than a handful of these documents simultaneously. The existing literature provides fertile ground for this project: extremely informative studies of individual khipu transcriptions, based in exegesis, have an over 40-year history.\textsuperscript{117} To complement these studies, I hope to provide some unifying principles of khipu transcriptions and khipu grammar preserved in ink and paper. That is, I suggest that further research into these varied documents would benefit from a typology—a \textit{primordial division} of the texts into working categories of analysis. Aggregate study of khipu texts offers a glimpse into trends in khipu construction that may only be discernible at high levels of abstraction. In the words of Van Dijk (1971):

“Classification in literature… can be made explicit only when we have a profound insight into the nature of textual structures and their underlying tales and \textit{categories in general}.”\textsuperscript{118}

Is there not a tinge of scientific reductionism evident in this whole exercise? A seemingly insatiable desire to categorize, classify and partition khipu texts into cleanly distinguishable buckets? Two observations are offered here. The first is that aggregation and close reading are not adversarial characters in a constant struggle for the attention of the researcher. Indeed, the findings of corpus linguistics techniques are often contextualized by close study of the texts contributing to the essence of a particular cluster; close reading is enhanced by the embedding of specific vocabularies within an overarching khipu transcription grammar. The techniques employed in the quantitative analysis echo this aim by being based in the texts. Put concisely by Boone: “it can be useful to explore characteristics that might lead to a typology of Amerindian


\textsuperscript{117} Several examples include: Brokaw 2003; Part II of Curatola Petrocchi and de la Puente Luna, eds. 2013; Loza 2001; Murra 1975; Platt 2002; Urton 1998.

\textsuperscript{118} Teun A. van Dijk, “Foundations for Typologies of Texts,” \textit{Semiotica} 6, no. 4 (1972): 320; emphasis added.
systems, a typology that embraces all but highlights their internal differences.”¹¹⁹ The second observation is that this study seeks to offer non-deterministic principles which best highlight the diversity of the transcriptions. This is not a Linnaean taxonomy of khipu transcriptions: while the principles offered are based in close reading and statistics, they by no means impose any single “identity” on a document. I offer a typology of these wonderfully diverse texts to study which linguistic elements differentiate them most strongly; and whether such elements could plausibly have been signed in khipu cords. To define “linguistic elements” I borrow Urton and Brezine’s understanding in their own study of archaeological khipus of “information” as “any detectable ‘difference,’ or variability… within or between samples—without difference, there is no information.”¹²⁰ In this sense, the value in typological thinking is contextual inference from the relative presence or absence of certain linguistic features: we can learn about the transcriptions from their own words. The study brings us closer to deciphering khipus by expanding our investigative vocabulary to take full advantage of the soundbites khipukamayuqs have left behind. Indeed, “the most successful approach to studying the khipu remains a comparative, systemic, and typological one.”¹²¹

It is also essential to state what this study does not attempt to probe. The analysis does not center the numerical content of the transcriptions: how many loads of chuño were delivered to an encomendero; the price of 100 eggs; etc. The granular details of specific lists—the quantities and names of particular items in particular places—are independently important, and it is argued that the broad principles suggested in this analysis might bring additional context to previous studies that have employed exegesis. In other words, although the explicit focus is not on different goods—the first mentions in the historical accounts of pigs or wheat—these categories are approached through aggregate study of lexical diversity in khipu transcriptions. The analysis does not focus on the mathematical operations involved in khipu recording. Recall that khipus were not calculating devices—the numerical cords held the results of particular calculations, performed by use of counting stones and yupanas. While counting and calculation

¹²⁰ Urton and Brezine, “Khipu Typologies,” 320; emphasis added.
appear quite often in the texts under study, a theory of khipu enumeration is not proposed here.\textsuperscript{122} The analytical study does not attempt to assign differential importance or significance to any particular transcription. Indeed, the corpus includes some transcriptions (e.g. Cobo’s recording of the Cuzco ceque system; memoria 3) which have numerous scholarly works dedicated to them, while others (e.g. the statement of the caciques of Machaca la Grande regarding their 1540s tax payments; memoria 68) count the Pärssinen and Kiviharju volumes as one of their first public appearances in a published volume. The transcriptions are evaluated with equal weighting for this exploratory purpose: the texts, beginning from a shared starting point, are allowed to differentiate themselves. Lastly, the study does not center the evolution of specific grammatical constructions in Old Spanish (e.g. the relative usage of the preterite versus imperfect tense across time and space). Instead, I use broad lexical categories to approach khipu transcriptions as flattened narrations—to detect underlying, shared patterns persisting underneath the confusion of sixteenth-century Spanish grammar.

This brings us back to the question of granularity: what is this study about, exactly, if not about the content of the narrations? At its core, this study is an analysis of the characteristic narrative structure of khipus. A khipu transcription typology allows us to uncover the extent to which khipu contents (whether wheat, laborers or llamas, etc.) were embedded within a relatively standardized vocal infrastructure discernible from the transcriptions of Inkaic khipukamayuqs narrating pre- and postconquest khipu cords. I aim to plot out the vehicle of khipu transcriptions, to better structure how we fill in its contents. The primordial division concept introduced above serves as an analytically and empirically evaluable typology—to approach the fundamental communicative operation of how khipus were vocalized by the former Inka cord keepers.

\textsuperscript{122} See Chirinos Rivera 2010; Curatola Petrocchi and de la Puente Luna 2013; Moscovich 2016; for the most thorough recent treatments of proportionality and calculation in khipu recording. Chirinos Rivera in particular offers an illustrative section of the manuscript dedicated to a theory of the procedures used with yupanas to perform arithmetic calculations.
Chapter 3

Methods: Formation of a Digital Khipu Transcription Corpus

3.1 Treating the *Textos Andinos* as a Corpus

Within the typology of natural language corpora, the khipu transcription corpus I compiled for this study is specific, dynamic and diachronic.\(^\text{123}\) The corpus is specific, as compared to general, since its aim is not to describe an entire language (e.g. Quechua or Spanish), but instead to represent a specific variety of Spanish: the scribal transcription of khipu narrations. The corpus is dynamic, rather than static: while the number of extant khipu transcriptions is fixed, the corpus is expected to expand as new transcription documents emerge from archives and private collections. The corpus is also diachronic, as compared to synchronic, since chronology is explored as a meaningful variable for describing linguistic variation.

I followed two guiding principles of corpus compilation: representativeness and balance.\(^\text{124}\) These ideals, or close variants of these terms, distinguish corpora from indiscriminate collections of things. Principally, a corpus is a sample of samples. This means that a corpus constitutes some fragment of existing records of a language variety or topic. The representativeness of the sample—the range of variation in the target language which is represented in the corpus—guides the breadth of the questions available to the researcher. There is an explicitly distributional consideration here: a corpus used to study the “language” of khipu narration ought to include—to all extents possible—transcriptions of both numerical and narrative khipu readings. For example, hypotheses regarding the recording of currency in khipu cords are not meaningfully drawn from a corpus lacking transcriptions of currency khipus. A representative corpus is also called balanced when the distribution of constituent texts faithfully approximates the actual proportion of genres in the target language variety. In our context, the guiding question is thus: does the compiled corpus of transcriptions from the *Textos Andinos* encompass the full range and distribution of khipu genres narrated by former Inka khipukamayuqs?

---


\(^{124}\) Ibid., 3.
Additional considerations surround corpus size. What Desagulier calls an “arms race” in the rate of corpus growth is largely attributable to the advent of computers, the internet and the decreasing price of electronic disk space, which have jointly made the compilation of large corpora increasingly simple. This overarching size imperative is evident in macrolanguage corpora—massive textual data sets used to study trends in the most prominent modern languages. In English alone, several include: The Corpus of Global Web-Based English (1.9 billion words); the British National Corpus (100 million words); and the Corpus of Contemporary American English (600 million words). One overarching aim of large corpora is to capture rare linguistic forms, which may otherwise escape undetected in smaller samples. Along with representativeness and balance, these principles broadly guide textual corpus compilation.

Do the *Textos Andinos* conform to these principles? The answer depends on the research question being suggested. This study aims to form an empirically-grounded typology of early colonial khipu transcriptions. Given this aim, the corpus is almost fully representative, since most-all known khipu transcriptions are included in the Pärssinen and Kiviharju volumes. In addition, conclusions about the language of khipu narration more broadly may also be suggested, since the corpus contains representative constituents of both numerical and narrative khipus, including enumerated lists, currency, lineages and narrative accounts. Section 2.1 gives a sense of the breadth of the documents included in this study.

The question of balance is more challenging. Whether the khipu transcriptions studied here form a balanced corpus depends again on the research question. Recall that balance describes the extent to which the distribution of genres in the corpus matches the distribution of genres in the target language. If the target language is khipu transcriptions, we might ask: does the relative saturation of this corpus with transcriptions from legal proceedings accurately represent the range of contexts in which khipus were transcribed in early colonial Peru? In this context, the corpus is strongly representative: few venues *other than* Iberian legal proceedings saw Spanish transcriptions of khipukamayuq narrations during this time period. Using a broader target of khipu narration, however, invites reflection. The reader is reminded that surviving transcriptions constitute a miniscule fraction of the sum total of khipu narrations that must have taken place in the Inka Empire and the early colonial period. Khipus were vehicles for

---

126 Brokaw, “La recepción del quipu,” 121, 140-141.
both horizontal and vertical transmission of information: the chroniclers tell us that most regions had at least two khipukamayuqs to record data independently from each other, and to interpret the khipus of neighboring regions to ensure that Inkaic labor tribute was equitably distributed. Yet, the vast majority of transcriptions bypass these civic narrations of daily life to represent khipu reading as a passive recital for Spanish testigos (witnesses).

Two observations are offered. The first is that we are limited in capability by the sparsity of the historical record. That is, we cannot add another fifty transcriptions to our corpus by traveling to Peru and watching khipu narrations during summer fieldwork! One of the staples of corpus linguistics is using aggregation (within the constraints of available resources) to reach profound hypotheses about the target language. Just as paleontologists comment on the behavior of dinosaurs from the miniscule fraction of their surviving, fossilized remains; or as the astrophysicist makes inferences about the universe from the small slice of it which is observable to us; so too does this study use the available record to study the target language of khipu narration in and after the life of the Inka Empire. The second observation is that the numerical/narrative distribution of the transcription corpus is broadly concordant with the body of archaeological khipus, which itself is considered a workable representation of khipu use during the Inka Empire. The overwhelming majority of transcriptions in the corpus (69/72 = 96%) includes some numerical information—counts of currency, llamas, chuño, etc. The most recent query of the KDB found that 92% of Inka-style khipu cords record a single numerical value (conforming to the Lockean base-10 numerical system). Recall that a more holistic division finds that about two-thirds of cataloged khipus can be confidently stated as numerical (see chapter 1). In addition, archaeological khipus do not communicate their genre (apart from calendrical and decimal tribute assignment khipus). As such, the archaeological record, which


129 This high-level measure considers the “primary narrative transcription” category from the exegetical typology to be the only explicitly non-numerical class of transcriptions.


is studied despite incomplete information, is offered as an analogy to making plausible conclusions from the khipu transcription corpus—an impressive collection of narrations spanning a wide variety of genres across time and space.

These observations are further contextualized by a numerical summary of the corpus. Table 3.1 includes a broad set of summary statistics about the 72 transcriptions in the corpus, including the distribution of high-level genre designations.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of transcriptions</td>
<td>72</td>
</tr>
<tr>
<td>Number of distinct author attributions</td>
<td>28</td>
</tr>
<tr>
<td>Number of distinct geographic locations</td>
<td>31</td>
</tr>
<tr>
<td>Number of lines</td>
<td>10,495</td>
</tr>
<tr>
<td>Number of word tokens</td>
<td>71,400</td>
</tr>
<tr>
<td>Number of distinct lemmas</td>
<td>5,219</td>
</tr>
<tr>
<td>Most frequent lemma</td>
<td>y (5,790)</td>
</tr>
<tr>
<td>Number of distinct word tokens</td>
<td>5,978</td>
</tr>
<tr>
<td>Mean number of lemmas per transcription</td>
<td>991.7</td>
</tr>
<tr>
<td>Median number of lemmas per transcription</td>
<td>388</td>
</tr>
<tr>
<td>Standard deviation (lemmas per transcription)</td>
<td>1,959</td>
</tr>
<tr>
<td>Mean number of distinct lemmas per transcription</td>
<td>155.9</td>
</tr>
<tr>
<td>Median number of distinct lemmas per transcription</td>
<td>102.5</td>
</tr>
<tr>
<td>Standard deviation (distinct lemmas per transcription)</td>
<td>185.6</td>
</tr>
<tr>
<td>Minimum number of distinct lemmas per transcription</td>
<td>27 (M52)</td>
</tr>
<tr>
<td>Maximum number of distinct lemmas per transcription</td>
<td>1,163 (M3)</td>
</tr>
<tr>
<td>Number (percentage) of primary narrative transcriptions</td>
<td>3 (4.2%)</td>
</tr>
<tr>
<td>Number (percentage) of primary numerical transcriptions</td>
<td>29 (40.3%)</td>
</tr>
<tr>
<td>Number (percentage) of summary transcriptions</td>
<td>32 (44.4%)</td>
</tr>
<tr>
<td>Number (percentage) of elicited transcriptions</td>
<td>8 (11.1%)</td>
</tr>
<tr>
<td>Number (percentage) of preconquest transcriptions</td>
<td>41 (56.9%)</td>
</tr>
<tr>
<td>Number (percentage) of postconquest transcriptions</td>
<td>31 (43.1%)</td>
</tr>
<tr>
<td>Number (percentage) of currency transcriptions among postconquest transcriptions</td>
<td>20 (64.5%)</td>
</tr>
</tbody>
</table>

Table 3.1: Transcription Corpus Summary Statistics

---

132 Table modeled after Jacques Savoy, “Authorship Attribution: A Comparative Study of Three Text Corpora and Three Languages,” *Journal of Quantitative Linguistics* 19, no. 2 (2012): 139 (table 1). The large standard deviations are attributable to highly variable excerpt length. I employed normalized frequency measures to mitigate the effect of text length (see chapter 4). However, it is emphasized that the study of khipu transcriptions requires a maximally-inclusive philosophy, given the scarcity of textual data which survives today.
It is emphasized that expansive definitions of language corpora are important in the study of endangered scripts (a term I use broadly to describe scripts with few available samples). In practice, this means leveraging the existing record in as powerful a way as possible, given the constraints of scarcity and resources. The immediate consideration is corpus size: can a 71,400-word corpus be used to make any defensible claims about a target language, when other corpora count millions—sometimes billions—of words among their ranks? Large corpora are not unambiguously “better” corpora. While the study of rare linguistic forms is a noted benefit of massive compilations, these same corpora face unique challenges. These include: the costs of quality control (the text annotation strategy described below relies in part on manual checks of the text, which are infeasible in large corpora); linguistic noise (massive corpora often mine the internet and unknowingly collect advertisements and metadata); and representativeness (the more automated the corpus aggregation, the more difficult to maintain a harmonious distribution of genres and topics), among others. Small corpora have already shown great promise in the study of language instruction, dialects in Lancashire, England, indigenous Dena’ina (the southcentral Alaskan language) and Texas German (a dialect spoken by the descendants of mid-nineteenth-century German immigrants in Texas). My analysis employs statistical methods to verify the adequacy of the textual sample size in interpreting the results. More broadly, however, study of the Textos Andinos continues with these principles in mind: powerful conclusions can be drawn from a relatively small subset of historical narration.

Before describing the digitization of the two volumes, I briefly detail the identification of the transcriptions. How were the beginnings and endings of excerpts identified during close reading? Recall from the introduction to the volumes that Pärssinen and Kiviharju called their compilation a corpus of Inka and colonial khipu texts—rather than a corpus of khipu transcriptions. The editors emphasized that the included texts include writings both from (transcriptions) and about khipus. As such, some documents were not included in my own

---

133 Desagulier, Corpus Linguistics and Statistics, 6.
transcription corpus because they do not plausibly contain the narrated data of khipu cords. However, even those texts that were included are excerpted. For example, in the transcription of an enumerated list of items given by the natives of Conchumayo to Gómez Arias de Ábila, I included only lines 135-224 of the original document (memoria 5; see appendix B). A direct excerpt, with line breaks preserved, illustrates the decision of starting and ending point:

/132 The said chieftains were asked what, /133 until the present, they give and have given /134 to the Captain Gómez Arias de Ábila and they said that /135 they give him, every 15 days, 100 pieces /136 of cotton clothing which every month constitutes /137 200 pieces in shirts and blankets. /138 And, they said that from now onward they will give /139 80 pieces every 15 days /140 because they are unable to give him 100…. /222 In addition, they said that will have given to the said /223 Gómez Arias, from workers and Indians /224 of service, 29 pieces [of unspecified goods]. /225 And, they said that all of these /226 said things is what they give and have given /227 to the said Gómez Arias and that it is what /228 they can give to him from now onward.138

The underlined portion retained for this study constitutes the best discernible approximation of where the elicitation ends and where the khipu narration begins. The distinction is signed in this account (and others in the corpus) with an introductory phrase such as “they said that,” followed by an enumerated list hundreds of lines long. Narrative endings are more easily traced by the break with the symmetrical rhythm of the transcription. In other cases, the narration ends once the speaker reaches the conclusion of the numerical data.

This process of discernment was repeated for each of the chosen transcriptions. The example highlights the active judgment involved in corpus compilation: the specificity of the target language (khipu narration) requires mining the text of interest from its surrounding Spanish technical/legal script. Crucially, I have recorded the bounds of each excerpt (the beginning and ending lines from the Pärssinen and Kiviharju volumes) in the metadata compilation (appendix B), to allow other researchers to study, in granular detail, the edges of the corpus. This choice allows for challenge and revision of transcription designations—that a

138 Pärssinen and Kiviharju, Textos andinos, vol. 1, 143, 145: lines 135-140, 222-224 (memoria 5). Original: “dixeron /135 que le dan cada quinze días ciento pie-/136 ças de ropa de algodón que son cada /137 mes dozientas pieças manta e camyseta. /138 E dixeron que de aqui adelante darán o-/139 chenta pieças cada quynze días por-/140 que ciento no las pueden dar… /222 Más dixeron que abrán dado al dicho /223 Gómez Arias de anaconas e yndias /224 de seruiçio veynte e nueue pieças.”
memoria begins one line earlier or later than proposed, perhaps—a process of academic iteration that, it is hoped, will converge on the most faithful representation of the target language. I describe the digitization and preparation of the corpus in detail in the subsequent sections.

3.2 Data Preparation and Transcription Cleaning

Digitization relies on active decisions made by the researcher which shape the future of the corpus. I discuss some considerations here regarding the conversion of the *Textos Andinos* to machine-readable text. Recall that Pärssinen and Kiviharju employed a liberal definition of khipu texts in their corpus (rightly so, for completeness), choosing to include texts that describe the use or curation of khipus, even if not containing a transcription of a khipu narration event. This decision is, in part, why this study extracted only a fraction of the total lines from the two volumes to form the corpus of khipu transcriptions. The two compilations (Pärssinen/Kiviharju’s and my own) achieve different goals. The larger collection (the two-volume set) maximizes breadth: the editors aimed to include as many documents as possible written both from (transcriptions) and about khipus. The collection, including the editors’ introductory notes for each document, serves as an invaluable informative resource for students of khipus. The subset of documents that I extracted narrows the scope of analysis to the documents—and only some fraction of lines from those documents—that contain numerical and/or narrative information from khipu cords.

After identifying the endpoints of the transcription texts in the printed volumes, I scanned the two volumes in their entirety to produce electronic versions of the texts. Figure 3.1 shows a sample page of text from the digitized *Textos Andinos*. 


The relevant sections of the scanned volumes were subjected to optical character recognition (OCR) and the texts extracted to a text editor, separated by headers identifying their source document. I manually corrected the OCR text, including reintroducing missing words and point correction of misinterpreted letters or typographic symbols (e.g. cameros → carnernos).

---

[“llamas”]). The texts included in this study track the order of the documents in Pärssinen and Kiviharju’s two volumes, such that the first transcriptions in the corpus (e.g. memorias 1-3) come from the beginning of the 2004 volume, while the last transcriptions in the corpus (e.g. memorias 70-72) come from the end of the 2010 volume. However, the numerical labels I assigned to the transcriptions do not correspond to the document numberings introduced by Pärssinen and Kiviharju. That is, memoria 34 is not necessarily from document 34 in the printed texts. There are two reasons for this choice. The first is that some document headers contain multiple transcriptions, since for the editors a “document” often means a legajo [“file”] from the AGI in Seville—which contains a collection of documents. The editors distinguished between documents in a file with alphabetic notation—e.g. document 32a, 32b, etc. I retained fully numerical labels for ease of processing. The second is that some documents do not contain khipu transcriptions. Documents in this category offer supplementary information regarding the uses of khipus in particular regions. The titles of the source documents are included in appendix B alongside my numerical labels, for ease of comparison with the printed volumes.

The extracted text required further cleaning before being subjected to analytical techniques. It is here where the attentiveness of the compiler is essential: my study works with transcriptions of transcriptions of khipu narrations. That is, I extracted text from Pärssinen and Kiviharju’s transcriptions of sixteenth-century Spanish documents, which are themselves transcriptions of khipus. As such, a detailed description of the steps taken in editing the texts is necessary not only for completeness, but also to explain my own analytical decisions in remaining maximally faithful to the volumes. I first removed metadata (included in the printed volumes) that is not intrinsic to the original document or my study. These data include folio numbers (assigned by the AGI for documentation purposes), in-text footnotes and artificial paragraph breaks (which the editors introduce for ease of reading).\footnote{See Pärssinen and Kiviharju, 	extit{Textos andinos}, vol. 1, 77-78 for the full set of transcription annotation notations employed in the printed volumes.} I then collapsed the line breaks and removed superimposed line numbers. The specific lines of the transcriptions are not studied here, since the division between lines appears to carry meaning in a middling fraction of cases. The familiar description of “line-by-line, cord-by-cord” khipu transcriptions is often difficult to detect in practice, since the entry for a cord may fill a line, with half of its final word spilling onto a second line. In reality, these two lines could be meaningfully combined and
studied as a single line. However, this implementation would then diverge from the line numbers of the original volumes. The lines of other transcriptions lack intrinsic meaning: the lines of a summary encomienda transcription, for example, represent aggregate summations of khipus from dozens of villages—in this situation, a line-by-line, cord-by-cord interpretation is not warranted, but the narrative structure remains relevant. As such, I focused on word choice, rather than sentence/line structure and punctuation, which are simultaneously influenced by scribal spatial concerns and the lack of uniformity in sixteenth-century Old Spanish.

In addition to non-content edits, I also incorporated the editors’ interpretations of illegible words and scribal mistakes. These include:

- Editor-introduced letters and words (denoted with angle brackets < > in the volumes) that replace partially-illegible, but unambiguous, tokens from the original documents.
- Suggested letters and words in cases of greater uncertainty (denoted with double angle brackets << >> in the volumes). While essentially the educated guesses of the editors, I incorporated the suggestions, given that a majority of the cases are quite intuitive. In memoria 3, for example, we are told that the Inkas offered a particular huaca (shrine) “carneros, ropa y con<<c>>has” [“llamas, clothing and shells”]. The editors found that the middle c in “conchas” is illegible, but suggested the word based on contextual information (e.g. that conch shells are offered to other shrines detailed in the list). There were 40 instances of this class of correction in total in the corpus.
- Suggested phrases in cases of highest uncertainty (i.e. when several illegible words can still be interpreted based on their context; denoted by [?] in the volumes). There were 18 instances in total in the corpus.
- Transcription errors either crossed out or erased by the Spanish scribe at the time of recording (denoted with braces { } and brackets [ ] in the volumes). While important in the edited volumes to remain faithful to the documents, I aimed to bypass noise introduced by Spanish scribal variation to study the underlying narration. As such, the 33 instances of erased or crossed-out words were removed, in line with the apparent wishes of the original scribes.
- In-line scribal corrections and additions were incorporated into the texts (denoted with slashes in the volumes /). Between-line and in-margin corrections were also incorporated, totaling 34 instances.
- Illegible words with uncertain count were removed (denoted by asterisks [***] in the volumes). There were six instances in the corpus.
- Countable numbers of fully illegible words were removed (enumerated with asterisks, e.g. [*2*] denotes two consecutive illegible words in the corpus). There were seven instances in the corpus.
- Words were removed when they contained enough illegible letters so as to make the word itself ambiguous. Four occurrences were found in the corpus. Words with a small minority of illegible letters (where [.1.] denotes one illegible letter in the corpus) were retained if their full meaning was unambiguous.
- References to blank spaces in the documents were deleted. The editors included a note in cases when the scribe left a significant amount of empty space, which is not of concern in this study.

The scribal edits in particular offer an intriguing window into the organic event of khipu narration, allowing a preliminary separation of narrative variation from scribal inconsistency. Some mistakes have little bearing on our analysis of khipu narrativity, the most frequent being variable spelling: one scribe spelled “huevos” [“eggs”] both as “hueuos” and “uebos” in the same line, only three words apart from each other. Another vacillated between “jaquima” and “xaquima” [“headstall” for horses] in close succession. Misspellings do not only apply to curious items like headstalls and eggs: one scribe was apparently ambivalent to corn, which he spelled both “mahíz” and “maíz.” Another spelled “ten” both correctly (“diez”) and incorrectly (“dyz”) on consecutive lines. Other in-document edits likely pertain to scribal mistakes. The repetition of one khipu narration impacted the scribe, who corrected the following mistake: “… in addition, [the tributaries gave] 15 chickens which were sold for one peso each. In addition, they gave four pesos ducks that were sold for one peso each. In addition, 540 eggs which were sold at five pesos for each 100 eggs.” The mistaken placement of “pesos” is likely due to homonymic confusion: the Spanish word for duck is “pato.” The symmetric repetition of the narration allows us to contextualize the performative act of khipu narration: we can envision

---

144 Ibid., 312-313: lines 603-604 (memoria 58).
the Spanish lengua and scribe so in rhythm, accustomed to repeated assignments of prices in pesos, that when another word beginning with a p emerged (“pato”), the scribe defaulted to muscle memory and accidentally described coinage instead of ducks.  

146 Similar mistakes arise in the midst of enumerated lists: one scribe, just having written of a sacred stream, accidentally called the subsequent shrine a stream, when it was in fact a hill; 147 another scribe, accustomed to the introductory phrase “in the first four years and four months…” accidentally included the “four months” clause when the introductory phrase transitioned to describing whole years only.  

The corrections are especially informative in numerical contexts, including the possibility of mistakes from the khipukamayuq himself. Take, for example, the following corrected transcription: “in addition, [the tributaries] gave to the said marquess one thousand two hundred and thirty-three pitchers of chicha [corn beer].” 149 This mistake seems unlikely to be attributed to a homonym or conditioned repetition: “doscientos” [“two hundred”] sounds unlike the rest of the number (“mil y treinta y tres”), and the only other mention of two hundred in this section occurs after the mistake. These factors make it appear likely that the mistake emerged (but was soon after corrected) during the khipukamayuq’s recitation from khipu cords. Other corrections are more thorough: “these said caciques of His Majesty have one thousand four hundred five hundred and seventy-nine fourteen alpacas and female sheep.” 150 If these mistakes emerged from the native narrator, might these totals have been adjusted at a subsequent date or in real time? Mistaking five hundred for four hundred is quite possible, especially in the tight knots of Inka-style khipu cords. This requires only miscounting the number of single knots at the hundreds level of the given cord: four single knots instead of five. However, the difference between 14 and 79 is more difficult to reconcile in real time, since this involves knots at different levels of the hierarchy of enumeration. In other words, a khipukamayuq would need to have missed the count of single knots in the tens place by six (one versus seven) and the units count by five (four versus 9), an improbable mistake by an expert cord-keeper. This assumes that the total (1,479 → 1,514) was recorded on a single cord. It is possible that the khipukamayuq employed a yupana for the

146 This specific mistake is not unique: the same mix-up occurs in Ibid., 307: line 231 (memoria 17).
147 Ibid., 135: lines 1322-1324 (memoria 3).
148 Ibid., 374: lines 338-339 (memoria 19).
149 Ibid., 202: lines 203-204 (memoria 11).
150 Ibid., 255: lines 125-127 (memoria 13).
aggregation of several cords, committing an algebraic error during the summation. It is also plausible that the khipukamayuq initially forgot to include the additional 35 livestock (alpacas and sheep), or that the 35 were added after the initial transcription \((1,479 + 35 = 1,514)\). A final example offers a similar mistake using smaller numbers: “… in another village called Cacajas [there are] 13 Indians of Pilco [the cacique]; in Ynputi [there are] 10 Indians of Pilco; in Quimayo, ten five Indians of Pilco…”\(^{151}\) Yet, in this example we arrive at a meaningful challenge to the claim that the khipukamayuq committed a narrative error. This discordance (five versus 10) is difficult to reconcile with the structure of a numerical khipu, since the conflated numbers differ by an order of magnitude. Namely, a five would have been recorded with a long knot of five turns, while 10 would have constituted a single knot at a higher horizontal level. Such a case of mistaken numerical identity seems improbable, to say the least.

These moments of correction, while signaling a mistake in one (or more) chains of the khipu narration event, are testaments to the on-the-ground challenge of a tripartite translation process. When a correction was entered into the transcription, and the incorrect value erased, was this due to (1) the scribe mishearing the lengua’s translation and clarifying the number verbally; (2) the lengua misunderstanding and clarifying the khipukamayuq’s Quechua narration; or (3) the cord-keeper making a mistake in reading the khipu and correcting himself, after the initial error had forced a linguistic cascade effect? The corpus offers examples reasonably attributed to all three cases. The numerical corrections offered above are exceptional. More often, adjusted numbers differ in the units place, often by one. In other words, the khipukamayuq must have missed (or added) a single turn of a long knot—a common mistake in the case of tightly-wound Inka khipu knots. However, larger discrepancies might be better attributed to the challenges of intercultural numeration—in particular, the difficulty exhibited by Spanish bureaucrats in understanding the Quechua numerical system.\(^{152}\) Regardless, the literary imperfections observed here motivate the study: to probe the narrative infrastructure of khipu readings against the linguistic noise endemic to the on-the-ground implementation of colonial taxation.\(^{153}\)

I conclude the section with a brief discussion of a final consideration: the possibility of typographic errors in Pärssinen and Kiviharju’s transcription of the documents. Rather than

---

\(^{151}\) Ibid., 417: lines 193-194 (memoria 22).

\(^{152}\) Pärssinen and Kiviharju, Textos andinos, vol. 1, 270.

\(^{153}\) Urton, Inka History in Knots, 97.
speculate in cases of ambiguity, I followed the work of the editors, with the exception of obvious typographical errors (e.g. changing “gAnado” [sic] to “ganado” [“livestock”]). The more challenging cases surround repeated words at the end/beginning of successive lines. For example: “[the witnesses] were asked how many chickens they gave in the said year to their said encomendero and [sic] to where [they delivered them].” Is the repeated “and” attributable to sixteenth-century scribal error or the 2004 corpus compilation? Regardless, the resulting sentence is grammatically incorrect—in this case, I removed the second “and” to approximate the natural language from the native source (which likely did not use two “ands” in a row), rather than capture scribal/modern linguistic variation. Given the relative infrequency (1) of this mistake (fewer than 10 occurrences in the corpus) and (2) of the mistakes listed above, these examples are offered as contextual considerations for preparation of the transcription corpus.

3.3 Spelling Standardization and Modernization

Khipu transcriptions present similar challenges as those that plague rare languages and scripts—namely, there are few (if any) available digital corpora, and those that do exist are not syntactically annotated. This last note is especially significant: an in-depth study of transcription grammars requires moderate linguistic granularity, else one faces a body of raw text without constituent categories. In this study, word type and lemma were employed as measures of linguistic diversity, necessitating both standardization and modernization of the transcription texts.

The memorias under study exhibit both synchronic and diachronic variability. That is, contemporaneous transcriptions employ different spellings and orthographic norms (e.g., the inconsistent use of the cedilla “ç” in “provinçia” [“region”]; different spellings of “huevo” [“egg”], etc.; see above), while these spellings and norms themselves evolve over the course of the near-century under study. In other words, I encountered intra-, inter- and intertemporal document variation, necessitating a rule-based approach to transcription stabilization and modernization. In practice, these two aims were in broad alignment.

The significant orthographic and grammatical variation observed is not unique to khipu transcriptions, since Old Spanish writ-large often lacks consistency. However, this inconsistency

---

155 Ibid., 329: lines 1003-1006 (memoria 17); emphasis added.
raises an impactful question in the study: do orthographic and morphosyntactic variations in the memorias reflect the idiosyncratic writing styles of particular Spanish scribes, or might these instead reflect the regional grammatical diversity of khipu narrations themselves?\footnote{Pedro Sánchez-Prieto Borja, “Desarrollo y explotación de un corpus de documentos españoles anteriores a 1700 (CODEA),” \textit{Scriptum Digital} 1 (2012): 428 applies this idea to thirteenth-century medieval Spanish texts.}

The historical record offers two notable insights into the above question. Central governmental efforts to stabilize Spanish did not emerge until 1713—with the chartering of the \textit{Real Academia Española}—leaving sixteenth-century scribes to follow the only existing set of spelling norms: those familiar from Renaissance Latin.\footnote{Cristina Sanchez-Marco et al., “Annotation and Representation of a Diachronic Corpus of Spanish.” In \textit{Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)}, eds. Nicoletta Calzolari et al. (Valletta, Malta: European Language Resources Association (ELRA), 2010), 2714.} Put more strongly, the sixteenth century was an “era of total anarchy in the use of [Spanish] spelling… attributable, above all, to phonetic insecurities… between historic graphical uses and a new phonetic reality.”\footnote{Beatrice Schmid, “«Es cosa dura hazer novedad»: Nebrija y la ortografía castellana,” in \textit{Colloquium zu Ehren von Germán Colón (Acta Romanica Basiliensia 9)}, ed. Maria Antonietta Terzoli (Basel: Romanisches Seminar, 1998), 60; my translation, emphasis added.} The first substantial academic effort to standardize Spanish orthography did not arise until 1517, with Nebrija’s publication of the \textit{Reglas de Ortografía}. Despite this relevant effort, why did our scribal protagonists remain consistently inconsistent? It is important to remember that standardization involved a process of diffusion: if there is any delay in widespread linguistic adoption, it might well arise among scribes under a satellite viceroyalty, transcribing indigenous khipu narrations almost 10,000 kilometers away from Nebrija’s Madrid residence. Further, spelling decisions were often influenced by paleographic factors, including the availability of page space and the particular choice of typography.\footnote{Sanchez-Marco et al., “Annotation and Representation,” 2714. See also Sánchez-Prieto Borja 2004.} These factors remain worthy of consideration as statistical study begins on the \textit{Textos Andinos}: grammar, handwriting and phonetics formed an organic and mutually-informing system of exchange that influenced the pen-strokes of Spanish scribes.\footnote{Sánchez-Prieto Borja, “Desarrollo y explotación,” 444.}

In order to normalize spelling within the corpus, I mapped historical tokens to modern Spanish, employing a rule-based strategy:

- Word tokens with one-to-one mappings were transformed directly, since their updated forms do not affect the word order or word count of any given sentence. Examples
include both familiar mappings like “cinco” → “cinco” [“five”]; and more subtle transformations like “yua” → “iba” [“went”].

- Antiquated word tokens lacking one-to-one mappings were transformed to their modern equivalents. Chief among these are clitic pronouns, which lack linguistic independence and require a verbal host in order to derive meaning. Examples include “dióles” → “les dio” [“he/she gave to them”] and “fueles preguntado” → “les fue preguntado” [“they were asked”]. Historical forms were included with their modern counterparts in an enriched-text markup script, in order to preserve indexing and continuity (see below).

Crucially, human annotation, while time-intensive, is one of the benefits of studying an endangered script: the small universe of available khipu transcriptions facilitates the “gold standard” of manual text markup.161 Text modernization in turn accomplished text stabilization, as modern Spanish exhibits substantial synchronic uniformity (that is, only one spelling exists for the most common word tokens at any given time) compared to its progenitor from the orthographic era of “total anarchy.” This text preprocessing proved crucial to performing targeted in-text searches, as it is not feasible to include all possible variants of a given word in each search.162 The modernized (and hence standardized) corpus was, in turn, submitted to an automatic linguistic annotation algorithm, described in further detail below.

3.4 POS-Tagging and Lemmatization

Part-of-Speech (POS) tagging (identifying and attaching the linguistic categories of word tokens to their hosts; e.g. labeling “Diego” as a proper noun, “correr” [“to run”] as a verb, etc.) and lemmatization (mapping word tokens to their dictionary entries; e.g. pairing “ate” and its lemma, “eat”) are available functions within several open-source software programs. These

161 Cristina Sánchez-Marco, Gemma Boleda and Lluís Padró, “Extending the Tool, or How to Annotate Historical Language Varieties,” in Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (Portland, Oregon: Association for Computational Linguistics, 2011), 4. Fully automated spelling normalization is currently only available for Old English (called Vard2; see Baron and Rayson 2008). This tool has been adapted for annotation of Old Spanish by Sánchez-Marco 2012. However, Sánchez-Marco 2012 finds that the adapted tool falls short of the automated linguistic mapping rules employed in the 2011 writing, mostly due to the larger contextual window (interval of words to the left and right) necessary for correct identification of the word under study. Since even the referenced linguistic mapping rules fall short of the manual “gold standard” employed herein, I elected not to use Vard2 for modernizing the corpus.

programs can annotate thousands of words in mere seconds, having been trained on billions of words in their development phases—most often consisting of modern news articles in the case of popular taggers. As a result, POS taggers and lemmatizers encounter challenges when presented with linguistic forms not used in news articles, Old Spanish being one such example.

Efforts to algorithmically process Old Spanish—a banner term that encompasses the medieval-to-early-modern Spanish language—are relatively recent, emerging in 2011 with an automated morphosyntactic/linguistic text enrichment software developed at the Universitat Pompeu Fabra in Spain. This study considered the central decision of the Pompeu Fabra research team: Is the most efficient means of linguistically annotating Old Spanish (a) to apply a modern tagger to the original historical text, followed by manual correction; (b) to adapt a modern tagger to interpret Old Spanish using a historical training set; or (c) to apply a modern tagger to modernized Old Spanish, retaining the historical forms as metadata? Option (a), while feasible in theory, grows quickly in complexity, given the sheer number of historical spelling variants detailed above. Option (b) has been implemented in Spanish to study massive diachronic variation (i.e. the evolution of Spanish over 500 years).

However, to adapt a POS tagger to a narrow chronological window (in this case, ~1532-1600 AD) would require a training set of words that is many orders of magnitude larger than those offered in available corpora of sixteenth-century Spanish texts. This concern may appear unfounded: would not the broad historical training set (e.g. twelfth- to sixteenth-century Spanish) of existing modified taggers improve annotation performance on texts from any given time slice within that interval? After enriching FreeLing—an open-source POS tagger—with an expanded dictionary of spelling variants, Sanchez-Marco et al. (2011) arrived at a decrease in accuracy of annotating Old Spanish compared to the gold standard of manual text annotation (termed “state of the art” by the authors). This curious result emerged from the architecture of POS taggers, which utilizes a probabilistic assignment algorithm to match word tokens with their most likely lemmas and parts of speech. In practice, an expanded dictionary requires the tagger to consider more potential word variants for any given word in context, raising the overall error probability; for example, words which evolved from lexical verbs in the twelfth century to auxiliary verbs by the sixteenth

164 Sanchez-Marco et al., “Annotation and Representation,” 2713.
Tag assignments are also influenced by choice of lemma: an incorrectly selected lemma will in turn raise the likelihood of an incorrect POS designation. The above-described dictionary ambiguity alone accounted for 90% of erroneous POS attributions in Sanchez-Marco et al.’s study of Old Spanish text.

Option (c) charts a middle path between options (a) and (b), capturing the benefits of manual annotation while utilizing a modern POS tagger. Specifically, I employed semi-automatic text annotation, applying a modern POS tagger to the modernized Spanish and manually checking the textual output. I also retained the historical word variants as metadata appended to the normalized text (discussed further below). While time-intensive, semi-automatic text annotation represents a gold-standard time improvement over fully manual text annotation, approaching the 95%+ accuracy rates of state-of-the-art POS taggers. Semi-automatic text annotation, as described above, has been applied successfully to a range of similar problems in historical linguistics in the last decade. A foregone benefit to adapting open-source taggers like FreeLing is that the improved tagger is available to the public for continued iteration and refinement. Given the relative infancy of automated spelling modernization studies in Old Spanish, larger training sets and the growth of digitalized Old Spanish corpora are eagerly anticipated for use in future studies of the Textos Andinos. The tagging process is described below.

I tokenized, lemmatized and POS-tagged the spelling-modernized khipu transcription corpus using the UDPipe repository, an R package which forms an Rcpp wrapper around the UDPipe C++ library. The package is language-agnostic, supporting tokenization (separation of text into its constituent word tokens), POS-tagging, lemmatization and additional functions to process raw text. The R wrapper is a 2018 release from the UDPipe development team at

---

166 Ibid., 8. The computational cost of this approach is formidable. The authors attribute assignment errors both to dictionary ambiguity and an insufficiently-sized training corpus. “Small” is relative: the training set consisted of 22.8 million word-tokens and the authors expanded FreeLing’s Spanish dictionary with over 32,000 word tokens.
168 See Baron and Rayson 2008; Ernst-Gerlach and Fuhr 2007; Fitzmaurice 2007; Pilz et al. 2008; Rayson et al. 2007; and Sánchez-Prieto Borja 2012. According to Sanchez-Marco et al., “Annotation and Representation,” 2714, this approach has also been adopted by the Penn Historical Corpora Project, the Corpus of Early English Correspondence (CEEC) and the York Toronto-Helsinki Parsed Corpus of Old English Prose.
Charles University, Prague, through the Institute of Formal and Applied Linguistics. Taking vectors of UTF-8-encoded raw text as input, the program can return R data frames with POS-tagged word tokens, employing Universal Part of Speech (UPOS) tags first developed by the Universal Dependencies Project—a cross-language text annotation initiative aiming to unify disparate word tag sets into a universal grammar.\footnote{See Nivre et al. 2016 for an introduction to the Universal Dependencies project. The text annotation grammar integrates prominent POS tag standards, including the Stanford dependencies and Google’s universal part of speech tags.} UDPipe was most recently updated for completion of a raw text-annotation task at CoNLL, SIGNLL’s (Special Interest Group on Natural Language Learning of the Association for Computational Linguistics) annual Conference on Computational Natural Language Learning and currently supports more than 50 languages.\footnote{Milan Straka, “UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task,” in Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies (Brussels, Belgium: Association for Computational Linguistics, 2018).}

The UD Pipe wrapper offers several advantages, as compared to other POS annotators and lemmatizers. The package does not rely on Python or Java (common language annotation platforms) and can be enacted directly from the R user interface. UD Pipe limits external R package dependencies to the tidyverse, minimizing exposure to version issues. The UPOS tagger (where “U” is used to denote “universal” from here on) and lemmatizer employ independent morphological guessers, solving the issue of aggregative lemma-POS errors evident in the FreeLing example above.\footnote{Milan Straka, Jan Hajič and Jana Straková, “UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing,” in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), eds. Nicoletta Calzolari et al. (Paris: European Language Resources Association (ELRA), 2016), 4293. Straka 2018 tested this claim for the 2018 UD Pipe 2.0 edition, comparing loosely- and tightly-joined annotation models (a proxy for separating the lemmatizer and the tagger). They found predictably that tightly joined (that is, where both lemma and POS are governed by the same algorithm) models performed worse at lemmatization; this is likely due to the aggregative error of interdependent tags and lemmas, discussed above in the case of FreeLing. In contrast, loosely joined models (i.e. UD Pipe) excelled at speech tagging and underperformed in dependency parsing (intuitive, given that two separate models will have greater difficulty communicating word embeddings between each other). This study does not consider dependency parsing, making use of UD Pipe net-beneficial. See also Milan Straka and Jana Straková, “Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UD Pipe,” in Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies (Vancouver, Canada: Association for Computational Linguistics, 2017), 91.} Lastly, the tagging output preserves compound verbs and punctuation within the data frame output, facilitating clean text reconstruction from tokenized input. The developers refer to the most recent version (UD Pipe 2.0) as a “single joint model,”
unifying CoNLL-U training data (UD Version 2) with a neural network-based UPOS tagger that leverages contextual word embeddings.¹⁷³

### 3.5 XML Conversion

Textual annotation yielded over 70,000 lemmatized and UPOS-tagged word tokens, constituting over 200,000 distinct points of textual data (i.e. tokens, lemmas and UPOS tags). However, this supplementary information serves little use as the basis for a computational linguistic study unless it is searchable in a structured way. Text markup languages offer a computer-legible grammar for achieving this goal, including a standardized documentation for textual annotations that groups—while distinguishing between—the original text and its accompanying attributes (UPOS, lemma, etc.). Armed with both word tokens and their descriptive metadata, the digital compiler faces an impactful decision in the life of the new corpus: should linguistic markup be stored adjacent to the original text (internally) or separately in a linked metadata document (externally)?¹⁷⁴ While a concern with internal representation is the resulting clutter of contiguous words, lemmas and part of speech tags, it is emphasized that markup-text is not intended to replace raw text as the primary vehicle for reading khipu transcriptions—Pärssinen and Kiviharju demonstrated the value of publishing the original texts as-is, scribal notations and all. Rather, markup languages impose a hierarchical structure on the text—embedding attributes within words, and words within documents—facilitating targeted searches of text under user-defined criteria.

I constructed an internally-represented Extensible Markup Language (XML) version of the khipu transcriptions. This textual compilation choice was inspired by recent efforts, also involving diachronic Spanish corpora, to standardize encodings in a mutually-intelligible way that (a) facilitates future study of digitalized khipu transcriptions and (b) allows for subsequent expansion of the corpus by researchers without extensive training in linguistic annotation.¹⁷⁵ Internal representations of metadata are especially impactful in the case of historical documents,

---

¹⁷³ Straka, “UDPipe 2.0 Prototype,” 197-199. See also Straka and Jana Straková, “Tokenizing, POS Tagging, Lemmatizing and Parsing,” 88 regarding the training set.


as historical and modern word forms can be kept in direct association with each other. In contrast to external representations, which use a system of nested files for preserving different levels of hierarchy, internal XML markup allows for simultaneous, in-line action on multiple levels of the document under study. This flattened structure is especially useful in preventing “splash effects”—human errors resulting from forgetting to correct a single misplaced cedilla or accent mark in all relevant files of an externally-represented hierarchy. 176 Internally-represented XML facilitates faster editing, minimizes splash effects and captures a large set of linguistic attributes in a standardized and repeatable form.

I created 72 separate XML files (one for each memoria), consisting of a single <text> header which encloses within it a number of <word> tags. Each <word> (corresponding to a single word or punctuation mark in the original text) contains several attributes, shown in the example below for the word “piezas” [“pieces”]:

- `<w lemma="pieza" pos="NOUN" token_orig="piéças" doc_id="52">piezas</w>`

While potentially intimidating in appearance, word tags follow strict structural rules. Each line of each XML document records one word tag (initiated with “w”), each of which contains four attributes:

- lemma: this attribute lists the lemma (singular: pieza) of the modernized word token (piezas).
- pos: the UPOS tag is listed: “piezas” is a noun.
- token_orig: the historical (original) form of the word token is “piéças.” That is, while “piezas” is listed in the word tag, the word appears originally in the transcription as “piéças.” The token_orig attribute lists the original historical form and is independent of the lemma (i.e. if the actual word token is plural, then its modernized form will also be plural). In cases where a given word does not require modernization, token_orig is equivalent to the value of the word tag.
- doc_id: the indexed number of the memoria from which the word is chosen (see appendix B). This identification allows all 70,000+ word tokens to be combined and studied in

---

176 Sánchez-Marco, Fontana and Boleda, “Propuesta de codificación.”
aggregate, since each word tag carries with it its own source (i.e. the document from which the word is taken).

In prose, the above word tag would be read: piezas, a noun, is the next word in transcription 52. It appears originally as pieças in the transcription and its lemma is pieza. A larger excerpt from the XML-markup corpus is shown in figure 3.2.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<text>
  <w lemma="a" pos="ADP" token_orig="A" doc_id="52">A</w>
  <w lemma="el" pos="DET" token_orig="el" doc_id="52">el</w>
  <w lemma="dicho" pos="ADJ" token_orig="dicho" doc_id="52">dicho</w>
  <w lemma="Vrtum" pos="PROPN" token_orig="Vrtum" doc_id="52">Vrtum</w>
  <w lemma="Sánchez" pos="PROPN" token_orig="Sánchez" doc_id="52">Sánchez</w>
  <w lemma="él" pos="PRON" token_orig="le" doc_id="52">le</w>
  <w lemma="dar" pos="VERB" token_orig="dieron" doc_id="52">dieron</w>
  ... 
  <w lemma="doscientos" pos="NUM" token_orig="doscientas" doc_id="52">doscientas</w>
  <w lemma="veinte" pos="NUM" token_orig="veinte" doc_id="52">veinte</w>
  <w lemma="piezas" pos="NOUN" token_orig="piezas" doc_id="52">piezas</w>
  <w lemma="de" pos="ADP" token_orig="de" doc_id="52">de</w>
  <w lemma="ropa" pos="NOUN" token_orig="ropa" doc_id="52">ropa</w>
  <w lemma="de" pos="ADP" token_orig="de" doc_id="52">de</w>
  <w lemma="abasca" pos="NOUN" token_orig="abasca" doc_id="52">abasca</w>
  <w lemma="y" pos="CCONJ" token_orig="y" doc_id="52">y</w>
  <w lemma="trescientos" pos="NUM" token_orig="trescientas" doc_id="52">trescientos</w>
  ... 
  <w lemma="cinco" pos="CCONJ" token_orig="cinco" doc_id="52">cinco</w>
  <w lemma="fanegas" pos="NOUN" token_orig="fanegas" doc_id="52">fanegas</w>
  <w lemma="de" pos="ADP" token_orig="de" doc_id="52">de</w>
  <w lemma="chuño" pos="NOUN" token_orig="chuño" doc_id="52">chuño</w>
  <w lemma="poner" pos="VERB" token_orig="puto" doc_id="52">puesto</w>
  <w lemma="en" pos="ADP" token_orig="en" doc_id="52">en</w>
  <w lemma="." pos="PUNCT" token_orig="." doc_id="52">.</w>
  <w lemma="Potosí" pos="PROPN" token_orig="Potosi" doc_id="52">Potosí</w>
</text>
```

Figure 3.2: Excerpt of Annotated XML Transcription Corpus

The XML khipu transcription corpus facilitated analysis with R of systematically-organized and enriched text.

---

177 Pärssinen and Kiviiharju, *Textos andinos*, vol. 2, 261: lines 2-7 (memoria 52).
Chapter 4

A Multivariate Study of the Textos Andinos

4.1 Multivariate Motivation

The compiled khipu transcription corpus may appear daunting due to its diversity of content. Readers who brave the two volumes are confronted with an overabundance of genres, village names and titles of long-dead Spanish administrators. How ought we approach a primordial division of this textual universe into usable categories? Moments like these invite Urton and Brezine’s reflection on studying archaeological khipus: “perhaps no challenges are more fundamental and daunting than those concerning typology.”\(^\text{178}\) Indeed, these same moments reinforce the familiarity of exegetical typologies—such as the earlier-proposed “primary,” “elicited” and “summary” categories—which employ close reading techniques to intuit which categories best characterize a corpus of texts.

While typologies are daunting, they are at the same time fundamental. It is argued herein that statistical typologies offer several advantages to their exegetical counterparts, while at the same time enriching close reading techniques.

- Statistical techniques better uncover aggregate, often unconscious, patterns in language. The present study focuses on unpeeling the narrative infrastructure that facilitated khipu readings—that is, khipu transcriptions as enablers of content rather than the content themselves. I abstracted the question of the content of any given list—15 loads of chuño; 22 llamas, etc.—to answer what foundational narrative conventions made such enumerated lists possible. This is surely an uncomfortable jump for students of khipu transcriptions. We are, after all, conditioned naturally to focus on the content of what we read: chuño, llamas and all. The techniques employed here combat this human “content imperative” by reducing complex khipu narrations to their most basic building blocks: (1) the choice of native khipukamayuqs to use broad categories, whether nouns or verbs, adjectives or numbers; and (2) the unthought word tokens (“and” or “the,” for example) that serve as interstitial linguistic bridges, gluing narrative content together. These measures serve as proxies for narrative infrastructure: the conscious choice of which

\(^{178}\) Urton and Brezine, “Khipu Typologies,” 319.
word category to use, and the (often unconscious) choice of how a subsequent word category should be joined with its neighbor.

- Statistical techniques facilitate searches for significant text categorizations, which together compose what I label statistical typologies. In contrast to close reading, corpus linguistics techniques allow the researcher to quantify their uncertainty, mediated through familiar measures such as p-values and confidence intervals, to confirm or refute stated hypotheses—the same types of hypotheses that often constitute the endpoint of close reading. Similarly, statistical measures allow for testing of previous categorizations. Are textual families A and B, previously proposed by researcher C, truly (and significantly) distinct from each other? Is the exegetical typology offered above a worthwhile motivator for future studies of khipu transcriptions? A statistical typology offers a novel reasoning by which to answer these questions.

- Statistical techniques expand our exploratory potential. Corpus linguistics offers insights into underlying linguistic patterns that cut across thousands of lines of text—tendencies on a scale too dispersed for observation by human readers. It is shown that while a statistical typology does not always match our expectations, there is value in hearing the document speak for itself, in a vacuum, without any pre-conceived notions of content or form. In this sense, aggregate textual analysis with digital methods gives ultimate agency to the text of khipu narrations, rather than our confidence in reading such text.

It is important to note that the proposed methodology does not aim to replace the close-reading techniques employed for decades by students of khipu transcriptions. Rather, I proffer the statistical study of khipu transcriptions, with a case study attesting to its explanatory power, as an additional strategy to enrich historical interpretation and evaluate previous typologies. An explicitly synergistic approach is employed, following Zuidema and de Boer’s principles of model parallelization: the results of multiple statistical implementations are compared and aggregated to approach external validation. Desagulier’s view is dutifully echoed: “[the] corpus is a central part of research, but not the alpha and omega of linguistics.” An interpretive

180 Desagulier, Corpus Linguistics and Statistics, 10.
openness is encouraged during these applications: my own “‘early time’ approaches” to the aggregate study of khipu transcriptions.181

4.2 Linguistic Feature Choice

The opening move of khipu transcription clustering was to represent each text excerpt by a multivariate linguistic characteristic. That is, what 10-15 variables might be used to “measure” a transcription? This study, following established corpus linguistics methods, represents each memoria by two measures, also called linguistic features:182

- The respective percentages of 13 Universal Part of Speech (UPOS) tag assignments within a given transcription. For example, in Cobo’s transcription of the ceque system of Cuzco (memoria 3), nouns make up 19.98% of total words; verbs constitute 16.64%, etc. This percentage breakdown was repeated separately for all 72 excerpts (that is, memorias were recorded independently).

- The percentage of the 15 most frequent word lemmas across the entire corpus. That is, what percentage of the words within each transcription are words with any of the following lemmas: y, el, de, él, que, a, en, dar, uno, indio, ser, con, otro, pueblo, dicho. For example, in an enumerated list of Inkaic tributary subjects in Chucuito (memoria 23), various forms of indio [“Indian”] (indios, indias, etc.) comprise 5.3% of words; forms of ser [“to be”] (fue, fueron, etc.) comprise less than 1% of words. These percentages were recorded separately for the 72 excerpts.

It is emphasized that these two measures—UPOS-tag composition and relative lemma frequency—were prepared in isolation from each other, with the aim of testing whether the results derived from each feature aligned with each other.

Beyond their recognition as mainstream linguistic features for text analysis, parts of speech and frequent lemmas were chosen because they are relatively content-agnostic. That is, measures of nouns and adjectives or the frequency of “indio” do not reflect the researcher’s preconceived notions of what comprises meaningful contextual information in the memorias. Instead, the corpus speaks for itself—I recorded (1) the 15 most frequent lemmas and (2) constituent parts of speech, letting these most basic numerical measures describe their parent. Put

181 Urton and Brezine, “Khipu Typologies,” 349.
182 Savoy, “Authorship Attribution,” 133.
more strongly, there is a remarkable meaningfulness to latent *contextual* patterns identified within the transcriptions, when those patterns emerge from solely *content-agnostic* measures. Might the most basic building blocks of khipu narration—interstitial conjunctive glue and elemental part of speech composition—yield a meaningful statistical typology? Linguistic features are active choices made by the researcher in approaching this question.

An additional consideration is the coarseness of the chosen features. For instance, the UPOS tag set employed by UDPipe consists of only 17 categories; examples include VERB, AUX (auxiliary), CCONJ (coordinating conjunctions) and PART (participle). Why rely on a relatively small number of tags, when other language-independent speech-parsing programs use upwards of 50, often allowing for differentiation of verbs into over 20 independent categories? This is a context-dependent question: in the present study, the breadth of UPOS tags is their greatest strength for two reasons. The first is mathematical. Given the aim of minimizing tagging error, utilizing broad UPOS categories improved the confidence of any given assignment and simplified the manual checking component of semi-automatic text annotation. The second is at the heart of our attempt to form khipu transcription typologies. Excessive granularity risks conflating meaningful khipu narration and Spanish scribal variation. The present study is concerned with analyzing the infrastructure of khipu vocalization, and less sixteenth-century scribal idiosyncrasies. In turn, the relative percentage of different word forms is considered more heavily than their specific forms. Recall that khipu narrations were, after all, organic events, often filled with verbal disagreement and negotiation. In light of this picture, it is postulated that we ought be concerned with the *choice*, rather than the *form* of particular words—to evaluate “dieron al encomendero” [“they gave the encomendero”] and “habían dado al encomendero” [“they had given the encomendero”] together as equals, reflecting a subject (they), verb (gave) and direct object (encomendero) potentially recorded in khipu cords. Given the diversity of orthographic and scribal inconsistencies already described, I minimized the influence of *consistent inconsistency* and aimed to capture high-level differences in transcription composition

---

183 Recall that I considered 13 UPOS tags in the introduction, while there are 17 total tags in the Universal Dependencies tag set. This is because four classes (interjections, punctuation, symbols and unidentifiable text) either do not appear in the corpus or are excluded (i.e. punctuation) intentionally.
across time and space. This approach risks flattening a potential khipu verb tense infrastructure. However, this risk is temporarily accepted to prevent a more concerning outcome: specious conclusions drawn from excessively granular data. In sum, it is the belief of the author that exploratory analysis should begin broad, and narrow its focus once informed by ongoing results.

This belief also influenced the choice of frequent lemmas as the second linguistic feature. It is first acknowledged that my measure differs slightly from a “stop words” methodology, in which non-content function words (e.g. the, an, for) are counted and used as proxies for authorial form—often employed in quantitative stylometry studies.\(^\text{186}\) Several function words appear in the vector of frequent lemmas: “a”, “de” and “que” comprising just a few examples. However, I also included “indio” and “pueblo,” two contextually meaningful word tokens. The hybrid approach captures the benefits of stop-words-style studies while maintaining historical context and maintains robust support.\(^\text{187}\) Modern Spanish stop words are not easily mapped onto sixteenth-century Old Spanish. That is, available lists of frequent (content-agnostic) word tokens reflect corpora of modern writing. This approach risks straining a sixteenth-century narrative through a twenty-first-century sieve. While my study aligns with the spirit of measuring function words, noting the power of common word tokens, I allow the corpus to speak for itself, including words like “indio” and “pueblo” because they appear in almost every transcription. Potential contextual bias is, in turn, partially mitigated by the ubiquity of such words in the corpus. It is also stressed that the present study does not purport to solve an authorship attribution problem—the situation in which a strict, stop-words-only methodology is often used. While similar techniques are employed, my work does not aim to prove (or disprove) the identity of a given khipu narrator. Instead, I utilized broad—even coarse—document features to intuit overarching themes that may prove helpful in our quest to construct khipu transcription typologies.

4.3 Supplemental Khipu Transcription Variables

In addition to processing the transcriptions, equally important was compiling accurate metadata about each memoria that can be linked to constituent word tokens in a flexible and


I compiled, in addition to the transcriptions, a metadata table (see appendix B) that consists of 72 observations of 13 illustrative variables. The variables consist of five quantitative measures and eight qualitative variables. The five quantitative measures are:

- **year**: the year to which a given khipu transcription pertains. It is noted that transcriptions often postdate the data of their narration by decades. For example: several Sakaka tribute restitution trial testimonies (memorias 14-16) occurred in 1571, over 20 years after the end of the alleged overpayments (1548-1551). To normalize the metadata, I assigned the earliest date by which a khipu could have plausibly contained the narrated data. In the Sakaka case, 1551 was used as the year label.

- **latitude**: the decimal degree latitudinal coordinate of a given transcription.

- **longitude**: the decimal degree longitudinal coordinate of a given transcription. I assigned geographic metadata based on the location to which a transcription refers, rather than the location where such a khipu was narrated. Several khipu transcriptions in the corpus describe khipukamayuqs summoned to testimony from disparate parts of the empire. As such, I focused on the narrative infrastructure that accompanied such khipus on the (often-long) trek from their places of origin—to test, for example, whether khipus from former Aymara territory were vocalized significantly differently than khipus from Cuzco, even if both narrated in the same location. In the case of encomienda titles, which often list dozens of villages separated by tens of kilometers, I approximated the geographic center point of the named villages as a proxy for the location of that khipu transcription.

- **dist_cuzco**: the distance (in km) between the given transcription’s latitude/longitude pair and the Coricancha of Cuzco, a temple which formed the ritual center of the Inka Empire (13.5202° S, 71.9752° W). This variable aims to capture the effects of (former) Inka administrative influence: were narrations of cords knotted far from the former Inka capital significantly different and more diverse, in their essential structure, than those in the heart of Cuzco?

---

188 Sanchez-Marco et al., “Annotation and Representation,” 2717 offers further discussion of the costs and benefits of maintaining metadata separated from the historical texts they describe.
189 See memoria 4, for example, in which khipukamayuqs from the provinces of the Chupayuchu ethnic group (much of the modern Huánuco province) were summoned to Conchumayo, the regional seat of power, to read their khipus as legal testimony. Recording the location of the narrated khipus as Conchumayo (a proxy for the valley; approximately 9.94°S, 76.2°W) introduced geographic diversity into the metadata of the resulting transcriptions.
• **altitude**: the altitude (in meters) corresponding to the stated decimal GPS coordinates. I calculated altitude values using Leaflet’s OpenStreetMap application, which is an open-platform JavaScript library. With this variable, I test whether altitudinal zones form a significant category by which to distinguish khipu transcriptions. This altitudinal variety constitutes an added benefit to studying the *Textos Andinos*: we can “observe” khipus from the Andean highlands. With the exception of 32 samples from Laguna de los Cóndores, not a single archaeological khipu sample has been reliably attributed to the highlands, and in turn, Cuzco, the former Inka capital; the organic cords in these areas have long-since disintegrated in the wetter climates. While this reality has frustrated many a decade of khipu scholar—the Spanish chronicles tell us that thousands of khipus were stored in massive historical archives, detailing the histories of the early empire—I analyzed, gratefully, the small family of surviving highland khipu transcriptions, likely the last surviving voices of these preconquest cords.

The eight categorical variables consist of the following:

• **type**: the assignments corresponding to my proposed exegetical typology. This variable has four levels (also called values): primary narrative (“pnar”), primary numerical (“pnum”), elicited (“elicit”) and summary (“summ”). The assignments were used to test the robustness of my proposed categories. For example, are elicited khipu transcriptions (recall the Sakaka trial example) significantly different, as measured by the chosen linguistic features, than primary numerical transcriptions? With this variable, I quantified the statistical validity of the exegetical typology.

• **pre_post_conquest**: whether the data recorded in the narrated khipu(s) pertain to before or after 1532. The variable has two levels: “pre” (preconquest) and “post” (postconquest). I considered early encomienda titles as postconquest compilations of preconquest cord data, and so labeled these transcriptions as “preconquest” in the compiled metadata.

---

• **language**: whether a khipu was narrated in Quechua or Aymara. To maintain high-level categorizations, I used as a proxy for Aymara all locations south of Lake Titicaca. The Aymara constituted a collection of pre-Inka states that occupied lands crossing northern Chile, most of Bolivia and southern Peru. The Spanish chronicles tell us that khipus (called “chinu” in this region) were narrated by chinukamayuqs in Spanish trial proceedings; the language variable, whose two levels are “quechua” and “aymara,” tests whether variations in transcription documents are (significantly) attributable to their source language.¹⁹³

• **int_unint**: whether a transcription suggests an (un)interrupted narration of khipu data. This assignment was often inferred: the transcriptions do not always state whether khipu narrations were in response to a query. In cases of ambiguity, I used transcription length as a proxy for non-interruption. That is, I considered a several-dozen-line narration of people and places, or numbers of things, to be an uninterrupted account. This variable was deployed to capture the effect of Spanish administrative/legal proceedings on khipu narrations. In other words, did call-and-response khipu narration—a phenomenon previously unknown to Inka khipukamayuqs and introduced only after the Spanish conquest—alter the essential structure of khipu readings? This statement was tested on the full transcription corpus. The variable has two levels: interrupted (“int”) and uninterrupted (“unint”).

• **pop_s_m_l**: the relative size of the population detailed in or corresponding to the transcription. I tested whether the relative population size of a given area is associated with significantly different vocalizations in the transcription corpus. This variable has three levels: small (“s”), medium (“m”) and large (“l”). It is emphasized that population was considered as a relative measure, given the challenges of assigning stable population figures to every region of the former Inka Empire at the precise time of its khipus’ construction.¹⁹⁴ In cases of non-census transcriptions—the enumeration of camelids

---

¹⁹³ The most-often cited example of *chinu* readings is the collection of documents produced during the Sakaka tribute restitution trial. See Platt 2002; Urton 1998 for the most thorough treatments of these court proceedings.

¹⁹⁴ See Noble David Cook, *Demographic Collapse, Indian Peru, 1520-1620*, Cambridge Latin American Studies, no. 41 (Cambridge, United Kingdom: Cambridge University Press, 1981) for high-level figures charting Andean population decline in the early colonial period. These measures, while valuable to studies of broad population dynamics, do not capture the spatial and temporally-situated contexts of each memoria. While I could use more granular population measures (e.g. the Toledan census), its figures, while locked in the same time period, also vary
around Lake Titicaca (memoria 27) or Topa Inka Yupanqui’s military conquests (memoria 2), for example—I assigned a population size consistent with (1) other memorias that reference the same town/place; and (2) the numerical magnitudes of tribute and counts of other items referenced in the transcription. Among encomienda titles, which often reference dozens of different locales, I used the survey total as a proxy for regional population.

- **num_speakers_sng_mult**: whether the transcription contains the statement of a single speaker, or that of multiple simultaneously contributing speakers. This variable, in tandem with (non-)interruption, was deployed to test whether khipu narrations differ significantly when varying the conditions of their performance. That is, I tested whether interruption/non-interruption or the number of contributing speakers is a significant category for differentiating khipu transcriptions. The variable has three levels: a single speaker (“sng”); multiple speakers (“mult”); and not applicable (“na”). I labeled as not applicable khipu transcriptions consisting of broad compilations of local khipu accounts, most often in the case of encomienda titles. For instance, it is not obvious whether a “summary” encomienda title, to use the language of the exegetical typology, should be analyzed as a single enumerated list or as a concatenation of individual testimonies from dozens of villages. Even if multiple khipukamayuqs contributed to a given village’s khipu narration, these details are often omitted from early colonial summary transcriptions. Even in a scenario of full information, difficult assignments remain: should a summary transcription citing an aggregated mix of single- and multiple-speaker narration events be categorized as a single- or multiple-speaker document? These situations were deemed “not applicable” to our measure of speaker count, avoiding the graver risk of misclassification under incomplete information.

- **currency**: whether Spanish currency is referenced within the khipu transcription. Given the substantial cultural change endured in the Andes following the introduction of

---

195 There are of course exceptions to this statement. Take, for example, Curatola Petrocchi and de la Puente Luna, “Contar concertando,” 226. The commentary concerns a khipu transcription from Lucanas, narrated in 1581 legal testimony concerning tribute restitution: “from the 'quipu y cuenta' did not emerge the reading of a captured immutable message, but instead the oral enunciation, by way of multiple voices, of a message inseparable from its own process of construction and actualization” (my translation; emphasis added).
monetary principles and values—including documented instances of “currency khipus”—
money was tested as a potential contributor to khipu narration variability. The variable
is binary: its levels are “yes” and “no.”

- **lat_zone**: the relative latitudinal zone of the khipu transcription. In order to consider
  narrative diversity in the extreme north and south of the former Inka Empire, I imposed a
  tripartite subdivision on the transcription locations. The three levels of the variable are
  “south” (applied to transcriptions from south of Lake Titicaca); “north” (transcriptions
  attributed to north of -11.0° in decimal latitude) and “central” (the remaining regions of
  the former empire). This variable considers to what extent longitude, latitude and distance
  from Cusco in combination may yield a significant marker for explaining khipu
  transcription diversity. In other words: to what extent did varying khipu narratives
  emerge from the extremes of former Inka territory, precisely the regions subject to less
direct administrative oversight than their counterparts in and near Cusco?

The five quantitative and eight categorical variables detailed above comprise the metadata
analyzed in the present study. It is also stressed that these data are a subset of the compiled
transcription descriptions (see appendix B). Omitted variables include title, author and
beginning and ending lines (all three using Pärssinen and Kivihaaru’s notation). These
omissions reflect the aims of my study. Author designations are especially challenging: is the
author of a given transcription the khipukamayuq who narrated the content of his khipu, the
lengua who translated the Quechua testimony into Spanish, or the scribe who entered the
speech into writing? Transcriptions often include the name of only one—sometimes none—
of the participants in the original khipu narration event. Document titles, which are imposed
by Pärssinen and Kivihaaru, do not answer meaningful hypotheses. Similar reasoning holds in
the case of starting and ending lines, although these endpoints define transcription length,
which is potentially meaningful. I normalized document feature measures by using relative
percentages, correcting for the length of the different transcriptions. Nevertheless,

---

196 Curatola Petrocchi and de la Puente Luna, “Contar concertando,” 208, 208n19; Medelius, “El descargo en las
cuentas,” 246 offers a case study in Huamanga.

Approaches, eds. Aquilino Sánchez Pérez and Moisés Almela Sánchez, Studien zur romanischen
Sprachwissenschaft und interkulturellen Kommunikation (Frankfurt: Peter Lang, 2010), 272 offers an introduction
to the “relative frequency ratio” metric.
transcription lengths are represented in the summary statistics (see table 3.1). I hoped to approach meaningful contextual conclusions from explicitly non-contextual measures, considering 72 observations of thirteen illustrative metadata variables. The final illustrative variable set was joined to a matrix of relative feature measures and subjected to automatic textual clustering, described in detail below—the genesis of a statistical khipu transcription typology.

4.4 Introduction to Principal Component Analysis

Two tables of compiled data were studied. While both consist of 72 rows (individuals), each of which corresponding to a memoria, the tables differ in the linguistic feature used to summarize the transcriptions. The “lemma” table has 15 active variables corresponding to the percentage of words in each transcription (row) that possess one of the frequent lemmas. Added to the 13 illustrative variables described above, this table has 72 rows and 28 columns (= 15 + 13). The “UPOS” table has 13 active variables corresponding to the percentage distribution of part of speech within each transcription. Added to the 13 illustrative variables, this table has 72 rows and 26 columns (= 13 + 13). The lemma and UPOS tables share the same set of 13 supplemental/illustrative variables. However, the documents were clustered based on the active variables—the relative similarity/dissimilarity of their linguistic feature sets. The clusterings were performed independently on both tables, and the results compared—if both content-agnostic features yield concordant conclusions, we can be more confident in the robustness of our findings.

The quantitative study had three aims in practice: (1) grouping together similar memorias (individuals); (2) grouping together correlated lemma or UPOS profiles (variables); and (3) exploring the relationship between memorias, linguistic features and contextual metadata (altitude, currency, etc.). In practice, this was accomplished by subjecting both the lemma and UPOS tables to principal component analysis (PCA). PCA is a type of multivariate exploratory data analysis which attempts to faithfully represent a table of individuals and variables in a space with fewer dimensions—that is, a visual representation in which the distances between individuals are distorted as little as possible. The analysis is labeled multivariate due to the

---

198 Desagulier, Corpus Linguistics and Statistics, 243.
number of variables. The lemma table, for example, can be considered as 72 points in $\mathbb{R}^{15}$—a
15-dimensional vector space. Each row of the table implicitly defines a “document space,”
whereby the relative frequencies of 15 lemmas (de, el, etc.) define a 15-place
Dimensionality reduction decomposes the cloud into orthogonal planes of ranked importance,
where each plane offers a “flattened” view of multiple dimensions—15 dimensions projected
onto two, for example. If the document spaces/vectors defined by two transcriptions have similar
values in many of their components, we expect that the two transcriptions will also appear close
to each other on a Euclidean plane, which constitutes a faithful representation of the multivariate
document cloud and its “underlying statistics.” If performed properly, minimal information is
lost while reducing the memoria cloud—a high-dimensional space which is not visualizable by
humans—into a set of more manageable memoria planes. The number of such planes studied is
an active choice which must be defended by the researcher.

Specifically, the process begins by using the linguistic feature frequencies (lemma or
UPOS) to construct a square, pairwise correlation matrix. In this matrix, every entry corresponds
to the correlation coefficient between each pair of frequent lemmas or UPOS tags, which
themselves are vectors with 72 entries (for example, the relative frequency of noun tags across
the 72 transcriptions). The matrix is both square and symmetric: the correlation coefficient
between “de” and “el” (in the lemma table, for example) is equal to the correlation coefficient
between “el” and “de,” while the entries along the diagonal correspond to the overall variance of
each variable across the corpus. The collection of pairwise correlation coefficients forms the
correlation matrix, whose eigenvectors (as numerous as the row/column count) are called
principal components (PCs). PCs from normalized data are orthonormal (orthogonal and unit-

---

201 Ashish Kumar and Avinash Paul, *Mastering Text Mining with R* (Birmingham, United Kingdom: Packt
Publishing, 2016), 100. Faithful representations of the document cloud are also discussed in R. Harald Baayen,
*Analyzing Linguistic Data: A Practical Introduction to Statistics Using R* (Cambridge, United Kingdom: Cambridge
University Press, 2008), 120.
203 Oakes, *Literary Detective Work*, 39 offers an expanded, informal explanation of the implementation of PCA.
204 Recall that by the spectral theorem, the square symmetric correlation matrix $A$ can be diagonalized as
$B = S^{-1}AS$, where $S$ contains a new orthogonal coordinate system defined by the eigenvectors. The corresponding
eigenvalues in the orthogonal coordinate system appear in the diagonal of $A$, the original correlation matrix. Since
the correlation between off-diagonal entries of $A$ is now zero ($A$ is diagonal in the new coordinate system), the
variance of any given projection of the data on the PCs is maximized by choosing the largest corresponding
eigenvalue. The document features share a common scale (percentage of total words) and thus have already been
length) basis vectors which span the newly created space and are projected in the directions of highest variability.\textsuperscript{205} PCA supposes that the original data can be represented by the orthonormal eigenvectors, instead of the original axes.\textsuperscript{206} The set of PCs can in turn be used to iteratively reconstruct the original data in a lower-dimensional space. Crucially, each PC is a linear combination of the active variables, projected in a new space in which the direction of each PC represents a ranked list of decreasing variability.\textsuperscript{207} The eigenvalues thus represent the percentage of variance in the data explained by the corresponding eigenvector.

In short, PCA achieves a change of basis while storing the cumulative percentage of explained variance in an orthogonal coordinate system. The result is a methodology for producing easily interpretable visualizations in which we can see relationships between memorias and their features. That is, to gauge whether there are “systematic differences” in the narrative features of khipu transcriptions, as a function of time, region, currency or altitude, among other variables.\textsuperscript{208} Study of these underlying contributors to transcription variation allows us to construct a “typology of the individuals”; in this case, the individuals are called memorias.\textsuperscript{209}

4.5 Principal Component Analysis Implementation

I implemented PCA using FactoMineR, an R package enclosing a library of functions for multivariate exploratory data analysis.\textsuperscript{210} The package is extensively documented, incorporating simple function arguments and integration with a range of graphing extensions.\textsuperscript{211} I subjected the lemma and UPOS tables to PCA, choosing the feature measure columns as active variables and the thirteen metadata columns as supplementary/illustrative variables. While the latter do not contribute to the construction of the PCs, supplementary levels can be projected on the lower-
dimensional space to aid with visual interpretation. I separate the analyses by lemma and part of speech below.

4.5.1 Lemma Implementation

The appropriateness of PCA on the lemma data set was first gauged by three standard measures: Bartlett’s Test of Sphericity, the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy and the correlation matrix determinant.212

- Bartlett’s Test. PCA assumes a moderate level of correlation between the variables. Since the correlation matrix is reconstructed using the PCs, it is not sensible to perform PCA if most of the correlation matrix values are near-zero. This would imply that most of the variables are unrelated, foreclosing meaningful structure detection in the data. The null hypothesis ($H_0$) is that the correlation matrix resembles an identity matrix (in which observed nonzero correlations result from sampling error). $H_0$ is rejected ($\chi^2(105) = 445.01, p = 2.097e - 43$), and thus the test is significant: the observed correlation coefficients are not attributable to sampling error.

- KMO. Since correlation coefficients fluctuate between samples, the meaningfulness of PCA is dependent on the sample size—this is especially relevant in the case of endangered languages, in which a small universe of samples is probed using a relatively large set of variables. The KMO statistic, which varies between 0 and 1, quantifies the adequacy of sample size by measuring the proportion of variance among the factors which exceeds shared variance.213 High KMO values indicate the presence of distinct correlation patterns that may be meaningfully studied using PCA. The general KMO statistic value obtained is 0.7, which is an intermediate value. By Kaiser’s criterion, this value reflects an adequate sampling; by Hutcheson and Sofroniou’s criterion, the sampling is “good,” second only to “great” and “superb.”214 While the intermediate KMO value may be cause for concern, it is emphasized that the relatively small sample size is a function of the known khipu transcription universe (see section 3.1). Lacking the luxury of growing the corpus, we proceed.

212 See section C.2.2.1 for the implementation of Bartlett’s Test, KMO and the correlation matrix determinant.


Correlation matrix determinant. PCA’s assumption of a moderate level of correlation between variables presents a two-tailed problem. That is, correlations between variables may either be too low or too high. While Bartlett’s Test confronts the former condition, the correlation matrix determinant confronts the latter problem—that of variables correlating too highly, called multicollinearity. Following Field et al.’s heuristic, the determinant of the correlation matrix (0.00108) was confirmed to be larger than the reference value (0.00001), and thus multicollinearity is not observed in the lemma data.215

The diagonalization produced (1) a matrix of eigenvectors (the PCs) and (2) a set of eigenvalues, which constitute the percentage of total variance (also called inertia) explained by the corresponding PC. The latter is particularly important because it aids in determining the number of components (hence planes) to consider in the interpretation of output. Table 4.1 lists the cumulative percentage of inertia explained by the first five (of 15 total) factors. Figure 4.1, called a scree plot, offers a visualization of the table.

<table>
<thead>
<tr>
<th>Comp</th>
<th>Eigenvalue</th>
<th>Percentage of Variance</th>
<th>Cumulative Percentage of Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>comp 1</td>
<td>4.398777</td>
<td>29.32518</td>
<td>29.32518</td>
</tr>
<tr>
<td>comp 2</td>
<td>1.964053</td>
<td>13.09369</td>
<td>42.41887</td>
</tr>
<tr>
<td>comp 3</td>
<td>1.718270</td>
<td>11.45514</td>
<td>53.87400</td>
</tr>
<tr>
<td>comp 4</td>
<td>1.296895</td>
<td>8.645965</td>
<td>62.51997</td>
</tr>
<tr>
<td>comp 5</td>
<td>0.981588</td>
<td>6.543918</td>
<td>69.06389</td>
</tr>
</tbody>
</table>

Table 4.1: First Five Eigenvalues and Explained Variance (Lemma PCA Implementation)

---

215 Field, Miles and Field, Discovering Statistics, 771, 777.
The first four factors were chosen to reconstruct the data. This choice follows Kaiser’s criterion for factor retention, which advises selecting factors with eigenvalues larger than one—hence the shading of the first four rows of table 4.1. Kaiser’s criterion performs especially well when fewer than 30 variables are studied and when samples are relatively small. The criterion imposes a floor on the quality of representation: components with eigenvalues less than 1

---


represent strictly less data than any original variable of the table.\footnote{Husson, Lê and Pagès, \textit{Exploratory Multivariate Analysis}, 17; Natalia Levshina, \textit{How to do Linguistics with R: Data Exploration and Statistical Analysis} (Amsterdam and Philadelphia: John Benjamins Publishing Company, 2015), 355.} By the table, the first component explains 29.3\% of total variance; the second component 13.1\%, etc. Due to the orthogonality of the PCs, explained variance can be aggregated across successive components: the four retained components collectively explain 62.5\% of inertia in the original table.

This four-factor cutoff aligned with additional standards of factor retention. These include: Loewen and Gonulal’s criterion of selecting factors until 60\% of total variance is explained;\footnote{Loewen and Gonulal, “Exploratory Factor Analysis,” 194.} ceasing factor retention at the point of visible inflexion (“elbow”) of the scree plot (at component 4 in figure 4.1);\footnote{Field, Miles and Field, \textit{Discovering Statistics}, 781; Loewen and Gonulal, “Exploratory Factor Analysis,” 196.} and parallel analysis, in which only those eigenvalues are retained which exceed the 95\textsuperscript{th} percentile of eigenvalues computed from 1,000 random multinormal matrices of the same dimension as the original data.\footnote{Loewen and Gonulal, “Exploratory Factor Analysis,” 196-197. In this case, the parallel analysis suggests that three components should be retained (see figure A.2 in appendix A.1.1.2). Nevertheless, the prospective inclusion of the fourth component falls within one standard deviation of the mean simulated eigenvalue corresponding to PC4. Four PCs were retained, under the assumption that overrepresenting variance (by the inclusion of the fourth PC) is preferable to the converse. The fourth PC was also plotted along with PC3 in the second document plane and drew inevitable attention.}

The visual and numerical output of PCA was analyzed to study the relationship among individual transcriptions and among variables. Figure 4.2 plots the quantitative variables considered in the analysis, projected on the plane spanned by the first two PCs.
Figure 4.2: Variable Factor Map for Dimensions 1 and 2 (Lemma Implementation)

The variable plot offers a visual representation of transcription variation as a function of relative lemma frequency. Recall that the plane captures a faithful representation of the original 15 dimensions in two. This representation allows us to use distance in the plane as a proxy for closeness in the original space, since the first two factors (PCs) capture over 40% of the total inertia—that is, almost half of the total variation among khipu transcriptions is represented by the plane.\textsuperscript{222} The black variables are considered active, as opposed to the quantitative supplementary variables, which are plotted in blue (recall that the PCA is performed on only the active variables). The (a) lengths and (b) directions of the arrows correspond to (a) the relative magnitude of each variable’s contribution to the construction of the axes and (b) the spatial

\textsuperscript{222} Husson, Lè and Pagès, Exploratory Multivariate Analysis, 34.
projection of the variable on the plane. The variables are represented within a circle of unit radius, following the standardization imposed on the mean lemma ratios during preprocessing. Interpretation followed from a joint consideration of (1) shared direction between variables (signifying strong intercorrelation between those variables), (2) oppositely-facing variables (that is, strongly anticorrelated variables) and (3) relative magnitudes of contribution to the components.

The two most closely correlated groups of variables in the plane are “indio,” “pueblo” and “otro,” as well as “a,” “el” and “dicho.” Lacking additional context, it might be suggested that the first cluster of variables is characterized by lists (joined by the coordinating conjunction “otro”) of villages (“pueblo”) and the associated villagers (“indio(s”)”). The second group brings together key words in subject-object giving relationships—“a” and “el” combine to form “al,” which often introduces a direct object (the recipient of service, goods, tribute, etc.) described by “dicho” (e.g. taxes paid to the “said” ruler). These abstract relationships are contextualized by studying opposition in the first component (the x-axis). Along this axis, the component separates lemmas which center actors/subjects (the right-hand half-plane including the pronoun “él” and nouns/descriptors like “pueblo(s),” “indio(s),” “que,” etc.) from those lemmas that center objects acted upon/direct objects (the left-hand half-plane including verbs such as “dar” and articles such as “a,” “el,” and “dicho”).

Along the second component (the y-axis), coordinating conjunctions (“y” [“and”]; the lower half-plane) are distinguished most strongly from generative lemmas like “ser” [“to be”] and “dicho” (which modifies an accompanying noun) in the upper half-plane. This axis might be tentatively labeled as differentiating between flatter (lower half-plane) and more complex (upper half-plane) speech. The only quantitative supplementary variable which is significantly ($p < 0.001$) correlated with either component is “year,” which is associated with negative components of PC1: memorias which have large negative values on PC1 constitute later transcriptions (chronologically).

Dimensions 3 and 4 also merit consideration. The numerical variables are plotted in figure 4.3.
Figure 4.3: Variable Factor Map for Dimensions 3 and 4 (Lemma Implementation)

This second plane (spanned by PC3 and PC4) adds supplementary information to that described in the first. Along the third dimension, the subject-object duality observed in the first plane is further supported: moderate positive values for description of subjects (“indio,” “el,” “pueblo”) oppose negative values of direct objects (e.g. those implicitly defined by “dar” [“to give”], which is significantly anticorrelated with PC1; $p < 0.001$). Dimension 4 reinforces the simple/rich text description binary proposed on the first plane. Twelve of 15 frequent lemmas are best associated with positive values of PC4 (and hence individuals that fall in the upper half-plane); among frequent lemmas, a greater diversity of lemmas is frequent among transcriptions with positive values in dimension 4. A preliminary variable clustering, considering all four dimensions, might be: subject-centered generative lemmas (top-right quadrant); subject-centered flat lemmas
Preliminary variable clusterings were also informed by the plot of individuals and associated illustrative variables. As described by Husson, PCA produces a “joint representation,” where “differences between individuals can be explained by the variables, and relationships between variables can be illustrated by individuals.” Figure 4.4 represents the 72 transcriptions as points, sized and colored corresponding to their respective contributions to the PCs. Figure 4.5 plots the eight qualitative categorical variables described in the previous section. The ellipses drawn in figure 4.5 represent 99% confidence ellipses—that is, two non-overlapping ellipses are said be significantly different categories for describing khipu transcription variation at the 1% level.

Two observations are evident from the plot of individuals. The first is that the memorias do not contribute equally to the construction of the axes. In other words, some khipu transcriptions contribute disproportionately to total variance. Examples include memoria 17 (an excerpt from the Sakaka tribute restitution trial); memoria 25 (an enumerated list of tribute items from Chucuito paid to the Spanish in 1567); and memoria 33 (a 1549 encomienda title allocated by Pedro de la Gasca to Hernán Bueno). Recall that the large contributions of these items are not necessarily due to the content of their narrations—since a list of content-agnostic lemmas was used to perform the clustering. Rather, the broader narrative structure of these excerpts is exceptional—their relative mean frequencies are out-of-line with the baselines established throughout the corpus. The second observation is that certain transcriptions natively cluster together. Some of these clusters follow intuitively from the corpus: memorias 31 and 33, for example, are two transcriptions of the same encomienda title in Carumas (allocated to Hernán Bueno), completed seven years apart from each other. Other groupings offer more insight: memorias 38 and 44 are two encomienda titles from two different regions separated by almost 500 kilometers (Tarapaca and Ubinas). Might the shared characteristic among these two transcriptions—their status as encomienda titles—significantly impact their narrative structure within the corpus?

---

Figure 4.4: Contributions of Individuals (Lemma Implementation)
Figure 4.5: Categorical Supplementary Variables with 99% Confidence Ellipses (Lemma Implementation)
Figure 4.5 allows us to test these hypotheses. The plotted ellipses offer a visualization of whether the levels of each variable (also called categories) differ significantly (i.e. \( p < 0.01 \)) from each other. The function considers the data as if from a multivariate normal distribution, whereby the ellipses constitute 99% confidence levels relative to the category means.\(^{224}\) Perhaps most striking from the plot is the result that only one level of the “type” variable—corresponding to the exegetical typology proposed in section 2.2—is significantly distinct from the other levels. That is, primary numerical, primary narrative and elicited transcriptions have coordinates that are not significantly distinct from each other. However, so-called summary transcriptions (often encomienda titles) are significantly distinct \((p < 0.01)\) from all other “types” of khipu transcriptions. This designation is supported by the ellipses plot of the “num_speakers” variable: the “na” level is significantly distinct \((p < 0.01)\) from either one or multiple speakers. This level is nearly one-to-one with the “summary” level of the exegetical typology: 30 of 32 “summary” transcriptions have “na” levels for the “num_speakers” variable.

The close-reading typology carried a mere 1-in-4 “success” rate in identifying meaningful transcription categories, when tested statistically under the parameters described above. Yet, novel typological categories emerge from the plot of illustrative variables. The pre/postconquest variable—reflecting whether two khipu transcriptions pertained to pre- or postconquest data—is strongly significant along PC1 \((p << 0.01; \text{correlation coefficient} = 0.50)\) but not significant along PC2. In contrast, while the currency variable is significant along PC1 \((p < 0.01; \text{correlation coefficient} = 0.24)\) the mention of Spanish currency is the only significant chronological category along dimension 2 \((p < 0.01; \text{correlation coefficient} = 0.13)\).\(^{225}\) In sum, the analysis yielded a set of empirically-distinguishable typological categories for consideration in clustering khipu transcriptions. What do these tests contribute in practice?

---

\(^{224}\) Husson, Lê and Pagès, *Exploratory Multivariate Analysis*, 24-25, 37. Specifically, Student’s t-test is performed to compare the coordinate average of individuals described by the given level on a single PC to the overall coordinate average of all individuals on the same PC. Reported \( p \)-values consider the null hypothesis \( H_0 \), that the mean coordinates of individuals possessing the level are not meaningfully different from the population mean coordinates. The test is significant at the 1% level (and thus \( H_0 \) is rejected) for \( p < 0.01 \).

\(^{225}\) Other significant categorizations include “small” versus “large” population centers along dimension 1 and “south” versus the group of “central” and “north” latitude zones along dimension 2. See appendix A.1.1.1 for the report of factor loadings and significance levels. The population categorization relies on high-level, informal population assignments and is subject to labeling error, barring a more thorough study of historical population. The “south” latitude level largely reflects the abundance of currency-related khipu transcriptions from south of Lake Titicaca—moving largely in-step with tribute restitution trials in the mid-sixteenth century. Latitude is not significant along dimension 1 (unlike both pre/postconquest and currency) and was thus omitted from the analysis.
Husson’s joint representation offers a summary of the methods discussed above. Having studied the relationship among variables and individuals, we join the analyses. Figure 4.6 is a biplot—a chart that overlays the rows and columns of the original data.\textsuperscript{226}

\textsuperscript{226} Levshina, \textit{How to do Linguistics with R}, 365. Only the highest contributing categorical variable levels are plotted.
Figure 4.6: Biplot (Lemma Implementation)
The biplot offers a visual metric for transcription clustering, including a synthesis of previous suggestions. As in figure 4.4, the individuals are sized and colored corresponding to their contributions to the plane. In addition, active, supplementary quantitative and categorical variables are overlaid. It is important to note that the biplot representation itself is merely illustrative, since the two clouds correspond to different vector spaces. However, the directions of variables (and category levels) are meaningfully compared to the placement of individuals. This exercise supports the earlier interpretation. **PC1 separates preconquest transcriptions with subject-focused linguistic structure (right-hand half-plane) from postconquest, object-focused enumerations involving giving items to others (left-hand half-plane). PC2 separates low-complexity, non-monetary repetitive narrations (bottom half-plane) from more complex, monetary data.** The memorias take high values in terms of the variables with which they share a coordinate direction. Therefore, while the superimposed distances are distorted, the quadrants of figure 4.6 are broadly labeled:

- Top-left quadrant: postconquest currency transactions in relatively complex narrations (e.g. queries in tribute restitution trials, etc.).
- Bottom-left quadrant: postconquest enumerated lists in simple narrative form; usually non-monetary goods/services provided to conquering Spanish bureaucrats.
- Bottom-right quadrant: compiled preconquest data in early, repetitive summary transcriptions (e.g. encomienda titles). Collections of “indio(s)” and “pueblo(s).”
- Top-right quadrant: preconquest narrative-historical and numerical khipu data of varying complexity.

The most pronounced resulting opposition is between pre- and postconquest cords (dimension 1). The categorical levels “pre” and “post” fall opposite from each other along PC1; the supplementary quantitative variable “year” takes high values among individuals in the left-hand half-plane. A plot of the chronologically-segmented transcriptions (see figure 4.7) confirms the PC1 interpretation: the right and left half-planes are associated with pre- and postconquest

---

227 Husson, Lê and Pagès, *Exploratory Multivariate Analysis*, 27. The plot of variables is in $\mathbb{R}^{72}$, while the plot of individuals is in $\mathbb{R}^{15}$. 
data, respectively. Figure 4.8 contrasts the recording of currency data along PC2: the lower half-plane corresponds to non-monetary data, while the upper half-plane records currency data.

---

228 The 20 transcriptions with the highest individual contribution to inertia are plotted.
Figure 4.7: Pre/Postconquest Transcription Segmentation (Lemma Implementation)

Figure 4.8: Currency Transcription Segmentation (Lemma Implementation)
It is acknowledged that these two categorical variables—pre/postconquest and currency—are interdependent. This is intuitive graphically: since currency was introduced to South America only after the Spanish conquest, we would not expect currency-related khipu data to appear in the upper right quadrant of figure 4.8 (which is associated with preconquest cord data). This is confirmed in the figure. Despite their interdependence, the two categories remain meaningfully distinct: PC2 exhibits a notable separation of monetary/non-monetary khipu data, even when restricted to postconquest cords (see left half-plane of figure 4.8).

The linked approaches above offered a glimpse into contextual measures elicited from largely non-contextual linguistic features. The significance of several novel categorizations was demonstrated empirically, including hierarchical demarcations—that is, a transcription is best classified first as either pre- or postconquest, followed by either currency/non-currency (among postconquest cords). The stability of these measures was tested by repeating PCA on a second, context-agnostic linguistic feature: the relative frequency of universal part of speech (UPOS) tags across the transcription corpus.

4.5.2 UPOS Implementation

The appropriateness of PCA was probed using the Bartlett Test, the Kaiser-Meyer-Olkin (KMO) Measure and the correlation matrix determinant. Recall that 13 UPOS tags were introduced in section 3.4. The resulting correlation matrix was computationally singular: the determinant was not meaningfully different from zero, failing to meet the determinant reference value of 0.00001 introduced above. Following Field et al.’s procedure for this situation, I analyzed the correlation matrix and omitted three UPOS tags which were highly collinear (correlation coefficients > 0.9), preserving one tag in cases of highly correlated tag pairs. I also omitted PART (a fourth UPOS tag), which had an individual KMO value of 0.32 (far below Kaiser’s 0.5 threshold and the only active variable with a KMO value < 0.5). The nine active UPOS variables used in the study are: DET, ADJ, ADP, AUX, NOUN, NUM, PRON, SCONJ, VERB.

- Bartlett. The null hypothesis \( H_0 \) was rejected \( (\chi^2(45) = 305.10, p << 0.001) \), and thus the test is significant: the observed correlations do not result from sampling error.

---

229 See section C.2.2.2 for the implementation of Bartlett’s Test, KMO and the correlation matrix determinant.
230 The abbreviations signify: determiner, adjective, adposition, auxiliary, noun, number, pronoun, subordinating conjunction and verb, respectively.
- KMO. The general KMO statistic value obtained is 0.72, which is an intermediate value (and 0.02 higher than the KMO statistic in the lemma implementation). By Kaiser’s criterion, this value reflects an adequate sampling; by Hutcheson and Sofroniou’s criterion, the sampling is “good.”
- Determinant. The determinant of the correlation matrix (0.0106) is larger than Field et al.’s reference value (0.00001), and thus multicollinearity is not observed in the UPOS data.

Table 4.2 lists the cumulative percentage of inertia explained by the factors, which are extracted from the diagonalization. See figure 4.9 for the corresponding scree plot.

<table>
<thead>
<tr>
<th></th>
<th>Eigenvalue</th>
<th>Percentage of Variance</th>
<th>Cumulative Percentage of Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>comp 1</td>
<td>3.629157</td>
<td>40.32397</td>
<td>40.32397</td>
</tr>
<tr>
<td>comp 2</td>
<td>1.909806</td>
<td>21.22006</td>
<td>61.54403</td>
</tr>
<tr>
<td>comp 3</td>
<td>0.946697</td>
<td>10.51886</td>
<td>72.06289</td>
</tr>
<tr>
<td>comp 4</td>
<td>0.811386</td>
<td>9.015397</td>
<td>81.07829</td>
</tr>
<tr>
<td>comp 5</td>
<td>0.731859</td>
<td>8.131761</td>
<td>89.21005</td>
</tr>
</tbody>
</table>

Table 4.2: First Five Eigenvalues and Explained Variance (UPOS PCA Implementation)
By the table, the first component explains 40.3% of the total variance; the second component 21.2%, etc. Two components (shaded in table 4.2) collectively explain 61.5% of inertia in the original table. As such, the first two factors are chosen to reconstruct the data.

This choice was in accordance with (1) Kaiser’s criterion for factor retention, which advises selecting factors with eigenvalues larger than one; (2) Loewen and Gonulal’s criterion of selecting factors until 60% of total variance is explained (here, 61.5% is explained); (3) ceasing factor retention at the point of inflexion (“elbow”) of the scree plot (at component 2 in figure
4.9); and (4) parallel analysis, by which only the first two eigenvalues, which exceed the 95\textsuperscript{th} percentile of randomized eigenvalues in 1,000 trials, are retained.\footnote{Loewen and Gonulal, “Exploratory Factor Analysis,” 196-197. See figure A.4 in appendix A.1.2.2 for the plot of simulated values. The choice of two factors is unambiguous in this case by parallel analysis (unlike in the lemma implementation).}

The choice of two components (versus the four studied using the lemma frequency measure) allows us to focus on the first plane—spanned by PC1 and PC2. The first plane captures nearly as much of total inertia as do two planes for lemma frequencies. Therefore, the study of individuals, variables and the joint representation (biplot) is repeated using only the first two dimensions. Figure 4.10 plots the nine active quantitative variables, as well as the same supplementary quantitative variables used in the lemma study.

Figure 4.10: Variable Factor Map for Dimensions 1 and 2 (UPOS Implementation)
The factor map allows us to study the relationship between the active and supplementary quantitative variables, since the plane represents over 60% of the original variance among the khipu transcriptions. I jointly studied variable clusters, anticorrelated variables and the relative magnitude of contributions from the variables. The most closely correlated variables include the groups of subordinating conjunctions and pronouns (“SCONJ” and “PRON”), as well as auxiliaries, adjectives and determinatives (“AUX,” “ADJ” and “DET”). At first glance, the first group might be thought of as collections of subordinate clauses with disproportionate pronominal usage. The second group might be labeled descriptive narrations with accompanying verb diversity. Once again, these abstract observations are concretized by studying the opposition of the variables along both PCs.

Along the x-axis (PC1), pronouns, subordinating conjunctions and verbs stand in strong opposition to the use of nouns and numbers. This division is tentatively described as that of explanatory narratives—characterized by use of two-thirds of the UPOS tags under study on the right-hand half-plane—as opposed to enumerated lists—collections of things (nouns) and quantities (numbers) on the left-hand half-plane. The “year” variable is significantly ($p < 0.05$) associated with negative values of PC1, suggesting that individuals on the left half-plane (negative values on PC1) tend to constitute later transcriptions.

The second dimension lacks strongly anticorrelated variables, with the strongest opposition apparent between determinatives and numbers. This opposition is contextualized by the broader diversity of parts-of-speech which contribute to narrative complexity in the upper half-plane: that is, six of nine UPOS tags project onto positive values of PC2.

By the UPOS measure, the preliminary variable clustering is introduced as a list of ranked linguistic diversity. I took UPOS-tag concentration as an informal proxy for narrative complexity, under which it is proposed that if a greater number of variables project onto a given quadrant, that individuals plotted in that quadrant exhibit greater variance due to the diversity of usage of the given parts-of-speech. By this measure, the top-right quadrant (characterized by 4 UPOS tags) represents linguistically complex transcriptions; the bottom-right and upper left quadrants (2 UPOS tags each) represent moderately complex narration, distinguished by nominal details (bottom-right) versus collections of things (top-left). By this proxy, transcriptions in the bottom-left quadrant, characterized by only one UPOS tag (number), are the least complex in their narrative structure—characterized by enumerated lists.
Individuals and illustrative variables offer a filter for assessing the preliminary variable clusterings.\textsuperscript{232} Figure 4.11 plots the 72 transcriptions in the space spanned by the first two PCs, proportionally sized and colored to reflect the contributions of the individuals to variance. Figure 4.12 plots the same eight categorical variables discussed in the lemma implementation. The ellipses represent 95\% confidence ellipses.

\textsuperscript{232} Recall Husson, Lê and Pagès’ (2017) notion of joint representation.
Figure 4.11: Contributions of Individuals (UPOS Implementation)
Figure 4.12: Categorical Supplementary Variables with 95% Confidence Ellipses (UPOS Implementation)
The plot of individuals offers insight into potential clusterings of memorias, weighted by their respective contributions to total variance. Transcriptions deemed large contributors include memoria 44 (a 1541 encomienda title allocated by Francisco Pizarro to Marcos de Retamoso in Tarapaca); memoria 10 (an enumerated list of items provided to the Spanish captains of the Mantaro Valley during the early years of the Spanish conquest); and memoria 17 (the same excerpt from the Sakaka tribute restitution trial that contributed heavily to axis construction during the frequent lemmas PCA). Native groupings are observed again among the individuals. Some of these groupings are intuitive: memorias 8, 9 and 10, for example, all constitute enumerated lists of goods given to conquering Spanish military captains in the Mantaro Valley. Other groupings are more curious: while memoria 20 and 39 are both encomienda titles, their locations of transcription (Cuzco and Majes, respectively) span a distance of over 300 kilometers. Given an analogous clustering of encomienda titles in the bottom-right quadrant among lemma data (see figure 4.4), further support is lent to the hypothesis that summary transcriptions constitute a significant category for explaining khipu narrative diversity.

The ellipses chart (figure 4.12) quantifies these observations using a normalized metric. Recall that non-overlapping ellipses denote significantly \( p < 0.05 \) distinct categories for explaining transcription variation. The “type” variable (corresponding to the exegetical typology) here suffers from sample size. The sizable ellipse around primary narrative transcriptions is attributable to the few available texts of this type. However, if these transcriptions (“pnar”) are grouped with the primary numerical transcriptions (as first proposed in the exegetical typology), then summary khipu transcriptions are indeed significantly \( p < 0.05 \) distinct from the other categories originally introduced.\(^{233}\) Once again, the exegetical typology was tested statistically, yielding a consistent conclusion with the lemma implementation: while summary transcriptions constitute a significantly distinct khipu transcription category, visualizations offered novel categories by which to cluster the excerpts in the corpus. The “num_speakers” variable supports this conclusion: the “na” level is once again significantly distinct \( p < 0.05 \) from both one- and multiple-speaker transcriptions.

\(^{233}\) Due to the small sample size among narrative transcriptions (“pnar”), their significant difference \( p < 0.05 \) compared to primary numerical transcriptions (“pnum”) is omitted from the present discussion, at the risk of overattributing meaning to potential sampling bias.
The novel categories proposed under UPOS-feature clustering are in accordance with those reached using lemma features. The pre/postconquest variable—relating pre/postconquest khipu data—is significant along PC1 \((p < 0.01; \text{correlation coefficient} = 0.19)\) but not significant along PC2. Especially concordant is the currency variable, which is not significant along PC1, but is the only significant categorical variable along PC2 \((p < 0.01; \text{correlation coefficient} = 0.17)\). In other words, PCA on UPOS features yielded a stronger version of the axial dichotomies reached using the lemma features: the first PC separates pre/postconquest data, while the second PC separates monetary from non-monetary khipu data.\(^{234}\) These empirically distinguishable categories emerged from two independent and content-agnostic measures of khipu transcription diversity.

The individuals and variables (both active and supplementary) are joined in figure 4.13, the biplot.\(^{235}\) The relationship between rows and columns can be explored and compared to the conclusions drawn from the lemma implementation.

\(^{234}\) There are fewer additional significant categorizations in the UPOS implementation than in the lemma implementation. See appendix A.1.2.1 for the table of factor loadings and significance levels. The “type” and “num_speakers” variables are discussed above. Along PC1, the population variable is significant in one of its three levels, weakly separating “large” population centers as drivers of khipu transcription diversity. The variable—given its reliance on high-level informal categorizations, was omitted from further analysis. Along PC2, no other qualitative variables or categories emerge as significant variable labels. It is interesting to note that the correlation coefficient of the currency variable along PC2 \((0.17)\) is stronger in the UPOS implementation than in the lemma implementation \((0.13)\). The segmentation of pre/postconquest and monetary/non-monetary khipu transcriptions is more pronounced in the UPOS implementation than with lemmas: pre/postconquest chronology is significant along PC1 but not along PC2, while currency is significant along PC2 but not along PC1.

\(^{235}\) Only the highest contributing categorical variable levels are plotted.
Figure 4.13: Biplot (UPOS Implementation)
To what extent did the spatial placements of individuals and variables support (or refute) the transcription clusters earlier proposed? Overwhelmingly, the UPOS implementation was in accordance with the lemma implementation. Recall that although the biplot is an illustrative construct (as individuals and variables exist in spaces of different dimensions), we can compare the quadrants and relative directions of the projections. The qualitative categories (plotted in purple in figure 4.13) demonstrate the similarities in results to the lemma PCA implementation. PC1 separates preconquest khipu data, rich in verbs, subordinate clauses and description (see “pre” level in right-hand half-plane), from postconquest data, which is saturated with rote repetitions of nouns and numbers (“post” level in left-hand half-plane). PC2 separates non-monetary symmetric lists (see “no” [non-currency] level in bottom half-plane) from rich monetary texts (upper half-plane). Direct comparison of the qualitative levels reinforces the concordance: in both implementations (lemma and UPOS), the bottom-right quadrant hosts the “no” [currency], “pre[conquest]” and “summ” [summary type] levels; the bottom-left quadrant hosts the “north” level; and the top-left quadrant shares the “post[conquest]” and “yes” [currency] levels. There is a 1:1 correspondence between the relative directions of the most significant qualitative variables, supporting the quadrant-by-quadrant interpretation reached above:

- Top-left quadrant: postconquest currency dialogues detailing lists of things (e.g. tribute restitution trials).
- Bottom-left quadrant: postconquest enumerated lists (see especially the strong influence of the “NUM” variable in this space; no other part of speech significantly distinguishes this quadrant).
- Bottom-right quadrant: preconquest data narrated in summary transcriptions (i.e. encomienda documents).
- Top-right quadrant: Lexically-diverse preconquest narrations (including narrative-historical and preconquest numerical transcriptions).

The strong opposition between pre- (right-hand half-plane) and postconquest (left-hand half-plane) data emerges again under the UPOS feature (dimension 1). Note the projection of “year” (a quantitative supplementary variable) on the left-hand half-plane of the biplot: individuals on this side of the plane take high chronological values (that is, later, postconquest transcriptions).
See figure 4.14 for the chronologically-segmented plot of transcriptions, which again confirms the PC1 interpretation. Figure 4.15 contrasts (albeit more moderately) the vertical division of currency (top half-plane) from non-monetary (bottom half-plane) khipu data.

The 20 transcriptions with the highest individual contributions are plotted. The currency variable presents more visual noise than under the lemma implementation. The 20 transcriptions with the highest individual contributions are highlighted to demonstrate the similarity to the lemma result. This visual noise is curious: How can the currency variable be more highly corelated with PC2 in the UPOS implementation, yet the individuals cluster less sharply along that axis? It is noted that the pre/postconquest axis can be interpreted as rotated slightly in this implementation: an approximately 30° line drawn with positive slope through the origin provides an analogous segmentation, among postconquest data khipus, between non-monetary (bottom-left quadrant) and monetary (top-left quadrant and left edge of top-right quadrant) khipus. However, the spatial opposition between levels is of primary interest here. I do not employ factor rotation (e.g. varimax) in this PCA implementation for this reason, since the relative spatial relationships between variables are preserved under rotation. Further, despite the noise in the visualization, the differentiation represented by the currency variable is significant at the 1% level in both implementations.
Figure 4.14: Pre/Postconquest Transcription Segmentation (UPOS Implementation)

Figure 4.15: Currency Transcription Segmentation (UPOS Implementation)
Quadrant-by-quadrant characteristics were synthesized across both PCA implementations to inform figure 4.16—a primordial division of khipu transcriptions by means of a statistical typology.

![Diagram of Statistical Khipu Transcription Typology](image)

Figure 4.16: Statistical Khipu Transcription Typology

The analyses above constitute linked approaches toward a statistical typology of khipu transcriptions. PCA produced a set of novel descriptive categories—chronology and currency among the most significant—not previously considered as “deciding” categories in the exegetical typology. A quadrant-by-quadrant segmentation was used to propose an empirically-tested categorization of khipu transcriptions. Yet, how might the memorias themselves be allocated to the proposed categories? Does the native clustering align with the typological ideal? In practice, how good is the statistical typology? These questions were probed using hierarchical agglomerative clustering—a related technique that further contextualizes the results of PCA.

### 4.6 Hierarchical Agglomerative Clustering Implementation

The biplots and ellipses plots demonstrated that on average, the statistical typology segments the transcription corpus (consisting of 72 individual memorias) into statistically

---

238 Recall that PCA is a closed, exploratory analysis. However, here one of the luxuries of an endangered language is that the sample corpus is (nearly) the population, and so exploratory studies of the corpus constitute a commentary on surviving khipu transcriptions. Nevertheless, whether these results can be extrapolated to khipu grammar more broadly is a matter of interpretation. The transcriptions we have today are a subset (those which have survived in archives) of a subset (those narrations which were transcribed) of khipu narrations. Nevertheless, the typology proposed herein is informative, despite the constraints of taphonomy and document preservation.
distinct categories: (1) summary transcriptions, (2) postconquest enumerated lists, (3) postconquest currency khipus and (4) preconquest narrative/numerical khipus. How well do these categories—as well as broader designations, such as whether transcriptions record pre- or postconquest data—persist when the memorias are sequentially grouped according to their closeness? I employed hierarchical clustering to visualize the order in which transcriptions plotted in the lemma and UPOS PCA implementations might be grouped, under a metric of maximizing within-group similarity while maintaining between-group difference. The subsequent visualizations contextualize the PCA plots in two ways. The first is that both approaches (PCA and hierarchical clustering) share the goal of visually representing variation in a data set, differing only in the means through which variation is plotted. In PCA, two transcriptions are deemed “similar” if their projections appear close on the plane(s) spanned by the PCs. In hierarchical clustering, this difference is captured by a tree diagram called a dendrogram, in which two transcriptions (leaves) are proximate if they share a common branch. The second is that dendrograms offer a discrete partition of the transcriptions according to user-defined criteria. This offers additional visualizations of the hypotheses generated in PCA—that the “pre/postconquest” variable significantly differentiates the corpus of transcriptions, for example—when transcriptions are subjected to mutually exclusive, indexed groupings rather than plotted planes of the Euclidean cloud (as in PCA).

Hierarchical agglomerative clustering (HAC) is an exploratory analysis technique that groups individuals through sequential mergers. The method constitutes a bottom-up approach: the algorithm begins by treating the 72 transcriptions as discrete texts and successively grouping together the most similar individuals (or groups of individuals), as measured by user-defined distance and amalgamation rules. In this implementation, PCA and HAC were applied to the same active variables: the matrix of 72 rows and 13 (UPOS) or 15 (lemma) columns, in which each cell contains the relative mean frequency. Interpretation of the resulting dendrograms informs our assessment of the stability of the axial interpretation in the PCA. This combined

---

240 Ibid., 192.
241 Recall that nine UPOS tags were retained in the PCA implementation due to multicollinearity in the correlation matrix. Since the Euclidean distance metric is applied to the rows of the matrix (i.e. the transcriptions), rather than the columns (which are found to exhibit multicollinearity), I retained the 13 original universal parts of speech for the hierarchical clustering analysis. As such, the set of active variables is slightly larger here (13) than in the PCA implementation (9).
approach actualizes Zuidema and de Boer’s model parallelization, whereby the resulting (dis)similarity between two exploratory techniques (PCA and HAC) allows us to augment the reliability of the conclusions.\textsuperscript{242}

I implemented HAC in R using the stats and dendextend packages.\textsuperscript{243} I subjected the \( k \) active columns from the lemma and UPOS tables—the same tables introduced in the PCA implementation—to the Euclidean distance metric, in which the row vectors (the transcriptions) are treated as points in \( k \)-dimensional space. The two resulting distance objects (one using lemmas and the other using parts of speech) constitute symmetric dissimilarity matrices with zeros along the diagonal (because a transcription is not distant from itself).\textsuperscript{244} For example, the cell with (row, column) coordinates (“memoria 10,” “memoria 24”) records the Euclidean distance between the vectors of two individuals: memorias 10 and 24. I used Ward’s amalgamation method, which considers intra-cluster variance, for clustering the transcriptions.\textsuperscript{245} Ward’s method has broad popularity in corpus linguistics studies because it generates moderately-sized clusters.\textsuperscript{246} The algorithm produced two dendrograms (one based on lemmas and the other based on parts of speech) from the transcriptions, where each of the 72 texts forms a leaf in a nested hierarchy of branches. To what extent did the two trees support the separation of pre/postconquest transcriptions described along PC1 of the UPOS and lemma PCA implementations? Figures 4.17 and 4.18 compare the distribution of pre- and postconquest transcriptions in the clustered groups to the theoretical ideal: a clean separation of the trees into pre- and postconquest branches.

\textsuperscript{242} Husson, Lê and Pagès, \textit{Exploratory Multivariate Analysis}, 192; Zuidema and de Boer, “Modelling in the language sciences,” 428.
\textsuperscript{243} Tal Galili, “dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering,” \textit{Bioinformatics} 31, no. 22 (2015). The stats package is part of base R.
\textsuperscript{244} Desagulier, \textit{Corpus Linguistics and Statistics}, 278-279.
\textsuperscript{246} Desagulier, \textit{Corpus Linguistics and Statistics}, 279.
Figure 4.17: Lemma Implementation of Pre/Postconquest Dendrogram
Figure 4.18: UPOS Implementation of Pre/Postconquest Dendrogram
The two dendrograms offer both vertical and horizontal spatial information. In these representations, the closeness of two transcriptions is understood as the vertical distance traveled from the leaves upward until reaching the common parent node. Intuitively, the vertical axis is a measure of the closeness of the groups of individuals. Note that each node represents an agglomeration event, in which two groups, individuals, or one of each are joined to form a larger cluster. Further, we can say that a transcription is misclustered if its label is a different color than the set of branches which emanate from it. For example, in both figures, memoria 6 (postconquest data; first red label from the left) is incorrectly associated with preconquest data.

In aggregate, the two figures support the findings of both PCA implementations: there is a visually distinguishable differentiation between a preconquest cluster (left-hand branches of the dendrograms) and a postconquest cluster (right-hand branches). It is acknowledged that the differentiation does not match the ideal: the right-hand branches of both figures, in particular, exhibit some intermixing of pre- and postconquest data. Nevertheless, the temporal segmentation is broadly supported by the figures. It is also noted that the dendrograms, which are each supportive of the PCA results, are also supportive of each other. This is observed by listing the shared misclusterings of transcriptions: both dendrograms incorrectly cluster preconquest transcriptions 3, 4, 23, 27, 56 and 61 among postconquest khipu data (right-hand branches). In other words, the consistent misclustering of these transcriptions implies that the narrative structure of these texts—and less the idiosyncrasies of the document features chosen to represent them—can be meaningfully associated with postconquest narrations. Yet, how can this concordance be formalized?

It is this question for which dendrograms offer an additional benefit as compared to PCA. Recall that in PCA, the concordance of the lemma and UPOS implementations was attested to by interpretation: e.g. in both PCA results, the pre/postconquest categorical variable significantly differentiated transcriptions along PC1. In contrast, hierarchical clustering outputs can be explicitly compared numerically. The basis for this comparison is figure 4.19, called a tanglegram, which positions the two dendrograms opposite to each other and connects their leaves with auxiliary lines. The colored lines represent groups of transcriptions which are common to both dendrograms. The dashed lines constitute branches whose leaves do not match one-to-one between both implementations. The labels are shortened due to spatial considerations. Therefore, the label “45” corresponds to “memoria_45,” for example.
Figure 4.19: Tanglegram
The conclusion is somewhat mixed. Both dendrograms agree, for the most part, in their separation of pre- and postconquest transcriptions. As plotted in figure 4.19, this division corresponds to the upper (postconquest) versus lower (preconquest) half-planes/dendrograms. Even leaves on non-identical branches can be connected by means of auxiliary lines which stay within the respective cluster: that is, rarely are groups of transcriptions associated with the preconquest cluster in one dendrogram and the postconquest cluster in the other. The two dendrograms are moderately harmonious visually. However, the quantity of dashed lines suggests that the tree topologies are far from identical. That is, while the two implementations successfully distinguish between pre- and postconquest narrations, the order and hierarchical relationships of groupings within those same categories differ markedly. Can we conclude, by statistical means, that the two trees are significantly faithful to each other, and thus that their differentiation of pre- and postconquest accounts can serve as supporting evidence for the segmentation reached by PCA?

This conclusion was supported by several means of comparing hierarchical clustering results, including Baker’s Gamma Index, the Mantel Test on cophenetic correlations and the Fowlkes-Mallows Index.\textsuperscript{247}

- **Baker’s Gamma.\textsuperscript{248}** This index constitutes the Spearman rank correlation between the ordered stages of agglomeration for each pair of transcriptions within each dendrogram.\textsuperscript{249} The index can assume values between -1 and 1, where a near-zero gamma suggests no similarity. The null hypothesis $H_0$ is that there is no meaningful association between the trees: one dendrogram can be expressed using the randomly shuffled labels of the other. In other words, the two trees are no more significantly similar to each other than either one of the trees is similar to a randomly shuffled version of itself. A value of $\gamma = 0.371$ is obtained. The statistical significance of the rank correlation is probed

\textsuperscript{247} See section C.2.4 for the Baker’s gamma, Mantel Test and Fowlkes-Mallows implementations.


\textsuperscript{249} Tal Galili, dendextend: Extending ‘dendrogram’ Functionality in R, R package version 1.10.0. https://CRAN.R-project.org/package=dendextend, 2019. There are $\binom{72}{2}$ unique pairings of the 72 transcriptions. For each of these pairings, and for each of the dendrograms, the largest $k$ is found such that when the given dendrogram is cut into $k$ clusters, the given pair of transcriptions remain part of the same cluster of that dendogram. The $k$ value is found simultaneously for the other dendrogram. Once compiled, the Spearman correlation is taken of the two ordered vectors of $k$ values, each with length $\binom{72}{2}$. 
through a permutation test, where the obtained $\gamma$ is compared to the null distribution of Baker’s Gamma under repeated resampling. The test is significant ($p << 0.001$) and we reject $H_0$; the two dendrograms are significantly related to each other and one dendrogram is not solely a randomly shuffled version of the labels of the other.\(^\text{250}\) The observed $\gamma$ is not exceptional, further reinforcing the notable concordance between the two dendrograms: a 95% bootstrap confidence interval of $[0.325, 1.000]$ is computed for $\gamma$ under 1,000 replications.\(^\text{251}\)

- **Mantel Test.**\(^\text{252}\) This test assesses the significance of inter-dendrogram agreement through random permutations of the rows and columns of the respective cophenetic correlation matrices underlying each tree. The cophenetic correlation between two transcriptions is defined as the correlation between the cophenetic distance (the height on the tree at which the two transcriptions are joined) and the original distance between the two transcriptions in the dissimilarity matrix.\(^\text{253}\) In practice, this test corresponds to random relabeling of both trees. The null hypothesis $H_0$ is that the pair of dendrograms is no more similar than pairs of randomly relabeled dendrograms generated from their underlying matrices. The test is significant ($p < 0.001$) and $H_0$ is rejected: we conclude that the two dendrograms are meaningfully similar.\(^\text{254}\)

- **Fowlkes-Mallows.**\(^\text{255}\) The index $B_k$ can assume values between 0 and 1, where higher values suggest meaningful similarity between two hierarchical cluster results. The test

---

\(^{250}\) See figure A.5 in appendix A.2.1 for the null distribution of Baker’s Gamma. If the observed correlation between the UPOS and lemma dendrograms were spurious, we would expect that the observed $\gamma$ could be exceeded more than a threshold percentage of the time in repeatedly sampled Baker indices between either dendrogram and shuffled samples of itself. The observed index in this case falls far above all values of the reference distribution. See also Michael Greenacre and Raul Primicerio, *Multivariate Analysis of Ecological Data* (Bilbao, Spain: Fundación BBVA, 2013), 104, 213-227 for examples of bootstrap sampling and permutation tests applied to hierarchical clusterings, and dendrogram comparison more broadly.

\(^{251}\) See figure A.6 in appendix A.2.1 for the bootstrap distribution of Baker’s gamma. See appendix C.2.4 for the confidence interval calculation.


\(^{253}\) Lapointe and Legendre, “Comparison Tests for Dendrograms: A Comparative Evaluation,” *Journal of Classification* 12 (1995): 270. The authors reached the conclusion that the Mantel Test is too conservative in its application to dendrograms—that permutation tests ought shuffle not only the labels (leaves), but also the tree topology and the cluster heights. The test was retained in this study for completeness, although it is acknowledged that recent studies (e.g. Harmon and Glor 2010) have found the test to have high type-I error (erroneous rejection of a true null) in phylogenetic tree applications in biology.

\(^{254}\) See appendix C.2.4 for the Mantel test implementation.

gauges the observed deviation from the expected number of concordant clusters under the null hypothesis, which is that the two dendrograms host independent clusters of fixed size.\textsuperscript{256} In practice, the labels of the two trees are randomly reshuffled, and the index $B_k$ is computed from a “matching matrix,” in which the entries correspond to the counts of matching constituents when the two dendrograms are cut to form $k$ clusters. The cutting is repeated for $k = 2, \ldots, n - 1$ (all nontrivial clusters), such that a plot of the index against the number of clusters, called a $(k, B_k)$ plot, offers a visual representation of the similarity of the two hierarchical cluster results across all $k$. In this implementation, $H_0$ is rejected ($p < 0.05$) for $k = 2$, which is the proxy for the pre/postconquest designation observed in the dendrograms.\textsuperscript{257} Figure 4.20, the corresponding $(k, B_k)$ plot, adds contextual visual support for the conclusion. In the figure, the index $B_k$ is plotted for each $k$. The black dashed line corresponds to the expected $B_k$ under $H_0$, which posits no relationship between the dendrograms. The red line represents the one-sided critical rejection value under $H_0$, based on the asymptotic distribution of $B_k$ and a 99% confidence level. Lastly, the blue line simulates the one-sided critical rejection threshold under 10,000 replications of $B_k$ for each $k$.


\textsuperscript{257} See appendix A.2.2 for the calculation of the critical rejection value for $k = 2$. 
The plot suggests that there is a significant similarity between the lemma and UPOS hierarchical clustering implementations not only under the goal of forming two clusters, but also in more granular divisions, even when cutting both trees to form almost \( k = 71 \) nontrivial clusters.\(^{258}\)

The linked approaches above further bolstered the primary division proffered in this study: that of a significant and meaningful division in the narrative structures of pre- versus postconquest khipu transcriptions. The introduction of tree robustness-testing measures contextualized the explicit comparison of clustering results, adding additional depth to the axial-chronological interpretation from the principal component analysis. These concordant findings

\(^{258}\) Fowlkes and Mallows, “A Method for Comparing,” 557. The language of \( k < 71 \) (the last nontrivial clustering) is also attested to by the authors (557), who note a “precipitous fall off at the very highest values of \( k \).”
offer increased confidence in a subsequent case study: explicitly testing one of the oldest extant hypotheses in khipu transcription studies.

4.7 “From Knots to Narratives”: Statistical Support for Urton (1998)

Among the few comparative analyses of khipu transcriptions is Urton’s study of syntactic structure in three transcribed narrations of pre- and postconquest khipu data. Spanning (a) preconquest labor tribute, (b) early colonial taxation and (c) trial proceedings contexts, the three transcriptions are included in the corpus as memorias (a) 4, (b) 7 and (c) 17. Comparing the three accounts, Urton concluded that the early colonial period saw a “radical transformation… in the recording techniques, information systems, and grammatical constructions of khipus in the transition from the Inkaic to the Spanish tribute system.” Specifically, it is argued that the recording of Spanish monetary and non-monetary tribute (1) flattened the diversity of preconquest linguistic categories recorded in khipus; and (2), as a result, vastly underutilized the record-keeping capacity of khipu cords. Urton reached this conclusion through comparison of the Spanish syntax in the three transcriptions, in which the shift from preconquest provision of tribute labor to the Inka state (memoria 4) to postconquest invoices of monetarily-valued things (memorias 7 and 17) was evidenced by the relative lexical poverty of the postconquest narrations. Specifically, Urton hypothesized that “early colonial transformations of the khipu record-keeping system” included:

1. “the virtual elimination of fully grammatical… narrative constructions in favor of attenuated, non-narrative clauses composed primarily of nouns and numbers”;  
2. retaining a “large array of representations of nouns and numbers, but essentially only one verb: ‘to give’ [‘dar’]”;  
3. and “the elimination of… native classificatory terms denoting actions required of subjects in the Inkaic tribute system.”

This study provides (to my knowledge) the first statistical evidence supporting statements (1) and (2) of Urton’s hypothesis. Urton’s broader claim of an early colonial (3) “radical

---

259 Urton, “From Knots to Narratives.”
260 Ibid., 422; emphasis added.
261 Ibid., 427.
262 Ibid., 410, 427.
transformation” in the grammar of khipu narration in the study is also found to be empirically justified. To see why, I synthesize the PCA and HAC results case-by-case below.

1. The left-hand half-plane of the UPOS biplot (figure 4.13), which captures postconquest khipu data, is saturated with only three parts of speech: nouns, numbers and “adpositions,” which in the UPOS syntax refers to prepositions and postpositions (often introducing simple noun phrases). Nouns and numbers are each strongly anticorrelated with PC1: their correlation coefficients are -0.76 and -0.81, respectively. These two variables, in their opposition to more linguistically rich features like verbs, adjectives, pronouns and subordinating conjunctions, are essential to explaining inter-transcription variance along PC1. Collectively, the opposition between linguistically “rich” preconquest text and attenuated postconquest narration accounts for over 90% of the variability in khipu transcriptions along PC1—an axis which significantly separates \((p < 0.01)\) pre- and postconquest transcriptions of khipu narrations.

Figure 4.21 plots the variable contributions (bars) against the expected average contribution (red horizontal line) under the uniform hypothesis (that each variable contributes equally to inertia).

---

263 See appendix A.1.2.1 for the full description of variable correlation coefficients.
264 This measure combines the contribution values of numbers, nouns, pronouns, subordinating conjunctions and verbs. Under the uniform contribution hypothesis, these five variables would collectively explain only ~56% of the total variance along PC1, markedly less than the 90.44% of actual explained variance along this axis. The contribution values for the “NUM” and “NOUN” variables are 18.128 and 15.887, respectively, which in sum constitute 34.01/90.44 = 37.61% of the contribution to the linguistically rich/poor binary. See table A.6 in appendix A.1.2.1 for the full list of variable contributions.
As is evident from the figure, the transition from pre- to postconquest khipu cords is observed to be a tale of diminishing linguistic complexity. Urton’s observation of a breakdown in fully grammatical khipu construction is supported by the UPOS variable plot (figure 4.10): two-thirds of the universal parts of speech necessary for complete grammars are overrepresented in preconquest accounts (right-hand half-plane). And, while three parts of speech are overrepresented among postconquest accounts (left-hand half-plane), those parts of speech that are present (nouns, numbers and adpositions) in this period are, by themselves, restricted in generative capacity to the creation of “attenuated, non-narrative clauses.” While postconquest accounts are by no means lacking in lexical diversity (they are, after all, distinguished by the sheer variety of nouns read out from khipus), the generative capacity of khipu narration was flattened in the early colonial period, under the metrics proposed herein. That is, much of the linguistic
diversity in colonial khipu narration can be attributed to new categories of things introduced by the conquering Spanish—currency, wheat, pigs, etc.—and less the intrinsic lexical complexity of the khipu narrations themselves. As Urton suggested in his study, such enumerated lists of objects given to conquering recipients might have needed just a single verb and direct object entry—“gave” and “to Pizarro,” for example—with all subsequent cords recording two data points: a noun for each tribute item and a number describing its quantity and/or monetary value. While it is acknowledged that such simple “primarily mnemonic constructions” are not restricted to the colonial period, the data suggest that, in aggregate, a disproportionate share of early colonial data recordings underutilized the expressive capacity of khipu cords.265

2. Urton hypothesized that the recording of Spanish tribute data could proceed requiring only the verb “dar” [“to give”). The lemma biplot (figure 4.6) supports this interpretation. One of only two verbs among the most frequent 15 lemmas across the corpus is “dar,” which is unambiguously associated with postconquest khipu data (projection on the left-hand half-plane of figure 4.6). The verb “dar” is significantly \( p < 0.01 \) anticorrelated with PC1 (but more moderately than in the case of nouns and numbers): its correlation coefficient is -0.43. Along the first axis, “dar” is second only to “de” [“of”] in the strength of its anticorrelation to PC1. In other words, transcriptions with similarly negative coordinates along PC1 are strongly (1) postconquest data and (2) saturated with orders “to give” different goods, often in enumerated lists. To make this argument more explicit, observe in figure 4.6 the close association of “dar” (an active variable) with “year” (the supplementary quantitative variable, plotted in blue). The close correspondence between the two projections bolsters Urton’s argument: even among postconquest accounts, the passage of time and the increasing relative frequency of “dar” move in step with each other. Neither “dar” nor “year” are significantly correlated with PC2 at the 1% level, suggesting that PC1 (a separation of pre- from postconquest khipu accounts) is an informative locus on the plane by which to understand the usage of “dar” in postconquest accounts. The “dar” variable is moderately represented by PC1, with a

265 Urton, “From Knots to Narratives,” 424. See Urton’s reference to a simple enumerated list of camelids recorded on the khipu of a cacique from Tapacari as an example of preconquest data retained using a largely mnemonic construction.
Not only is “dar” associated with postconquest contexts, but the best variable to explain its relative oversaturation is the passage of time—following the imposition of an object-centered contributory economy of items with monetary value beginning in 1532. The plot allows us to comment on the relative saturation of accounts with “dar.” Despite the appearance of this specific verb in both pre- and postconquest accounts, the visualization aids in quantifying the relative growth over time of flattened, lexically-poor verbs such as “dar.” Giving to the state, in the form of labor tribute, was an essential component of life under Inka rule. Yet, cases in which ethnocategories of service were grouped together into the single, unmarked term “give” are best identified with narrations of postconquest data in the corpus. It is also noted that the generative capacity of lemmas differs markedly for positive and negative values of PC1. In the right-hand half-plane of figure 4.6 (largely preconquest data), complex forms like “que” [“that”] (often introducing verbal phrases), “otro” [“another”], “pueblo” [“village”] and “indio” [“Indian”] stand in contrast to postconquest giving (“dar”); collectively, the opposition between these five lemmas accounts for almost 60% of the variability in khipu transcriptions along PC1—which significantly separates (\(p < 0.01\)) pre- and postconquest khipu transcriptions (in accordance with the UPOS implementation). Under the same proxy used in the UPOS implementation—where the fraction of variables on each half-plane of the plot is used to approximate linguistic diversity—lemma frequencies support the narrative of diminishing linguistic complexity in khipus during the early colonial period: 60% of the most frequent lemmas are overrepresented among preconquest accounts (right-hand half-plane), better facilitating lexically diverse khipu narrations.

---

266 The cosine squared value represents the importance of the given PC (in this case, PC1) for interpreting the observations (the relative frequency of “dar” [“to give”] in the corpus). The contribution of “dar” to the formation of PC1 itself is relatively low (about 4.12% of inter-transcription variance along this axis is explained by the appearance of “dar”). Two comments can be made here. The first is that PC1 faithfully represents the variable—the moderate cos\(^2\) value of 0.184 supports the statement that PC1, which is significantly distinguished by pre- and postconquest data at the 1% level, is significantly anticorrelated with the relative frequency of “dar.” The second is that more variables (15) are used in this implementation than in the UPOS study (9). It follows that the relative expected contribution of any given variable will be lower than in the UPOS implementation. Nevertheless, it is the directional correlation of “dar” with postconquest khipu data that is emphasized in this interpretation.

267 See appendix A.1.1.1 for the full report of variable contributions and correlation coefficients. The collective contribution to variance along PC1 is 57.6%, which is markedly more than the percentage of explained variance with five variables under the uniform contribution hypothesis (33.3%).
3. Do the joint conclusions of both PCA and hierarchical clustering implementations support Urton’s statement of a “radical transformation” in khipu grammar following the Spanish conquest? The concordant segmentations of khipu transcriptions produced in this study bolster Urton’s hypothesis. Put more strongly: at the 1% level, and by two content-agnostic measures (lemma and UPOS), the occurrence of the Spanish conquest significantly differentiates the narrative structure of the largest existing corpus of khipu transcriptions. The results from each PCA bolster each other: a general trend is observed from lexically rich preconquest khipu narrations (with a greater variety of frequent lemmas) to flattened postconquest narrations, in which a smaller diversity of frequent lemmas operated within less generative part-of-speech categories. Two hierarchical agglomerative clusterings performed on the same active variables produced statistically significantly concordant segmentations of the 72 transcriptions, bolstering the axial interpretation of the PCA. A radical transformation, indeed.

The arguments above constitute the introduction of statistical methods to studying preexisting khipu transcription hypotheses. As such, I conclude the section by outlining a set of guiding principles for the benefits and limitations of these methods in khipu transcription studies.

The chosen approach demonstrates the importance of using multiple features (i.e. UPOS and lemma frequency) within PCA implementations—precisely because not every part of Urton’s hypothesis (numbered 1-3 above) can be tested using the same measure. Urton’s statement regarding the abundance of nouns and numbers in postconquest tribute recordings, for example, is not meaningfully tested using the 15 most frequent lemmas—counts of words like “a,” “el” or “de.” Instead, noun and number frequencies were tested directly, using a UPOS implementation that recognizes nouns and numbers as discrete aggregate categories. The PCA implementations were both individually significant and were supported by significantly concordant dendrograms from the hierarchical clustering—the UPOS implementation supports Urton’s statement (1) and the lemma implementation supports statement (2). Jointly, the statistically significant findings were aggregated to provide empirically-tested justifications for the portion of Urton’s hypothesis that cannot be tested statistically: whether 1532 constituted a “radical transformation” of tribute data recording in khipus.

The results suggest that statistical typologies expand the vocabulary of argumentation to supplement exegetically-derived khipu hypotheses. In some cases, specific claims can be tested
statistically (e.g. whether postconquest accounts—in relation to those from before 1532—are characterized by nouns and numbers; whether certain khipu knots are found in widely conventionalized orientations in the corpus of archaeological khipus,\textsuperscript{268} etc.). Others require a cross-disciplinary mapping: Urton’s “radical transformation,” under one interpretation, can be identified with “a statistically significant ($p < 0.01$) differentiation under pre- and post-1532 segmentation.” This interdisciplinary translation, within the confines of the chosen parameters, can empower our historical arguments: we can further defend the use of terms like “radical” by means of hypothesis testing and measures of statistical significance. Ultimately, I do not propose replacing close-reading and interpretation with cold, unfeeling statistical methods, devoid of cultural context. Instead, the findings—which provide novel support for a 20+-year-old hypothesis, show the promise of an expanded vocabulary for the study of khipu transcriptions, and the study of khipus more broadly. A wealth of testable hypotheses is available from previous studies; a small handful were engaged with in this study. Yet, the exploratory methods I have implemented offer both inroads into preexisting research questions, as well as novel insights into suggesting new questions. Such questions could include study of changing compositions/quantities of items (llamas, chuño, wheat, etc.) demanded in early colonial tribute documents; the full spectrum of verb diversity in khipu transcriptions as a function of time; or aggregate study of prices and monetary value assigned to different goods throughout different areas of the early colonial republic, facilitated by the tokenized, lemmatized and syntactically annotated corpus compiled in this study. The tools introduced herein, as well as the long-overdue digitization and compilation of available khipu transcriptions, establish the necessary infrastructure for providing quantitative insights into the questions entered into the academic record decades ago.

At the same time, it is imperative to state what studies like this one do not accomplish. Principally, the conclusions I have reached are valid (1) under a chosen set of parameters and (2) for the specific transcriptions included in the corpus. For this reason, it is speculative to say that the results “confirm” or “prove” Urton’s hypothesis. Urton included a warning to this very point in his study, which is dutifully acknowledged: “uncertainty cautions against overinterpreting the meager evidence available in the Spanish transcription with respect to specific grammatical

\textsuperscript{268} See Clindaniel 2019 for promising statistical approaches to analyzing archaeological khipus, including hypothesis testing inspired by previous studies of postconquest khipu signs.
elements that may have been recorded on the khipus.\textsuperscript{269} Rather than “proving” or “confirming,” my results lend support to Urton’s hypothesis. I have also attempted to hedge against spurious correlations by using broad categories—“verbs” instead of pluperfect, imperfect and preterite—and appropriate sample sizes to flatten potential circumlocution introduced by Spanish lenguas and scribes; and to approach, however tentatively, the underlying khipu narration. Similarly, Urton’s statement that “the record-keeping system of the khipus included verbs” is difficult to comment on under the parameters that I have chosen. Commentary was offered on the relative mean frequency of verbs across time and space, among the 72 transcriptions in the corpus (that is, under a specific set of parameters). However, broader extrapolations are not borne out from statistical analysis alone—PCA is an exploratory technique precisely because it yields insights into patterns within the data set under study.\textsuperscript{270} In addition, neither this study, nor Urton’s, focuses on numerical quantities in tribute accounts—amounts of goods or levels of prices.

Urton’s exegetical study was offered as a case study in applying PCA and HAC—exploratory multivariate data analysis tools—to specific hypotheses. Aggregate lexical diversity was studied using novel methods to probe how khipu transcriptions might inform researchers regarding the performative act of khipu narration. The discussion of broader considerations and challenges facing this study, and those facing computational historical linguistics more broadly, is pursued in the subsequent section.

\textsuperscript{269} Urton, “From Knots to Narratives,” 428.
\textsuperscript{270} See Desagulier 2017.
Chapter 5
Considerations and Challenges

The results invite considerations regarding both the application of corpus linguistics techniques and the wholesale aggregation of khipu transcriptions. I engage with these groups of questions separately below, noting the importance that questions of methodology hold in studies which use multivariate techniques.

5.1 Challenges to Corpus Linguistics

Much support has been offered for the ways in which textual statistics can add novel visualizations and generate new hypotheses. However, a more fundamental question arises: is the application of statistics to natural language itself an exercise worth pursuing? There is indeed a certain eeriness to embedding the narrations of khipukamayuqs within the language of eigenvector and dendrograms. One locus for criticism is the use of hypothesis testing methods in the study. For example, in the PCA implementations, I employed Bartlett’s Test of Sphericity, whose null hypothesis $H_0$ is that observed nonzero correlations between linguistic variables result from sampling error and randomness. However, it might seem intuitive that the relationship between linguistic variables is never completely random. As Kilgariff has noted: “language is not random, so the null hypothesis is never true.”

This observation has particular relevance to the UPOS implementation: might we not expect that a transcription saturated with numbers would also be saturated with nouns, since the two are so clearly associated in human speech? In enumerated lists (one of the most common backdrops for the transcriptions), we rarely hear isolated numbers: 12, 20, etc. Instead, the numbers are counts of people and things: 12 llamas, 20 villagers, etc. While less obvious in the frequent lemma context, we might expect that transcriptions saturated with “indios” might also be saturated with “dar”—since the indigenous villagers were almost always the ones giving, whether camelds, currency, or blankets. Spurious correlations—nonrandom but arbitrary associations between features—are also of concern due to variable sample size. Natural languages are characterized by a Zipfian distribution: a relatively large number of rare tokens coexist with a small number of very

---

frequent tokens. In practice, this raises concerns over infrequent tokens: their smaller sample sizes make any single association a larger percentage of total frequency.

A mitigating strategy employed here was to focus on high-frequency features, which are not as clearly subject to spurious correlations. For example, I retained only the 15 most frequent lemmas in the corpus. This has the added effect of diminishing the presence of sparse matrices, which have distributional effects on parametric statistical tests. It is dutifully acknowledged that the correlations between linguistic features are nonrandom: this was shown by the Bartlett Test in the lemma and UPOS PCA implementations. However, this specific test was used less as a measure of a “special” relationship between variables, as much as a checkpoint in determining the appropriateness of PCA. Indeed, lack of a correlation is not the fear: four active variables were omitted from the implementation because of the strength of their correlation with other variables. Instead, hypothesis testing was used to establish moderate correlations, which were then contextualized through multivariate visualizations. The reader is challenged to consider whether the process of statistical treatment of language seems uncanny, simply because most of human history has involved thinking language through interpretation and exegetical experience. A conceptual shift is challenged and encouraged, consistent with the broader themes of this essay.

I employed multivariate methods that take moderate correlations as the starting point—not the endpoint—of analysis. One can simultaneously acknowledge the interdependence of linguistic features (as is done in this study) while suggesting that within the morass of textual data, certain relationships stand out. In this study, it is argued that pre- versus postconquest cords and the recording of currency are unexpectedly strong differentiators of khipu narration. The relationship between variables is considered outside of isolated pairwise cooccurrences. Pairwise correlation is important nonetheless and was noted throughout the analysis. However, the most robust and interesting conclusions emerge from the spatial contrast of large groups of variables—that postconquest noun and number clauses are significantly differentiated from linguistically richer preconquest narrations, for example. The types of hypothesis testing employed here are deeply contextualized and remain reflective to errors of arbitrariness.

Kilgarriff, “Language is never, ever, ever, random,” 266.
Ibid., 57.
Probabilistic understandings of language usage remain extremely important—this study moves beyond non-randomness to suggest the non-arbitrariness of the linguistic relationships. In fact, the non-randomness of language is precisely the reason for introducing dimensionality reduction techniques like PCA and HAC—to represent the data faithfully in fewer planes and identify the primary contributors to moderately correlated variable levels.

Dimensionality reduction also engages with concerns over contributions from multiple variables. This issue was immediately relevant during PCA when assessing multicollinearity in the correlation matrices. While it is indeed difficult to isolate specific variables, linear algebra offers a vocabulary for quantifying the contributions of these same variables in constructing the coordinate axes. Just as HAC offers an illustrative hierarchy of the khipu transcriptions, recasting the individuals in terms of linear combinations of the active variables (i.e. PCA) also offers a comparative vocabulary by which to assess the importance of specific labels (chronology, altitude, etc.) for differentiating the data. Dimensionality reduction offers a discerning lens through which we can apply both exegetical and statistical techniques.

The centrality of features and frequencies to corpus linguistics (see above) raises the question of whether UPOS labels and lemmas are defensible document feature measures. As the starting point of corpus analysis, choice of feature remains an area of ongoing academic debate, although there is moderate consensus that high-frequency measures (e.g. lemmas in this study) generate reproducible and robust differentiations among texts. In practice, this means that the chosen words are both more likely to appear than other word tokens, while at the same time more likely to vary in a discernible and meaningful way. This same logic applies to the choice of UPOS categories as the second linguistic feature. High-level categorizations of words, in addition to mitigating the impact of Spanish scribal variation, yield insights into the aggregate compositions of document genres—whether narrations that are clear in their contextual difference are also differentiable in their fundamental narrative difference. In addition to the long-documented reputation of these feature measures, it is emphasized that they are chosen for

---

their applicability to the specific research questions of the study—examining how form is traceable to underlying differences in the function of khipu grammar.  

It is also acknowledged that many of the sources consulted here are from the field of stylometry. What is the relevance of authorship attribution techniques when I stated earlier that the study is not concerned with (dis)proving the identity of any given khipu narrator? The principles of authorship attribution are helpful because of their careful consideration of literary style. It is important to remember that while I employed stylometric techniques (most notably the usage of function words for text clustering), these were implemented (intentionally) at moderate levels of granularity. As described above, I used the 15 most frequent lemmas—including somewhat contextual (non-function) words like “Indian” and “to give”—to acknowledge the ubiquity of these terms throughout the corpus. I employed 13 parts of speech—a far cry from available part-of-speech taggers offering over 50 different syntactic tags—to introduce specificity in the analysis while avoiding the mere isolation of scribal style. What I call a hybrid approach, and what Zuidema and de Boer (2013) call model parallelization, leverages the established strategies of multiple disciplines to study messy historical data.

In summary, simultaneous use of multiple models allows us to think of corpus linguistics as a “probabilistic continuum, where individual pieces of evidence can increase or reduce support for a given hypothesis.” Model parallelization develops an explicitly comparative literary acumen: content-agnostic linguistic features, in line with the principles of corpus studies, facilitate deeper analysis of historical texts. While “language is never, ever, ever, random,” this study identifies meaningful variation within variation.

5.2 Challenges to Aggregate Interpretation of Khipu Transcriptions

The results raise additional questions surrounding the aggregation of khipu transcriptions. To begin with, does the temporal differentiation of khipu transcriptions found in the study just reflect the introduction of the Spanish language to the Andes? How confident can we be that the results are not solely attributable to Spanish scribal variation?

---


282 See Kilgarriff 2005.
The concept of linguistic flattening introduced by Urton (1998) is helpful in approaching this question. Indeed, I find that transcriptions of preconquest khipu cords differ significantly from narrations of postconquest data. This is, in part, attributable to subject matter: while I have used prominent content-agnostic measures, no linguistic feature is immune from influence by the subject matter of a particular transcription. We would expect, for example, that a narrated currency khipu—an enumerated list of prices and quantities—would contain more numbers as a percentage of total words than a preconquest account of Inkaic labor tribute. As a result, one could argue that the exercise as a whole is tautological: when times change, I find that times have changed. The historical record itself attests to the linguistic challenges faced by lenguas attempting to synthesize Quechua and Spanish prose: the Inka Garcilaso noted the “difficulty of the general language of Peru” in his early seventeenth-century account, posing a provocative question when recounting the challenges faced by the early interpreters: “if this ignorance still existed after… years of familiarity and intercourse between Indians and Spaniards, how much greater must have been the deficiency of the interpreter when the only contact and concern was arms and warfare?”283

It is important to remember that the transcriptions in the corpus all took place after 1532. While the three-person khipu narration event was itself vulnerable to misunderstanding, our surviving transcriptions of pre- and postconquest cords were subjected to the same basic translative process. One (many) native khipukamayuq(s), summoned to narrate his (their) accounts in legal proceedings, was (were) interpreted by a lengua, whose verbal output was in turn recorded according to scribal norms. To the extent that this process did confuse khipu narration, I argue that narrations of both pre- and postconquest cords would have suffered from the same confusion. One might point to the high frequency of Quechua proper names in narrated preconquest cords as a prime site of confusion. Yet, the Andes did not surrender its language: even simple, postconquest enumerated lists of items taken by Spanish encomenderos list the Quechua-named villages from which particular goods were provided. The lengua and the Spanish scribe, in their imperfections, formed an unintentional controlling variable in this sense, since all of the transcriptions proceeded through them. Their ubiquity might well explain why the “language of narration” variable employed in PCA did not significantly differentiate the corpus texts: Spanish participants in khipu narrations flattened the speech of khipukamayuqs—whether

283 de la Vega, Royal Commentaries, 683.
in Quechua or Aymara—into notarial Old Spanish. In combination, these factors mitigate some concern. By the mid-sixteenth century, different khipu cords—some produced before 1532 and others after—produced significantly and structurally different narrations, even when recited contemporaneously. This suggests that the Spanish conquest introduced more than Christianity and wheat; the recording of information in khipus itself reflected a new grammar—one preoccupied with grand totals and rote enumeration, “completely integrated with the Spanish political economy [and] interested only in transferring wealth from indigenous to Spanish hands and, eventually, to Europe.”

A final observation concerns linguistic diversity. Despite the hypothetical flattening of khipu narration by Spanish interpreters and scribes, the study finds a strong chronological trend in the relative saturation of khipu narrations by different parts of speech. Postconquest accounts were lexically rich but relied on different categories of words—nouns and numbers—than preconquest accounts—adjectives, pronouns, etc.—to achieve their depth. Cords narrated sometimes in the same year told markedly different stories.

In what language did khipukamayuqs narrate their khipus to the Spanish interpreters?

What should be a straightforward question suffers from a lack of information, since the transcriptions often do not state whether the accounts were read out in the “general language” of Peru (i.e. Quechua). In some cases, archival research and ethnographic study can provide confident answers. The Sakaka khipus, which were narrated in Aymara, are perhaps the best example of this strategy (see above). Khipu accounts can plausibly be associated with the regions to which they pertained; I have attempted this in practice with the “language” variable in the PCA implementation, which labels transcriptions from south of Lake Titicaca—historical Aymara territory—as Aymara narrations. However, language of narration is acknowledged to be a legitimate consideration. It is added that the translative process again constitutes a filter for this analysis by unifying multiple steps of interpretation: the output of two-thirds of the participants (the lengua and scribe) was consistently in Spanish. Translation emerges as both the essential challenge of, and a helpful standardizing screen for, interpreting khipu transcriptions. Under this interpretation, Spanish involvement transformed Quechua and Aymara khipu accounts to the

---

point where they even can be directly compared. Pending further research into each khipu transcription’s archival footprint, I supplemented the language proxy variable (i.e. language of narration) with additional measures like latitude and distance from Cuzco to give utmost acknowledgment of this consideration.

Does a khipu transcription corpus flatten cultural context?

This consideration, while important, is not unique to khipu narration. One cannot help but wonder, however, whether an Inka khipukamayuq would approve of the mass compilation of khipu narrations. There is a notable risk that the linguistic “flattening” perpetrated by the Spanish bureaucrats pales in comparison to the level of abstraction—parts of speech, function words, etc.—employed in this study.

The methods I have introduced to khipu transcription studies raise larger questions regarding the appeal of subjecting natural language to statistics and linear algebra. Two observations are offered to end this section. The first is that while corpus linguistics techniques do not make direct use of exegesis, the visualizations they produce can only enhance close reading. The spirit of Inka khipu recording—horizontal comparison of accounts between cord keepers and vertical aggregation of accounts for high-level analysis—is in line with the principles of corpus linguistics and literary computing. Bottom-line totals (in our case, a khipu transcription typology) can be subdivided to consider deeper levels of granularity, whether on a regional (genre) or village (document) basis. The categories generated by statistical methods generate a sharper lens through which to read individual transcriptions. The second observation is that corpus linguistics offers a vocabulary for describing the scale of one’s conclusions. Measures of sampling adequacy (recall Kaiser-Meyer-Olkin from PCA) can be integrated in real-time to bolster arguments which rely on interpretation. Consider the opposite case: Is there not an unfairness in reaching impactful conclusions from close reading, without leveraging existing statistical tools to quantify the confidence of those very results? I present this study, and the subsequent summary of results, as a call to study the Textos Andinos as a balanced and representative sample of samples. The techniques proposed provide an additional method for analyzing the small surviving population of these diverse texts in rigorous and novel ways.
Chapter 6

Summary of Results and Future Directions

6.1 Toward a Statistical Khipu Transcription Typology

This study presents a novel, multivariate approach to generating khipu transcription typologies. The academic study of khipu transcriptions is well-established, offering informative works based in exegesis—close reading and literary analysis—that date back over four decades. Following this tradition, I proposed my own exegetical typology, following a reading of the largest existing corpus of khipu transcriptions. The output of this close reading—a primordial division of the 72 transcriptions—formed the input to a multivariate statistical analysis of the texts based in corpus linguistics techniques.

Principal component analysis (PCA) and agglomerative hierarchical clustering (HAC)—two exploratory multivariate data analysis methods—were utilized to suggest a statistical typology, an empirically-tested framework for evaluating narrative diversity in the early colonial Andes. Crucially, aggregate text analysis required representing the documents by a small set of universal, content-agnostic linguistic features. I chose two measures in line with established corpus linguistics techniques—the relative appearance of frequent lemmas and the distribution of universal parts of speech—as proxies to measure essential narrative infrastructure. These choices of variables are understandably challenging to interpret; I chose coarse, non-contextual measures that abstract the content of any given transcription to study the narrative vehicle itself, while minimizing the influence of Spanish scribal variation. There is, after all, profound meaning when non-contextual measures facilitate contextual inference.

I introduced 13 supplementary variables to test a wide range of potential hypotheses concerning khipu transcription variation. Hypotheses were evaluated using a joint representation of individuals and variables: whether transcriptions (individuals) were differentiated by chronology, altitude, distance from the former Inka capital, etc. (variables). PCA offered two statistically distinguishable typological categories for consideration in clustering khipu transcriptions: chronology and currency. This two-level segmentation corresponds to an axial interpretation of the PCA results. In the first plane, the first principal component significantly separates preconquest from postconquest khipu data. Along the second principal component, non-monetary enumerated lists are significantly distinguished from currency khipus—an
example of Andean accommodation to European valuation and quantification introduced during
the Spanish conquest. While these variables are interdependent, they are also shown to produce
surprising synchronic segmentation. That is, postconquest khipu narrations, in their essential
linguistic structure, are significantly distinguishable by the mere presence or absence of the
Spanish peso. Hierarchical agglomerative clustering, another exploratory multivariate technique,
was applied to the linguistic features to reinforce the axial interpretation found in PCA. Two
dendrograms, indexed hierarchies of ordered groupings (one based on lemmas, the other on parts
of speech), enriched our understanding of the Euclidean cloud of PCA. The two trees were
shown (1) to meaningfully distinguish pre- from postconquest khipu data and (2) to be
significantly faithful to each other, suggesting that their differentiation of pre- and postconquest
accounts bolsters the case for PCA’s axial/chronological segmentation. The mutually-informing
processes, inspired by principles of model parallelization, support the primary division proffered
in the study: that of a significant and meaningful segmentation in the narrative structures of pre-
versus postconquest khipu transcriptions.285

The four linked results (two implementations each of PCA and HAC) were applied in a
case study to provide the first statistical evidence for Urton’s (1998) hypothesis, which posited a
“radical transformation” in khipu narration—namely, a flattening of the expressive capacity of
khipus into ledger books, comprised of mere nouns, numbers and records of payment—following
the Spanish conquest. The testable statements of Urton’s exegetical study were supported
statistically under the chosen parameters, constituting novel evidence for a 20+-year-old theory
(and several of the few testable hypotheses in this subfield of khipu studies). Jointly, the findings
were unified to provide empirically rigorous justifications for the component of Urton’s
hypothesis that cannot be tested statistically: whether 1532 in fact constituted a “radical
transformation” in khipu recording and narration. The importance of this finding attests to the
power of multivariate methods and the wealth of testable hypotheses available in previous
studies of khipus.

The synthesized results demonstrate that four exploratory multivariate implementations
are in accordance with each other and collectively inform a statistical typology, a quadrant-by-
quadrant segmentation of khipu transcriptions into statistically distinguishable categories: (1)
postconquest currency khipus, (2) postconquest enumerated lists, (3) preconquest

285 See Zuidema and de Boer 2013.
narrative/numerical khipus and (4) summary transcriptions. It is acknowledged that my original exegetical typology achieved a mere 25% “success” rate—summary transcriptions emerged as the only category that is significantly distinguishable as a contributor to inter-transcription variance. This is perhaps the most surprising result: that an exegetical typology—grounded in close reading and interpretation—might lack statistical rigor. It is emphasized that the two typologies (exegetical and statistical) are not adversarially opposed. That is, these two results need not be reconciled, but rather synergized. I offer a statistical typology as an exercise in methodological breadth—a cross-disciplinary mapping—whereby statistical techniques expand our *vocabulary of argumentation* to bolster exegetically-derived hypotheses. It is hoped that the tools employed in this study lay the enabling infrastructure for studying khipu transcriptions in the technological age. The two typologies are contrasted directly in figure 6.1, with the concordant category (summary khipus) highlighted in green.

![Figure 6.1: Comparative Khipu Transcription Typologies](image)

A lingering question remains: Do the linked typologies, acknowledging their apparent divergence, justify the preliminary assignment of principles of a khipu transcription “insistence”? That is, do the four archetypes of khipu transcriptions—(1) symmetric repetition, (2) elaboration, (3) accommodation and (4) variable scale—collectively describe the khipu transcription corpus?
The applicability of these principles grows throughout the study. The overarching symmetry of khipu transcriptions—whether postconquest enumerated lists of goods or preconquest lineages of Inka kings—emerges both from individual texts (exegetical typology) and from the compiled corpus (visual output of PCA and HAC; statistical typology). A content-agnostic rhythmic harmony is evident. The separation between preconquest narrative description and postconquest attenuated noun and number phrases highlights the relevance of variable elaboration to the performative act of khipu narration. The expressive capacity of khipus—repositories of numerical and linguistic information—was flattened in the sixteenth century once twisted cords were put on trial, queried like indigenous SQL databases. These moments of elicitation were projected against the overarching linguistic fracture introduced in 1532, separated not only along the axes of dendrograms or biplots, but in the knotted vocalizations of khipu cords themselves. We might distinguish the lexical richness of the Cuzco ceque system from rote testimony in the Sakaka restitution trial along an axis of elaboration versus summation: the long march of khipus from narratives to balance sheets. An added testament to khipu recording is the ease with which the script accommodated Spanish currency, a system of valuation not previously employed in precolumbian South America. While the introduction of prices and currency indeed radically altered the structure of khipu narration—saturating performance with nouns and numbers of pigs, wheat and other newly introduced things—khipus possessed more-than-enough expressive capacity to incorporate Spanish invoices, while preserving numerical accuracy at highly variable scales. Indeed, the khipu testimonies attest to the quantitative breadth of khipus as repositories of mathematical results. These results often subdivided massive totals of regional population or items given to Spanish bureaucrats, recorded efficiently in knotted cords.

In summary, these principles of khipu transcription insistence emerge as the “defining cultural coherencies” of narrated khipu cords—idiosyncratic mannerisms identified using close reading and statistical analysis techniques on the largest corpus of khipu transcriptions.286 These principles, in addition to the issues proposed below, ought to guide future research on this extraordinary body of textual data.

---

6.2 Implications for the Future: Outlining a Formal Khipu Transcription Corpus

This penultimate section charts the study of khipu transcriptions in the digital age with guiding principles for a Khipu Transcription Corpus (KTC)—a novel online repository of early colonial khipu transcriptions. This proposal is inspired by the Khipu Database (KDB), hosted in the Department of Anthropology at Harvard. The project, initiated in 2002, but itself motivated by the pioneering khipu cataloging work of the Aschers since the 1960s, aims to consolidate fully descriptive observational entries (i.e. knots, cord colors, spin direction, etc.) for the full body of archaeological khipus in a publicly accessible online repository. In practice, the project includes a public interface where users from around the world can download excel spreadsheets pertaining to individual khipus, as well as the full relational database in the form of MySQL data tables.

My own KTC proposal builds on the structure of the KDB to accommodate natural language data. A precondition for public accessibility is making available annotated, internally-represented XML files for each transcription that include ample metadata. It is anticipated that the XML files used in this study will soon be made publicly available, pending further analysis of publication rights and licensing agreements. The KTC interface will include both individually-downloadable files, as well as the compiled corpus. Transcriptions will be available in both XML (.xml) and UTF-8 plain text (.txt) formats, which can be edited and opened with any text editor (e.g. Notepad++). The interface may incorporate corpus visualization tools, including native web browser searches and summary statistics with dynamic graphs (e.g. frequent tokens, word collocations, etc.). A research bibliography and linked publications will further integrate the reader into the state of the art. In their implementations, the KTC and KDB are unified by guiding principles of khipu studies:

- Public accessibility is linked to equity. The beauty of the KDB is its universality: any researcher who is curious to test a particular hypothesis can download the relational database to analyze over 50,000 rows of numerical cord data directly. This accessibility imperative is not new to the field: even during its development in 2003, Urton emphasized that the nascent database would soon allow others to test his and other researchers’ theories. A discipline borne out of colonization should be accessible to all.

---

287 Urton, “Khipu Database Project.”
Digitization is particularly important so that the transcriptions can outlive the paper on which they are scrawled. The annotated text—itself subject to open-source iterations and academic study—will survive for future research.

- Digitization facilitates continual addition of new data. In the context of archaeological khipus, numerical cord information, provenience details and color information can be added to the KDB, which follows the standard notational system first formalized by the Aschers. The KTC will expand with the addition of newly discovered khipu transcriptions, which will be syntactically annotated and included in the corpus in XML and plain text formats. Digital availability itself is powerful: historical Spanish in particular lacks the wealth of open-source textual corpora that are available to students of Old English, Greek or Latin, for example.\(^{289}\) The KTC, while not contributing millions or billions of tokens, offers further granularity to the body of historical Spanish texts available to researchers around the world. It is for this reason that neither the KDB nor the proposed KTC are static databases. Many known museum khipus around the world remain uncatalogued by the KDB; new archaeological khipus have been excavated and entered into the KDB within the last five years. The quantity of unpublished khipu transcription documents is no doubt sizeable, many likely hidden in scores of uncatalogued legajos [“files”] in the AGI.\(^{290}\) The centralized digital compilation of writings from and about khipus promises a self-energizing research corpus with substantial academic potential.

- Interactive data facilitates public outreach. Centrally available khipu transcriptions open exciting new avenues for the visualization of khipu data. The thousands of counts and quantities transcribed in enumerated lists (e.g. 24 llamas, 134 pesos, etc.) can be checked in aggregate against the numerical data in the KDB to search for matches and near-matches among archaeological khipus—to widen the set of possible candidates for a “Rosetta khipu.” As the 200-year anniversary of Champollion’s decipherment of hieroglyphs approaches, this possibility is particularly exciting. Additional versions of the documents will explicitly aid outreach: full English translations will open the field to study by a wider range of researchers, while an interactive map of South America,

---
\(^{289}\) Sánchez-Marco, Fontana and Domingo, “Anotación automática.”
\(^{290}\) Gary Urton, personal communication, March 15, 2019.
modeled as an evolved version of figure 2.1, could have clickable icons identifying each transcription with its original location, year and topic. Embedded document visualization tools and three-dimensional khipu renderings (in conjunction with the KDB) offer additional inroads for making khipu transcriptions, and khipu studies more broadly, responsive to the digital age.

The KTC is outlined as a dual project. The first is to make sustainable the continual, publicly-informed study of this extraordinary collection of documents. In practice, this involves the standardization of the digitized transcriptions, hosted in an easily expandable medium. The sheer volume of text offers inroads to exegetical, statistical and parallel approaches to khipu transcription studies. The second is to facilitate novel studies while remaining attuned to public outreach. The implementations of PCA and HAC offered here are testament to the power that emanates from even modestly-sized corpora of natural language. The basic compilation of textual data about khipus—an imperative over 15 years in the making in the case of digitized KDB data—facilitates explicit hypothesis testing of the wealth of available research on khipus based in close reading and literary analysis. At the same time, interactive content and bilingual translation offer near-horizon goals to accompany the online publication of the KTC—informing by the guiding principles of khipu transcription typologies.
Chapter 7

Conclusion: Toward a Khipu Transcription “Insistence”

7.1 Final Remarks

This aggregate study of khipu transcriptions capitalizes on academic energy built up over the past decades. In 1998, Urton offered an optimistic message: “It is hoped that, in time, more khipu transcriptions will become available from the stores of Spanish documents that remain in Seville and elsewhere.” Following a close reading of four khipu transcriptions from legal proceedings, Loza (2001) motivated robust corpus compilation, choosing transcriptions from different decades and areas of the Andes to “ensure a sufficiently wide empirical baseline.” In a subsequent critique of Loza’s study, Brokaw (2013) argued that “the number of direct khipu transcription documents that Loza presents is quite limited… [especially when] a broad corpus has now been published of judicial-administrative documents that, while being less significant, have numerous references to khipus.” These excerpts are unified by the recognition of the importance of sample size when studying khipu transcriptions. However, this study is offered in part as a retort to Brokaw’s lukewarm acknowledgment of the Pärssinen/Kiviharju corpus. The minutiae of khipu transcriptions—administrative proceedings and enumerated lists, detailing llamas, chuño and wheat—operated within a linguistic infrastructure brought to life with each public narration of khipu cords. The “less significant” documents of the corpus created in this study are precisely its strength: the study of khipu grammars comes to life through the aggregate and content-agnostic study of surviving transcriptions. Brokaw’s assignment of relative significance to the transcriptions is symptomatic of the pitfalls of our natural human “content imperative.” How can researchers begin a typology of khipu transcriptions when some transcriptions are deemed “less significant” before the analysis has even begun?

Corpus linguistics offers a novel technique for constructing a primordial division of khipu texts—bypassing contextual assumptions to form a nascent transcription typology. The introduction of exploratory multivariate statistics provides novel answers that both support previous theories—Urton’s 1998 hypothesis among them—as well as inform novel questions and

291 Urton, “From Knots to Narratives,” 431.
293 Brokaw, “La recepción del quipu,” 121-122; my translation, emphasis added.
principles for a khipu transcription insistence. Exegetical and statistical typologies were
introduced in close succession to one another to further bolster the case for a synergistic and
parallel approach to khipu transcription studies.

The techniques in this study capitalized on the syntactic annotation of the largest existing
corpus of digitized colonial-era khipu transcriptions. By introducing an explicitly comparative
analysis of close reading and document aggregation, it is hoped that this study has established a
path forward that is mutually informing, the enabling infrastructure laid for dialogues between
the archaeologist and the linguist; the statistician and the historian. Noting the promise of recent
statistical studies of khipus, the study motivates computational approaches to natural language
processing in khipu transcription studies.\textsuperscript{294} The analysis forms a call for textual statistics—
grounded in historical corpus linguistics techniques—in the study of early colonial khipu texts.

Crucially, this study returned the focal point of attention to the primary text of the khipu
transcriptions themselves. Over 70,000 words were compiled to capture the largest surviving
corpus of khipu narration—royal lineages and holy shrines, the camelids of Lake Titicaca and
disputed payments under corrupt Spanish encomenderos. To further illustrate the idiosyncrasies
of these wonderfully diverse texts, the Aschers’ concept of “insistence” was deployed, informed
by aggregate analysis of the texts as products of organic khipu “writing events.” Syntactic
annotation and rich visualizations focused our attention on the narrative infrastructure of khipus.
Blueprints for an expandable and enriched Khipu Transcription Corpus (KTC) were introduced,
to preserve these transcriptions beyond the life of their paper hosts and to facilitate both
exegetical and statistical research in the digital age. This study has emphasized the primacy of
the khipu transcription text, to hear the echoes of khipukamayuqs within the garble of Old
Spanish scribal variation. Indeed, within this digital corpus survive the only written recordings of
the vocal infrastructure of early colonial khipu narration. A corpus-based approach to these
distressingly understudied texts motivates a dual emphasis on khipu preservation, both in and out
of the archive. While archaeological khipus are “imprisoned in museums worldwide,” their
voices and testimonies survive in fading ink and moldering portfolios.\textsuperscript{295} We ought to listen to
their stories.

\textsuperscript{294} See Clindaniel 2019.
\textsuperscript{295} John V. Murra, “The Ethno-categories of a Regional Khipu,” partial English translation of Murra 1975,
mimeographed, revised and updated, 1990 [1975], 10.
Bibliography


Pizarro, Hernando. “A los Señores Oydores de la Audiencia Real de Su Magestad.” In Informaciones sobre el antiguo Perú, edited by Horacio H. Urteaga, 16-180. Colección


Appendix A

Supplementary Data and Figures

A.1 Principal Component Analysis
A.1.1 Lemma Implementation
A.1.1.1 Factor Loadings and Contributions

FactoMineR’s dimdesc( ) function natively identifies the variables (quantitative and qualitative) that are most significantly associated with the specified principal components at a given significance threshold. In other words, the function provides an automatic description of the specified dimensions. I include the full significant variable readout for the 1% significance threshold and dimensions 1-4 below. The resulting list is sorted within each dimension by sign and p-value: categories most significantly associated with positive values of the PC and those most significantly associated with negative values of the PC. The result is an ordered list which can be interpreted to assess the levels which are most important to each side (positive and negative) of the specified dimension.

```r
# print significant factor loadings using dimdesc
dimdesc(mem_tok.pca,axes=c(1:4),proba = 0.01)
```

Dimension 1

$Dim.1$quanti

<table>
<thead>
<tr>
<th></th>
<th>correlation</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>otro</td>
<td>0.8568800</td>
<td>7.937972e-22</td>
</tr>
<tr>
<td>que</td>
<td>0.8409736</td>
<td>2.391553e-20</td>
</tr>
<tr>
<td>pueblo</td>
<td>0.7880284</td>
<td>2.140485e-16</td>
</tr>
<tr>
<td>con</td>
<td>0.6922029</td>
<td>1.645761e-11</td>
</tr>
<tr>
<td>él</td>
<td>0.6764361</td>
<td>6.964282e-11</td>
</tr>
<tr>
<td>indio</td>
<td>0.5360868</td>
<td>1.214978e-06</td>
</tr>
<tr>
<td>en</td>
<td>0.3458246</td>
<td>2.925561e-03</td>
</tr>
<tr>
<td>dicho</td>
<td>-0.3155993</td>
<td>6.924136e-03</td>
</tr>
<tr>
<td>a</td>
<td>-0.3985278</td>
<td>5.257860e-04</td>
</tr>
<tr>
<td>dar</td>
<td>-0.4287687</td>
<td>1.714307e-04</td>
</tr>
<tr>
<td>year</td>
<td>-0.4664240</td>
<td>3.634754e-05</td>
</tr>
<tr>
<td>de</td>
<td>-0.6225589</td>
<td>5.264156e-09</td>
</tr>
</tbody>
</table>

$Dim.1$quali

<table>
<thead>
<tr>
<th></th>
<th>R2</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>num_speakers_sng_mult</td>
<td>0.6683541</td>
<td>2.905862e-17</td>
</tr>
<tr>
<td>type</td>
<td>0.6705498</td>
<td>2.208117e-16</td>
</tr>
<tr>
<td>pre_post_conquest</td>
<td>0.4956753</td>
<td>5.237919e-12</td>
</tr>
<tr>
<td>currency</td>
<td>0.2352992</td>
<td>1.571881e-05</td>
</tr>
<tr>
<td>pop_s_m_l</td>
<td>0.2517632</td>
<td>4.511774e-05</td>
</tr>
</tbody>
</table>

$Dim.1$category

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>p.value</th>
</tr>
</thead>
</table>


Dimension 2

<table>
<thead>
<tr>
<th>Dimension 2</th>
<th>$\text{Dim.2}_\text{quanti}$</th>
<th>correlation</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dicho</td>
<td>0.6058580</td>
<td>1.710830e-08</td>
</tr>
<tr>
<td></td>
<td>a</td>
<td>0.5939210</td>
<td>3.813840e-08</td>
</tr>
<tr>
<td></td>
<td>ser</td>
<td>0.4893488</td>
<td>1.288269e-05</td>
</tr>
<tr>
<td></td>
<td>el</td>
<td>0.4552584</td>
<td>5.868094e-05</td>
</tr>
<tr>
<td></td>
<td>lon</td>
<td>0.3731626</td>
<td>1.244665e-03</td>
</tr>
<tr>
<td></td>
<td>él</td>
<td>0.3513445</td>
<td>2.476567e-03</td>
</tr>
<tr>
<td></td>
<td>uno</td>
<td>0.3163591</td>
<td>6.782717e-03</td>
</tr>
<tr>
<td></td>
<td>lat</td>
<td>-0.3132409</td>
<td>7.379674e-03</td>
</tr>
<tr>
<td></td>
<td>y</td>
<td>-0.4515880</td>
<td>6.844124e-05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimension 2</th>
<th>$\text{Dim.2}_\text{quali}$</th>
<th>R2</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>lat_zone</td>
<td>0.2166148</td>
<td>0.000219862</td>
</tr>
<tr>
<td></td>
<td>currency</td>
<td>0.1264441</td>
<td>0.002174317</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimension 2</th>
<th>$\text{Dim.2}_\text{category}$</th>
<th>Estimate</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>south</td>
<td>0.9359808</td>
<td>5.240756e-05</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>0.5563034</td>
<td>2.174317e-03</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>-0.5563034</td>
<td>2.174317e-03</td>
</tr>
</tbody>
</table>

Dimension 3

<table>
<thead>
<tr>
<th>Dimension 3</th>
<th>$\text{Dim.3}_\text{quanti}$</th>
<th>correlation</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>el</td>
<td>0.5085984</td>
<td>5.081060e-06</td>
</tr>
<tr>
<td></td>
<td>indio</td>
<td>0.4295398</td>
<td>1.663682e-04</td>
</tr>
<tr>
<td></td>
<td>de</td>
<td>0.4050288</td>
<td>4.168794e-04</td>
</tr>
<tr>
<td></td>
<td>pueblo</td>
<td>0.3302703</td>
<td>4.606117e-03</td>
</tr>
<tr>
<td></td>
<td>ser</td>
<td>0.3053023</td>
<td>9.112097e-03</td>
</tr>
<tr>
<td></td>
<td>él</td>
<td>-0.4924815</td>
<td>1.11487e-05</td>
</tr>
<tr>
<td></td>
<td>dar</td>
<td>-0.4996392</td>
<td>7.889574e-06</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dimension 3</th>
<th>$\text{Dim.3}_\text{category}$</th>
<th>Estimate</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mult</td>
<td>-0.5876371</td>
<td>0.009666627</td>
</tr>
</tbody>
</table>

Dimension 4
$\text{Dim. 4}$ quanti

<table>
<thead>
<tr>
<th></th>
<th>correlation</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ser</td>
<td>0.4598637</td>
<td>4.825833e-05</td>
</tr>
<tr>
<td>el</td>
<td>0.3799750</td>
<td>9.941473e-04</td>
</tr>
<tr>
<td>en</td>
<td>-0.3902287</td>
<td>7.023406e-04</td>
</tr>
<tr>
<td>uno</td>
<td>-0.6779967</td>
<td>6.061527e-11</td>
</tr>
</tbody>
</table>

Table A.1: Significant Factor Loadings (Lemma PCA Implementation)

I also include the full list of active variable contributions below.

<table>
<thead>
<tr>
<th></th>
<th>Dim.1</th>
<th>Dim.2</th>
<th>Dim.3</th>
<th>Dim.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>el</td>
<td>1.864787</td>
<td>10.55268</td>
<td>15.05422</td>
<td>11.13282</td>
</tr>
<tr>
<td>a</td>
<td>3.610649</td>
<td>17.95991</td>
<td>5.064763</td>
<td>1.514207</td>
</tr>
<tr>
<td>dar</td>
<td>4.179403</td>
<td>1.447135</td>
<td>14.52853</td>
<td>1.790456</td>
</tr>
<tr>
<td>de</td>
<td>8.811075</td>
<td>5.824257</td>
<td>9.547292</td>
<td>2.782592</td>
</tr>
<tr>
<td>dicho</td>
<td>2.264331</td>
<td>18.6891</td>
<td>1.239201</td>
<td>2.900387</td>
</tr>
<tr>
<td>él</td>
<td>10.40211</td>
<td>6.285111</td>
<td>14.11524</td>
<td>4.012391</td>
</tr>
<tr>
<td>en</td>
<td>2.718817</td>
<td>4.579888</td>
<td>2.067618</td>
<td>11.74177</td>
</tr>
<tr>
<td>indio</td>
<td>6.533385</td>
<td>0.992615</td>
<td>10.7378</td>
<td>0.267017</td>
</tr>
<tr>
<td>otro</td>
<td>16.69199</td>
<td>0.776528</td>
<td>2.436544</td>
<td>2.328292</td>
</tr>
<tr>
<td>que</td>
<td>16.07803</td>
<td>1.996803</td>
<td>4.047148</td>
<td>2.157302</td>
</tr>
<tr>
<td>ser</td>
<td>0.297831</td>
<td>12.19225</td>
<td>5.424613</td>
<td>16.30623</td>
</tr>
<tr>
<td>uno</td>
<td>1.496794</td>
<td>5.095743</td>
<td>2.416489</td>
<td>35.44463</td>
</tr>
<tr>
<td>y</td>
<td>0.040806</td>
<td>10.38321</td>
<td>2.566206</td>
<td>7.011554</td>
</tr>
<tr>
<td>con</td>
<td>10.89268</td>
<td>1.048891</td>
<td>4.406178</td>
<td>0.150414</td>
</tr>
<tr>
<td>pueblo</td>
<td>14.11731</td>
<td>2.175881</td>
<td>6.348157</td>
<td>0.459938</td>
</tr>
</tbody>
</table>

Table A.2: Active Variable Contributions (Lemma PCA Implementation)
The full variable correlation matrix, which serves as the basis for the PCA implementation, is reproduced below.

<table>
<thead>
<tr>
<th></th>
<th>cl</th>
<th>a</th>
<th>dar</th>
<th>de</th>
<th>dicho</th>
<th>él</th>
<th>en</th>
<th>indio</th>
<th>otro</th>
<th>que</th>
<th>ser</th>
<th>uno</th>
<th>y</th>
<th>con</th>
<th>pueblo</th>
</tr>
</thead>
<tbody>
<tr>
<td>cl</td>
<td>1</td>
<td>0.180</td>
<td>-0.022</td>
<td>0.320</td>
<td>0.430</td>
<td>-0.138</td>
<td>0.054</td>
<td>-0.062</td>
<td>-0.282</td>
<td>-0.170</td>
<td>0.310</td>
<td>-0.047</td>
<td>-0.110</td>
<td>-0.033</td>
<td>-0.114</td>
</tr>
<tr>
<td>a</td>
<td>0.180</td>
<td>1</td>
<td>0.287</td>
<td>-0.094</td>
<td>0.288</td>
<td>0.004</td>
<td>-0.099</td>
<td>-0.300</td>
<td>-0.377</td>
<td>-0.182</td>
<td>0.096</td>
<td>0.119</td>
<td>-0.114</td>
<td>-0.191</td>
<td>-0.498</td>
</tr>
<tr>
<td>dar</td>
<td>-0.022</td>
<td>0.287</td>
<td>1</td>
<td>0.098</td>
<td>0.148</td>
<td>0.079</td>
<td>-0.106</td>
<td>-0.189</td>
<td>-0.340</td>
<td>-0.246</td>
<td>-0.087</td>
<td>-0.284</td>
<td>-0.014</td>
<td>-0.293</td>
<td>-0.474</td>
</tr>
<tr>
<td>de</td>
<td>0.320</td>
<td>-0.094</td>
<td>0.098</td>
<td>1</td>
<td>0.011</td>
<td>-0.566</td>
<td>-0.257</td>
<td>-0.183</td>
<td>-0.494</td>
<td>-0.599</td>
<td>-0.034</td>
<td>-0.222</td>
<td>0.105</td>
<td>-0.358</td>
<td>-0.263</td>
</tr>
<tr>
<td>dicho</td>
<td>0.430</td>
<td>0.288</td>
<td>0.148</td>
<td>0.011</td>
<td>1</td>
<td>-0.064</td>
<td>0.098</td>
<td>-0.196</td>
<td>-0.321</td>
<td>-0.165</td>
<td>0.179</td>
<td>-0.046</td>
<td>-0.112</td>
<td>-0.106</td>
<td>-0.223</td>
</tr>
<tr>
<td>él</td>
<td>-0.138</td>
<td>0.004</td>
<td>0.079</td>
<td>-0.566</td>
<td>-0.064</td>
<td>1</td>
<td>0.159</td>
<td>0.078</td>
<td>0.682</td>
<td>0.812</td>
<td>0.161</td>
<td>0.110</td>
<td>-0.029</td>
<td>0.389</td>
<td>0.306</td>
</tr>
<tr>
<td>en</td>
<td>0.054</td>
<td>-0.099</td>
<td>-0.106</td>
<td>-0.257</td>
<td>0.098</td>
<td>0.159</td>
<td>1</td>
<td>0.128</td>
<td>0.162</td>
<td>0.252</td>
<td>-0.019</td>
<td>0.284</td>
<td>-0.161</td>
<td>0.222</td>
<td>0.277</td>
</tr>
<tr>
<td>indio</td>
<td>-0.062</td>
<td>-0.300</td>
<td>-0.189</td>
<td>-0.183</td>
<td>-0.196</td>
<td>0.078</td>
<td>0.128</td>
<td>1</td>
<td>0.326</td>
<td>0.242</td>
<td>0.123</td>
<td>0.091</td>
<td>-0.088</td>
<td>0.521</td>
<td>0.566</td>
</tr>
<tr>
<td>otro</td>
<td>-0.282</td>
<td>-0.377</td>
<td>-0.340</td>
<td>-0.494</td>
<td>-0.321</td>
<td>0.682</td>
<td>0.162</td>
<td>0.326</td>
<td>1</td>
<td>0.750</td>
<td>0.023</td>
<td>0.035</td>
<td>0.075</td>
<td>0.453</td>
<td>0.612</td>
</tr>
<tr>
<td>que</td>
<td>-0.170</td>
<td>-0.182</td>
<td>-0.246</td>
<td>-0.599</td>
<td>-0.165</td>
<td>0.812</td>
<td>0.252</td>
<td>0.242</td>
<td>0.750</td>
<td>1</td>
<td>0.161</td>
<td>0.125</td>
<td>-0.039</td>
<td>0.434</td>
<td>0.565</td>
</tr>
<tr>
<td>ser</td>
<td>0.310</td>
<td>0.096</td>
<td>-0.087</td>
<td>-0.034</td>
<td>0.179</td>
<td>0.161</td>
<td>-0.019</td>
<td>0.123</td>
<td>0.023</td>
<td>0.161</td>
<td>1</td>
<td>0.046</td>
<td>-0.153</td>
<td>0.208</td>
<td>0.036</td>
</tr>
<tr>
<td>uno</td>
<td>-0.047</td>
<td>0.119</td>
<td>-0.284</td>
<td>-0.222</td>
<td>-0.046</td>
<td>0.110</td>
<td>0.284</td>
<td>0.091</td>
<td>0.035</td>
<td>0.125</td>
<td>0.046</td>
<td>1</td>
<td>-0.199</td>
<td>0.250</td>
<td>0.114</td>
</tr>
<tr>
<td>y</td>
<td>-0.110</td>
<td>-0.114</td>
<td>-0.014</td>
<td>0.105</td>
<td>-0.112</td>
<td>-0.029</td>
<td>-0.161</td>
<td>-0.088</td>
<td>0.075</td>
<td>-0.039</td>
<td>-0.153</td>
<td>-0.199</td>
<td>1</td>
<td>-0.065</td>
<td>0.008</td>
</tr>
<tr>
<td>con</td>
<td>-0.033</td>
<td>-0.191</td>
<td>-0.293</td>
<td>-0.358</td>
<td>-0.106</td>
<td>0.389</td>
<td>0.222</td>
<td>0.521</td>
<td>0.453</td>
<td>0.434</td>
<td>0.208</td>
<td>0.250</td>
<td>-0.065</td>
<td>1</td>
<td>0.536</td>
</tr>
<tr>
<td>pueblo</td>
<td>-0.114</td>
<td>-0.498</td>
<td>-0.474</td>
<td>-0.263</td>
<td>-0.223</td>
<td>0.306</td>
<td>0.277</td>
<td>0.566</td>
<td>0.612</td>
<td>0.565</td>
<td>0.036</td>
<td>0.114</td>
<td>0.008</td>
<td>0.536</td>
<td>1</td>
</tr>
</tbody>
</table>

Table A.3: Variable Correlation Matrix (Lemma PCA Implementation)

I use the PerformanceAnalytics package to display the significance of the correlation coefficients among the active variables. This visualization enriches the correlation matrix. Above the diagonal are the absolute values of the correlation coefficients (from table A.3), including statistical significance levels derived from a correlation test expressed with asterisks (* = $p < 0.05$; ** = $p < 0.01$; *** = $p < 0.001$). Along the diagonal are the univariate distributions, in the form of histograms with kernel density overlays. Below the diagonal are bivariate scatter plots with fitted lines under locally estimated scatterplot smoothing.
Figure A.1: Variable Correlation Significance Plot (Lemma Implementation)

The full individual contributions matrix is reproduced below. These values are used to construct figure 4.4, the plot of individuals.

<table>
<thead>
<tr>
<th></th>
<th>Dim.1</th>
<th>Dim.2</th>
<th>Dim.3</th>
<th>Dim.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>memoria_01</td>
<td>0.692923</td>
<td>1.594641</td>
<td>0.304675</td>
<td>1.84486</td>
</tr>
<tr>
<td>memoria_02</td>
<td>1.382774</td>
<td>0.020083</td>
<td>0.064662</td>
<td>0.001115</td>
</tr>
<tr>
<td>memoria_03</td>
<td>0.041616</td>
<td>5.581131</td>
<td>1.607687</td>
<td>0.02953</td>
</tr>
<tr>
<td>memoria_04</td>
<td>1.14101</td>
<td>0.003419</td>
<td>0.050035</td>
<td>0.001214</td>
</tr>
<tr>
<td>memoria_05</td>
<td>Dim.1</td>
<td>Dim.2</td>
<td>Dim.3</td>
<td>Dim.4</td>
</tr>
<tr>
<td>--------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>memoria_06</td>
<td>0.552236</td>
<td>0.090428</td>
<td>7.050189</td>
<td>1.20616</td>
</tr>
<tr>
<td>memoria_07</td>
<td>0.488851</td>
<td>1.179083</td>
<td>0.05996</td>
<td>1.370645</td>
</tr>
<tr>
<td>memoria_08</td>
<td>0.968627</td>
<td>0.030316</td>
<td>4.559696</td>
<td>0.466551</td>
</tr>
<tr>
<td>memoria_09</td>
<td>1.66892</td>
<td>2.971271</td>
<td>0.01505</td>
<td>1.283531</td>
</tr>
<tr>
<td>memoria_10</td>
<td>1.326913</td>
<td>1.979748</td>
<td>0.320481</td>
<td>2.600004</td>
</tr>
<tr>
<td>memoria_11</td>
<td>1.79906</td>
<td>3.365871</td>
<td>0.020978</td>
<td>0.951648</td>
</tr>
<tr>
<td>memoria_12</td>
<td>0.673606</td>
<td>2.549578</td>
<td>2.202292</td>
<td>0.827865</td>
</tr>
<tr>
<td>memoria_13</td>
<td>0.552795</td>
<td>0.691883</td>
<td>6.04003</td>
<td>0.012594</td>
</tr>
<tr>
<td>memoria_14</td>
<td>1.223951</td>
<td>0.680419</td>
<td>2.30627</td>
<td>2.920292</td>
</tr>
<tr>
<td>memoria_15</td>
<td>1.865561</td>
<td>0.096575</td>
<td>0.030316</td>
<td>1.20616</td>
</tr>
<tr>
<td>memoria_16</td>
<td>0.276295</td>
<td>0.429206</td>
<td>0.003692</td>
<td>2.600004</td>
</tr>
<tr>
<td>memoria_17</td>
<td>1.582852</td>
<td>12.69463</td>
<td>0.122845</td>
<td>3.04792</td>
</tr>
<tr>
<td>memoria_18</td>
<td>0.673606</td>
<td>2.549578</td>
<td>1.766262</td>
<td>0.012594</td>
</tr>
<tr>
<td>memoria_19</td>
<td>1.223951</td>
<td>0.680419</td>
<td>2.30627</td>
<td>2.920292</td>
</tr>
<tr>
<td>memoria_20</td>
<td>1.277279</td>
<td>2.793157</td>
<td>0.049535</td>
<td>0.019269</td>
</tr>
<tr>
<td>memoria_21</td>
<td>1.085776</td>
<td>0.73452</td>
<td>0.003692</td>
<td>2.600004</td>
</tr>
<tr>
<td>memoria_22</td>
<td>0.840031</td>
<td>0.900455</td>
<td>2.069259</td>
<td>2.059321</td>
</tr>
<tr>
<td>memoria_23</td>
<td>0.726716</td>
<td>2.943664</td>
<td>3.340541</td>
<td>0.74202</td>
</tr>
<tr>
<td>memoria_24</td>
<td>0.552236</td>
<td>0.090428</td>
<td>7.050189</td>
<td>1.20616</td>
</tr>
<tr>
<td>memoria_25</td>
<td>0.488851</td>
<td>1.179083</td>
<td>0.05996</td>
<td>1.370645</td>
</tr>
<tr>
<td>memoria_26</td>
<td>0.968627</td>
<td>0.030316</td>
<td>4.559696</td>
<td>0.466551</td>
</tr>
<tr>
<td>memoria_27</td>
<td>1.66892</td>
<td>2.971271</td>
<td>0.01505</td>
<td>1.283531</td>
</tr>
<tr>
<td>memoria_28</td>
<td>1.326913</td>
<td>1.979748</td>
<td>0.320481</td>
<td>2.600004</td>
</tr>
<tr>
<td>memoria_29</td>
<td>1.79906</td>
<td>3.365871</td>
<td>0.020978</td>
<td>0.951648</td>
</tr>
<tr>
<td>memoria_30</td>
<td>0.673606</td>
<td>2.549578</td>
<td>1.766262</td>
<td>0.012594</td>
</tr>
<tr>
<td>memoria_31</td>
<td>1.223951</td>
<td>0.680419</td>
<td>2.30627</td>
<td>2.920292</td>
</tr>
<tr>
<td>memoria_32</td>
<td>1.865561</td>
<td>0.096575</td>
<td>0.030316</td>
<td>1.20616</td>
</tr>
<tr>
<td>memoria_33</td>
<td>0.276295</td>
<td>0.429206</td>
<td>0.003692</td>
<td>2.600004</td>
</tr>
<tr>
<td>memoria_34</td>
<td>1.582852</td>
<td>12.69463</td>
<td>0.122845</td>
<td>3.04792</td>
</tr>
<tr>
<td>memoria_35</td>
<td>0.673606</td>
<td>2.549578</td>
<td>1.766262</td>
<td>0.012594</td>
</tr>
<tr>
<td>memoria_36</td>
<td>1.223951</td>
<td>0.680419</td>
<td>2.30627</td>
<td>2.920292</td>
</tr>
<tr>
<td>memoria_37</td>
<td>1.865561</td>
<td>0.096575</td>
<td>0.030316</td>
<td>1.20616</td>
</tr>
<tr>
<td>memoria_38</td>
<td>0.276295</td>
<td>0.429206</td>
<td>0.003692</td>
<td>2.600004</td>
</tr>
<tr>
<td>memoria_39</td>
<td>1.582852</td>
<td>12.69463</td>
<td>0.122845</td>
<td>3.04792</td>
</tr>
<tr>
<td>memoria_40</td>
<td>0.673606</td>
<td>2.549578</td>
<td>1.766262</td>
<td>0.012594</td>
</tr>
<tr>
<td>memoria_41</td>
<td>1.223951</td>
<td>0.680419</td>
<td>2.30627</td>
<td>2.920292</td>
</tr>
<tr>
<td>memoria_42</td>
<td>1.865561</td>
<td>0.096575</td>
<td>0.030316</td>
<td>1.20616</td>
</tr>
<tr>
<td>memoria_43</td>
<td>0.276295</td>
<td>0.429206</td>
<td>0.003692</td>
<td>2.600004</td>
</tr>
<tr>
<td>memoria_44</td>
<td>1.582852</td>
<td>12.69463</td>
<td>0.122845</td>
<td>3.04792</td>
</tr>
<tr>
<td>memoria_45</td>
<td>0.673606</td>
<td>2.549578</td>
<td>1.766262</td>
<td>0.012594</td>
</tr>
<tr>
<td>memoria_46</td>
<td>1.223951</td>
<td>0.680419</td>
<td>2.30627</td>
<td>2.920292</td>
</tr>
<tr>
<td>memoria_47</td>
<td>1.865561</td>
<td>0.096575</td>
<td>0.030316</td>
<td>1.20616</td>
</tr>
<tr>
<td>memoria_48</td>
<td>0.276295</td>
<td>0.429206</td>
<td>0.003692</td>
<td>2.600004</td>
</tr>
<tr>
<td>memoria_49</td>
<td>1.582852</td>
<td>12.69463</td>
<td>0.122845</td>
<td>3.04792</td>
</tr>
<tr>
<td>memoria_50</td>
<td>0.673606</td>
<td>2.549578</td>
<td>1.766262</td>
<td>0.012594</td>
</tr>
<tr>
<td>memoria_51</td>
<td>1.223951</td>
<td>0.680419</td>
<td>2.30627</td>
<td>2.920292</td>
</tr>
<tr>
<td>memoria_52</td>
<td>1.865561</td>
<td>0.096575</td>
<td>0.030316</td>
<td>1.20616</td>
</tr>
<tr>
<td>memoria_30</td>
<td>memoria_31</td>
<td>memoria_32</td>
<td>memoria_33</td>
<td>memoria_34</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>Dim.1</td>
<td>Dim.2</td>
<td>Dim.3</td>
<td>Dim.4</td>
<td>Dim.1</td>
</tr>
<tr>
<td>4.947903</td>
<td>2.158661</td>
<td>3.624013</td>
<td>1.090309</td>
<td>10.33447</td>
</tr>
<tr>
<td>memoria_55</td>
<td>memoria_56</td>
<td>memoria_57</td>
<td>memoria_58</td>
<td>memoria_59</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>4.34E-06</td>
<td>0.037657</td>
<td>0.914125</td>
<td>1.3906</td>
<td>0.536016</td>
</tr>
<tr>
<td>3.918128</td>
<td>0.290383</td>
<td>1.609885</td>
<td>0.000369</td>
<td>0.262802</td>
</tr>
<tr>
<td>2.24389</td>
<td>0.00125</td>
<td>2.665006</td>
<td>0.259045</td>
<td>2.631355</td>
</tr>
<tr>
<td>0.880213</td>
<td>0.040803</td>
<td>0.307317</td>
<td>0.21445</td>
<td>0.050592</td>
</tr>
</tbody>
</table>

Table A.4: Individual Contributions (Lemma PCA Implementation)
A.1.2 UPOS Implementation
A.1.2.1 Factor Loadings and Contributions

I include the significant variable readout for the 1% threshold and dimensions 1-2 below. Recall that the resulting list is sorted within each dimension by sign and p-value: categories most significantly associated with positive values of the PC and those most significantly associated with negative values of the PC. Two dimensions are studied (rather than four, as in the lemma implementation), following the results of the various sufficiency tests (i.e. Kaiser’s criterion; Loewen and Gonulal’s criterion; scree plot elbow and parallel analysis).

```r
# print significant factor loadings using dimdesc
dimdesc(mem_POS.pca, axes = c(1, 2), proba = 0.01)
```
### Dimension 1

<table>
<thead>
<tr>
<th></th>
<th>correlation</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRON</td>
<td>0.8433218</td>
<td>1.481748e-20</td>
</tr>
<tr>
<td>SCONJ</td>
<td>0.8079555</td>
<td>9.726018e-18</td>
</tr>
<tr>
<td>VERB</td>
<td>0.7943419</td>
<td>8.339889e-17</td>
</tr>
<tr>
<td>ADP</td>
<td>-0.4925857</td>
<td>1.106015e-05</td>
</tr>
<tr>
<td>NOUN</td>
<td>-0.7593092</td>
<td>1.074374e-14</td>
</tr>
<tr>
<td>NUM</td>
<td>-0.8111124</td>
<td>5.766131e-18</td>
</tr>
</tbody>
</table>

### Dimension 1

<table>
<thead>
<tr>
<th></th>
<th>R2</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>num_speakers_sng_mult</td>
<td>0.2773670</td>
<td>1.357262e-05</td>
</tr>
<tr>
<td>type</td>
<td>0.2931671</td>
<td>2.802393e-05</td>
</tr>
<tr>
<td>pre_post_conquest</td>
<td>0.1900858</td>
<td>1.291121e-04</td>
</tr>
<tr>
<td>pop_s_m_l</td>
<td>0.1364717</td>
<td>6.332014e-03</td>
</tr>
</tbody>
</table>

### Dimension 1

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>na</td>
<td>1.2706597</td>
<td>1.071531e-05</td>
</tr>
<tr>
<td>summ</td>
<td>0.9043000</td>
<td>1.709764e-05</td>
</tr>
<tr>
<td>pre</td>
<td>0.8387015</td>
<td>1.291121e-04</td>
</tr>
<tr>
<td>l</td>
<td>-0.9652098</td>
<td>1.529683e-03</td>
</tr>
<tr>
<td>sng</td>
<td>-1.0957094</td>
<td>2.175150e-04</td>
</tr>
<tr>
<td>post</td>
<td>-0.8387015</td>
<td>1.291121e-04</td>
</tr>
<tr>
<td>pnum</td>
<td>-1.3133324</td>
<td>4.225171e-06</td>
</tr>
</tbody>
</table>

### Dimension 2

<table>
<thead>
<tr>
<th></th>
<th>correlation</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DET</td>
<td>0.7751486</td>
<td>1.330750e-15</td>
</tr>
<tr>
<td>AUX</td>
<td>0.6854868</td>
<td>3.076272e-11</td>
</tr>
<tr>
<td>ADP</td>
<td>0.5589112</td>
<td>3.351538e-07</td>
</tr>
<tr>
<td>ADJ</td>
<td>0.4295632</td>
<td>1.662168e-04</td>
</tr>
<tr>
<td>NUM</td>
<td>-0.3538858</td>
<td>2.362740e-03</td>
</tr>
</tbody>
</table>

### Dimension 2

<table>
<thead>
<tr>
<th></th>
<th>R2</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>currency</td>
<td>0.1745874</td>
<td>0.0002602244</td>
</tr>
</tbody>
</table>

### Dimension 2

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>0.6445948</td>
<td>0.0002602244</td>
</tr>
<tr>
<td>no</td>
<td>-0.6445948</td>
<td>0.0002602244</td>
</tr>
</tbody>
</table>

Table A.5: Significant Factor Loadings (UPOS Implementation)

I also include the full list of active variable contributions below.
The full variable correlation matrix, which serves as the basis for the PCA implementation, is reproduced below.

<table>
<thead>
<tr>
<th></th>
<th>DET</th>
<th>ADJ</th>
<th>ADP</th>
<th>AUX</th>
<th>NOUN</th>
<th>NUM</th>
<th>PRON</th>
<th>SCONJ</th>
<th>VERB</th>
</tr>
</thead>
<tbody>
<tr>
<td>DET</td>
<td>1</td>
<td>0.253</td>
<td>0.308</td>
<td>0.343</td>
<td>-0.130</td>
<td>-0.492</td>
<td>-0.005</td>
<td>0.056</td>
<td>0.295</td>
</tr>
<tr>
<td>ADJ</td>
<td>0.253</td>
<td>1</td>
<td>-0.045</td>
<td>0.188</td>
<td>-0.067</td>
<td>-0.287</td>
<td>0.053</td>
<td>0.043</td>
<td>0.149</td>
</tr>
<tr>
<td>ADP</td>
<td>0.308</td>
<td>-0.045</td>
<td>1</td>
<td>0.130</td>
<td>0.528</td>
<td>0.208</td>
<td>-0.424</td>
<td>-0.361</td>
<td>-0.218</td>
</tr>
<tr>
<td>AUX</td>
<td>0.343</td>
<td>0.188</td>
<td>0.130</td>
<td>1</td>
<td>0.096</td>
<td>-0.305</td>
<td>0.030</td>
<td>-0.101</td>
<td>0.293</td>
</tr>
<tr>
<td>NOUN</td>
<td>-0.130</td>
<td>-0.067</td>
<td>0.528</td>
<td>0.096</td>
<td>1</td>
<td>0.651</td>
<td>-0.523</td>
<td>-0.513</td>
<td>-0.399</td>
</tr>
<tr>
<td>NUM</td>
<td>-0.492</td>
<td>-0.287</td>
<td>0.208</td>
<td>-0.305</td>
<td>0.651</td>
<td>1</td>
<td>-0.516</td>
<td>-0.460</td>
<td>-0.600</td>
</tr>
<tr>
<td>PRON</td>
<td>-0.005</td>
<td>0.053</td>
<td>-0.424</td>
<td>0.030</td>
<td>-0.523</td>
<td>-0.516</td>
<td>1</td>
<td>0.826</td>
<td>0.642</td>
</tr>
<tr>
<td>SCONJ</td>
<td>0.056</td>
<td>0.043</td>
<td>-0.361</td>
<td>-0.101</td>
<td>-0.513</td>
<td>-0.460</td>
<td>0.826</td>
<td>1</td>
<td>0.602</td>
</tr>
<tr>
<td>VERB</td>
<td>0.295</td>
<td>0.149</td>
<td>-0.218</td>
<td>0.293</td>
<td>-0.399</td>
<td>-0.600</td>
<td>0.642</td>
<td>0.602</td>
<td>1</td>
</tr>
</tbody>
</table>

Table A.7: Variable Correlation Matrix (UPOS PCA Implementation)

I repeat use of the PerformanceAnalytics package to display the significance of the correlation coefficients among the active variables.
Figure A.3: Variable Correlation Significance Plot (UPOS Implementation)

The full individual contributions matrix is reproduced below. These values are used to construct figure 4.11, the plot of individuals.

<table>
<thead>
<tr>
<th></th>
<th>Dim.1</th>
<th>Dim.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>memoria_01</td>
<td>0.007067</td>
<td>5.875257</td>
</tr>
<tr>
<td>memoria_02</td>
<td>0.053936</td>
<td>1.037736</td>
</tr>
<tr>
<td>memoria_03</td>
<td>1.543776</td>
<td>2.49209</td>
</tr>
<tr>
<td>memoria_04</td>
<td>0.100219</td>
<td>0.856368</td>
</tr>
<tr>
<td></td>
<td>Dim.1</td>
<td>Dim.2</td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>memoria_05</td>
<td>0.928352</td>
<td>0.232468</td>
</tr>
<tr>
<td>memoria_06</td>
<td>0.954213</td>
<td>0.917969</td>
</tr>
<tr>
<td>memoria_07</td>
<td>0.00635</td>
<td>0.79506</td>
</tr>
<tr>
<td>memoria_08</td>
<td>9.530877</td>
<td>0.725973</td>
</tr>
<tr>
<td>memoria_09</td>
<td>9.593572</td>
<td>1.34953</td>
</tr>
<tr>
<td>memoria_10</td>
<td>10.65523</td>
<td>2.247156</td>
</tr>
<tr>
<td>memoria_11</td>
<td>0.226435</td>
<td>0.066144</td>
</tr>
<tr>
<td>memoria_12</td>
<td>0.012162</td>
<td>0.495935</td>
</tr>
<tr>
<td>memoria_13</td>
<td>0.164743</td>
<td>1.504649</td>
</tr>
<tr>
<td>memoria_14</td>
<td>0.006079</td>
<td>0.043018</td>
</tr>
<tr>
<td>memoria_15</td>
<td>0.060074</td>
<td>0.429771</td>
</tr>
<tr>
<td>memoria_16</td>
<td>0.001909</td>
<td>0.19021</td>
</tr>
<tr>
<td>memoria_17</td>
<td>1.359432</td>
<td>12.89064</td>
</tr>
<tr>
<td>memoria_18</td>
<td>0.145812</td>
<td>2.593729</td>
</tr>
<tr>
<td>memoria_19</td>
<td>0.014094</td>
<td>6.019252</td>
</tr>
<tr>
<td>memoria_20</td>
<td>3.5708</td>
<td>1.152983</td>
</tr>
<tr>
<td>memoria_21</td>
<td>0.260843</td>
<td>0.009468</td>
</tr>
<tr>
<td>memoria_22</td>
<td>0.007276</td>
<td>0.018924</td>
</tr>
<tr>
<td>memoria_23</td>
<td>2.291828</td>
<td>0.013744</td>
</tr>
<tr>
<td>memoria_24</td>
<td>1.508018</td>
<td>0.030188</td>
</tr>
<tr>
<td>memoria_25</td>
<td>2.873191</td>
<td>1.7E-06</td>
</tr>
<tr>
<td>memoria_26</td>
<td>3.287046</td>
<td>1.566598</td>
</tr>
<tr>
<td>memoria_27</td>
<td>1.082839</td>
<td>2.046737</td>
</tr>
<tr>
<td>memoria_28</td>
<td>2.451111</td>
<td>0.282468</td>
</tr>
<tr>
<td>memoria_29</td>
<td>2.18066</td>
<td>0.461677</td>
</tr>
<tr>
<td></td>
<td>Dim.1</td>
<td>Dim.2</td>
</tr>
<tr>
<td>--------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>memoria_30</td>
<td>5.921124</td>
<td>6.310025</td>
</tr>
<tr>
<td>memoria_31</td>
<td>0.542615</td>
<td>0.825689</td>
</tr>
<tr>
<td>memoria_32</td>
<td>0.00027</td>
<td>1.05713</td>
</tr>
<tr>
<td>memoria_33</td>
<td>0.418388</td>
<td>0.958918</td>
</tr>
<tr>
<td>memoria_34</td>
<td>0.209744</td>
<td>0.11405</td>
</tr>
<tr>
<td>memoria_35</td>
<td>0.387446</td>
<td>0.051241</td>
</tr>
<tr>
<td>memoria_36</td>
<td>0.023393</td>
<td>1.820452</td>
</tr>
<tr>
<td>memoria_37</td>
<td>0.01614</td>
<td>0.003701</td>
</tr>
<tr>
<td>memoria_38</td>
<td>1.503506</td>
<td>1.889773</td>
</tr>
<tr>
<td>memoria_39</td>
<td>3.905651</td>
<td>1.911054</td>
</tr>
<tr>
<td>memoria_40</td>
<td>0.269076</td>
<td>0.084814</td>
</tr>
<tr>
<td>memoria_41</td>
<td>0.047584</td>
<td>0.166261</td>
</tr>
<tr>
<td>memoria_42</td>
<td>0.21743</td>
<td>0.013389</td>
</tr>
<tr>
<td>memoria_43</td>
<td>0.184394</td>
<td>0.843392</td>
</tr>
<tr>
<td>memoria_44</td>
<td>12.46771</td>
<td>5.704977</td>
</tr>
<tr>
<td>memoria_45</td>
<td>0.736274</td>
<td>1.968307</td>
</tr>
<tr>
<td>memoria_46</td>
<td>0.631865</td>
<td>0.218231</td>
</tr>
<tr>
<td>memoria_47</td>
<td>0.220729</td>
<td>0.311823</td>
</tr>
<tr>
<td>memoria_48</td>
<td>0.00737</td>
<td>0.257087</td>
</tr>
<tr>
<td>memoria_49</td>
<td>0.262724</td>
<td>0.218004</td>
</tr>
<tr>
<td>memoria_50</td>
<td>0.051067</td>
<td>0.124945</td>
</tr>
<tr>
<td>memoria_51</td>
<td>0.000606</td>
<td>0.028552</td>
</tr>
<tr>
<td>memoria_52</td>
<td>1.649831</td>
<td>0.389626</td>
</tr>
<tr>
<td>memoria_53</td>
<td>0.015486</td>
<td>0.035641</td>
</tr>
<tr>
<td>memoria_54</td>
<td>0.006862</td>
<td>0.000167</td>
</tr>
<tr>
<td></td>
<td>Dim.1</td>
<td>Dim.2</td>
</tr>
<tr>
<td>----------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>memoria_55</td>
<td>1.982732</td>
<td>0.032949</td>
</tr>
<tr>
<td>memoria_56</td>
<td>0.978357</td>
<td>4.800736</td>
</tr>
<tr>
<td>memoria_57</td>
<td>0.512186</td>
<td>1.261881</td>
</tr>
<tr>
<td>memoria_58</td>
<td>1.354433</td>
<td>1.208196</td>
</tr>
<tr>
<td>memoria_59</td>
<td>0.13806</td>
<td>0.049714</td>
</tr>
<tr>
<td>memoria_60</td>
<td>0.525717</td>
<td>0.757507</td>
</tr>
<tr>
<td>memoria_61</td>
<td>2.022982</td>
<td>1.801253</td>
</tr>
<tr>
<td>memoria_62</td>
<td>0.146376</td>
<td>1.414668</td>
</tr>
<tr>
<td>memoria_63</td>
<td>0.32246</td>
<td>0.775111</td>
</tr>
<tr>
<td>memoria_64</td>
<td>0.006577</td>
<td>2.47957</td>
</tr>
<tr>
<td>memoria_65</td>
<td>0.006575</td>
<td>4.975812</td>
</tr>
<tr>
<td>memoria_66</td>
<td>0.06157</td>
<td>0.140297</td>
</tr>
<tr>
<td>memoria_67</td>
<td>0.004212</td>
<td>0.879918</td>
</tr>
<tr>
<td>memoria_68</td>
<td>0.129425</td>
<td>3.860584</td>
</tr>
<tr>
<td>memoria_69</td>
<td>4.836169</td>
<td>0.000446</td>
</tr>
<tr>
<td>memoria_70</td>
<td>0.184526</td>
<td>2.679088</td>
</tr>
<tr>
<td>memoria_71</td>
<td>0.324694</td>
<td>0.93847</td>
</tr>
<tr>
<td>memoria_72</td>
<td>1.887726</td>
<td>0.290802</td>
</tr>
</tbody>
</table>

Table A.8: Individual Contributions (UPOS PCA Implementation)
A.1.2.2 Parallel Analysis Plot

Parallel Analysis with 10,000 Simulations

Figure A.4: Parallel Analysis Plot (UPOS Implementation)
A.2 Hierarchical Agglomerative Clustering
A.2.1 Baker’s Gamma for Dendrogram Comparison

![Baker's Gamma Distribution under H0 with 1,000 Permutations](image)

Figure A.5: Baker’s Gamma Null Distribution Density under Permutation Test

It is clear from the above figure that the observed index (blue line) is strictly larger than any value of the null distribution, confirming the significance of the permutation test. The following plot (figure A.6) shows the bootstrap distribution of Baker’s gamma, including a 95% confidence interval (green lines), to contextualize the observed index value (red line) in the hierarchical clustering comparison.
A.2.2 Critical Rejection Value for \( k = 2 \) in Fowlkes-Mallow

The critical rejection value in the Fowlkes-Mallow analysis follows from the following readout:

```r
FM_index_R(cutree(pos.hclust,k = 2),cutree(tok.hclust,k=2))
# [1] 0.6843131
# attr(,"E_FM")
# [1] 0.5031147
# attr(,"V_FM")
# [1] 9.637995e-05
```

In the above, the three values refer to: the observed index (0.684), the expected index under \( H_0 \) (0.503; that one tree is a randomly shuffled version of the other’s labels) and the index variance under \( H_0 \) (9.638e-5). The calculation is as follows, under the asymptotic normal distribution:
\[ 0.6843131 > 0.5192642 = 0.5031147 + 1.645\sqrt{9.637995e - 05} \]

The observed index value is larger than the critical value under \( H_0 \). The one-sided test is significant, and \( H_0 \) is rejected (\( p < 0.05 \)): one tree is not a randomly shuffled version of the other’s labels.
## Appendix B

### Transcription Metadata

<table>
<thead>
<tr>
<th>memoria</th>
<th>title</th>
<th>author</th>
<th>line_start</th>
<th>line_end</th>
<th>year</th>
<th>lat</th>
<th>lon</th>
<th>dist_cuzco (km)</th>
<th>altitude (m)</th>
<th>city</th>
<th>type</th>
<th>pre_post_conquest</th>
<th>language</th>
<th>int_unint</th>
<th>pop_s_m_l</th>
<th>num_speakers</th>
<th>currency</th>
<th>lat_zone</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Memoria de los nietos de Topa Inga Yupangui</td>
<td>Capac Ayllu</td>
<td>304</td>
<td>315</td>
<td>1569</td>
<td>-13.532</td>
<td>-71.968</td>
<td>1.56</td>
<td>3337</td>
<td>Cuzco</td>
<td>pnar</td>
<td>pre</td>
<td>quechua</td>
<td>unint</td>
<td>l</td>
<td>mult</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>2</td>
<td>Memoria de las provincias que conquistó Topa Ynga Yupangui padre de Guaina Capac Ynga con sus hermanos</td>
<td>Capac Ayllu</td>
<td>383</td>
<td>481</td>
<td>1569</td>
<td>-13.532</td>
<td>-71.968</td>
<td>1.56</td>
<td>3337</td>
<td>Cuzco</td>
<td>pnar</td>
<td>pre</td>
<td>quechua</td>
<td>unint</td>
<td>l</td>
<td>mult</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>3</td>
<td>Relación de las guacas del Cuzco</td>
<td>Bernabé Cobo</td>
<td>159</td>
<td>1327</td>
<td>1653</td>
<td>-13.532</td>
<td>-71.968</td>
<td>1.56</td>
<td>3337</td>
<td>Cuzco</td>
<td>pnar</td>
<td>pre</td>
<td>quechua</td>
<td>unint</td>
<td>l</td>
<td>sng</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>4</td>
<td>Relación de lo que los Chupachos daban al Inga</td>
<td>Pauacaguaman, Querin, Julcacondor</td>
<td>59</td>
<td>127</td>
<td>1549</td>
<td>-9.942</td>
<td>-76.239</td>
<td>610.4</td>
<td>1925</td>
<td>Conchumayo</td>
<td>pnum</td>
<td>post</td>
<td>quechua</td>
<td>unint</td>
<td>l</td>
<td>mult</td>
<td>no</td>
<td>north</td>
</tr>
<tr>
<td>5</td>
<td>Relación de lo que los Chupachos dan a Gómez Arias de Ábila</td>
<td>Pauacaguaman, Querin, Julcacondor</td>
<td>135</td>
<td>224</td>
<td>1549</td>
<td>-9.942</td>
<td>-76.239</td>
<td>610.4</td>
<td>1925</td>
<td>Conchumayo</td>
<td>pnum</td>
<td>post</td>
<td>quechua</td>
<td>unint</td>
<td>l</td>
<td>mult</td>
<td>no</td>
<td>north</td>
</tr>
<tr>
<td>6</td>
<td>Quipo y memoria de los yndios del valle de los Chillos, encomendados a Francisco Bayz, sobre que an dado para llevar carga</td>
<td>Quipocamayos de los Chillos</td>
<td>72</td>
<td>82</td>
<td>1564</td>
<td>-0.278</td>
<td>-78.541</td>
<td>1633.64</td>
<td>2862</td>
<td>Valle de los Chillos</td>
<td>pnum</td>
<td>post</td>
<td>quechua</td>
<td>unint</td>
<td>s</td>
<td>mult</td>
<td>no</td>
<td>north</td>
</tr>
<tr>
<td>7</td>
<td>Memoria de los indios que yo Don Jerónimo Guacrapaucar di al marquez Don Francisco Pizarro desde que salió de Caxamamarca</td>
<td>Jerónimo Guacrapaucar</td>
<td>22</td>
<td>394</td>
<td>1558</td>
<td>-11.917</td>
<td>-75.333</td>
<td>405.55</td>
<td>3264</td>
<td>Mantaro Valle</td>
<td>pnum</td>
<td>post</td>
<td>quechua</td>
<td>unint</td>
<td>l</td>
<td>sng</td>
<td>yes</td>
<td>central</td>
</tr>
<tr>
<td>memoria</td>
<td>title</td>
<td>author</td>
<td>line_start</td>
<td>line_end</td>
<td>year</td>
<td>lat</td>
<td>lon</td>
<td>dist_cuzco (km)</td>
<td>altitude (m)</td>
<td>city</td>
<td>type</td>
<td>pre_post_conquest</td>
<td>language</td>
<td>int_unint</td>
<td>pop_s_m_l</td>
<td>num_speakers_sng_mult</td>
<td>currency</td>
<td>lat_zone</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>--------</td>
<td>------------</td>
<td>----------</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td>-----------------</td>
<td>-------------</td>
<td>------</td>
<td>------</td>
<td>------------------</td>
<td>-----------</td>
<td>------------</td>
<td>-----------</td>
<td>----------------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>8</td>
<td>Memoria de las cosas que Don Christóval Alaya y sus caciques e yndios dieron a los capitanes de su magestat para la guerra contra Francisco Hernández Girón</td>
<td>Christóval Alaya</td>
<td>398</td>
<td>455</td>
<td>1558</td>
<td>-11.917</td>
<td>-75.333</td>
<td>405.55</td>
<td>3264</td>
<td>Mantaro Valle</td>
<td>pnum</td>
<td>post</td>
<td>quechua</td>
<td>unint</td>
<td>1</td>
<td>sng</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>9</td>
<td>Memoria y relación de lo que yo Don Álvaro, cacique de los yndios de Carabantes gasté con los capitanes de su magestat en la guerra de Francisco Hernández y di para ella</td>
<td>Don Álvaro</td>
<td>459</td>
<td>505</td>
<td>1558</td>
<td>-11.917</td>
<td>-75.333</td>
<td>405.55</td>
<td>3264</td>
<td>Mantaro Valle</td>
<td>pnum</td>
<td>post</td>
<td>quechua</td>
<td>unint</td>
<td>1</td>
<td>sng</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>10</td>
<td>Memoria y relación de lo que con los capitanes de su majestad gasté yo Don Jerónimo en la guerra de Francisco Hernández y di para ella</td>
<td>Jerónimo Guacrapaucar</td>
<td>508</td>
<td>552</td>
<td>1558</td>
<td>-11.917</td>
<td>-75.333</td>
<td>405.55</td>
<td>3264</td>
<td>Mantaro Valle</td>
<td>pnum</td>
<td>post</td>
<td>quechua</td>
<td>unint</td>
<td>1</td>
<td>sng</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>11</td>
<td>Memoria de lo que Sulichaque, cacique principal que fue del repartimiento de Atunxauxa, dio al marqués Don Francisco Pizarro al tiempo que entró en este reyno e a los demás capitanes de su magestad para su auiamiento en todas las alteraciones que a ayudo</td>
<td>Francisco Cusichac, Cristóval Canchaya y Diego Enaspari</td>
<td>17</td>
<td>1433</td>
<td>1561</td>
<td>-11.77</td>
<td>-75.493</td>
<td>428.37</td>
<td>3369</td>
<td>Atun Jauja</td>
<td>pnum</td>
<td>post</td>
<td>quechua</td>
<td>unint</td>
<td>1</td>
<td>mult</td>
<td>yes</td>
<td>central</td>
</tr>
<tr>
<td>memoria</td>
<td>title</td>
<td>author</td>
<td>line_start</td>
<td>line_end</td>
<td>year</td>
<td>lat</td>
<td>lon</td>
<td>dist_cuzco (km)</td>
<td>altitude (m)</td>
<td>city</td>
<td>type</td>
<td>pre_post_conquest</td>
<td>language</td>
<td>int_unint</td>
<td>pop_s_m_l</td>
<td>num_speakers_sng_mult</td>
<td>currency</td>
<td>lat_zone</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>--------</td>
<td>------------</td>
<td>----------</td>
<td>------</td>
<td>------</td>
<td>-------</td>
<td>-----------------</td>
<td>--------------</td>
<td>--------</td>
<td>------</td>
<td>------------------</td>
<td>----------</td>
<td>-----------</td>
<td>-----------</td>
<td>------------------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>12</td>
<td>Declaración de quipocamayos de Mataguaci, en el repartimiento de los Huringuancas, sobre ciertos pesos de oro los que fueron dados a Bartolomé Díaz</td>
<td>Juan Apunpicho, Francisco Auquispicho y Pedro Yaldama</td>
<td>110</td>
<td>122</td>
<td>1570</td>
<td>-11.891</td>
<td>-75.392</td>
<td>412.53</td>
<td>3316</td>
<td>Sincos</td>
<td>pnum</td>
<td>post</td>
<td>quechua</td>
<td>unint</td>
<td>l</td>
<td>multi</td>
<td>central</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Relación del ganado que parece aver entregado Juan Vásquez de Tapia y Diego Pacheco a estos caçiques de Chucuyto y sus provincias</td>
<td>Caciques de Chucuito</td>
<td>5</td>
<td>419</td>
<td>1550</td>
<td>-15.894</td>
<td>-69.889</td>
<td>345.62</td>
<td>3876</td>
<td>Chacuito</td>
<td>pnum</td>
<td>post</td>
<td>quechua</td>
<td>unint</td>
<td>l</td>
<td>multi</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>14</td>
<td>La cuenta y quipo de los tributos y tasa que pagaua el repartimiento de Sacaca a Don Alonzo de Montemayor</td>
<td>Hernando Achacata</td>
<td>74</td>
<td>283</td>
<td>1551</td>
<td>-18.067</td>
<td>-66.394</td>
<td>781.28</td>
<td>3714</td>
<td>Sacaca</td>
<td>pnum</td>
<td>post</td>
<td>aymara</td>
<td>unint</td>
<td>m</td>
<td>single</td>
<td>yes</td>
<td>south</td>
</tr>
<tr>
<td>15</td>
<td>La cuenta y quipo de los tributos y tasa que pagaua el repartimiento de Sacaca a Don Alonzo de Montemayor</td>
<td>Luis Anba</td>
<td>513</td>
<td>678</td>
<td>1551</td>
<td>-18.067</td>
<td>-66.394</td>
<td>781.28</td>
<td>3714</td>
<td>Sacaca</td>
<td>pnum</td>
<td>post</td>
<td>aymara</td>
<td>unint</td>
<td>m</td>
<td>single</td>
<td>yes</td>
<td>south</td>
</tr>
<tr>
<td>16</td>
<td>La cuenta y quipo de los tributos y tasa que pagaua el repartimiento de Sacaca a Don Alonzo de Montemayor</td>
<td>Pedro Horuro</td>
<td>994</td>
<td>1178</td>
<td>1551</td>
<td>-18.067</td>
<td>-66.394</td>
<td>781.28</td>
<td>3714</td>
<td>Sacaca</td>
<td>pnum</td>
<td>post</td>
<td>aymara</td>
<td>unint</td>
<td>m</td>
<td>single</td>
<td>yes</td>
<td>south</td>
</tr>
<tr>
<td>17</td>
<td>La cuenta y razón de lo que pagaron al dicho Don Alonzo de Montemayor y a sus mayordomos por sus quipos</td>
<td>Fernando Achacata y Luis Conba</td>
<td>88</td>
<td>1218</td>
<td>1551</td>
<td>-18.067</td>
<td>-66.394</td>
<td>781.28</td>
<td>3714</td>
<td>Sacaca</td>
<td>elicite</td>
<td>post</td>
<td>aymara</td>
<td>int</td>
<td>m</td>
<td>multi</td>
<td>yes</td>
<td>south</td>
</tr>
<tr>
<td>memoria</td>
<td>title</td>
<td>author</td>
<td>line_start</td>
<td>line_end</td>
<td>year</td>
<td>lat</td>
<td>lon</td>
<td>dist_cuzco (km)</td>
<td>altitude (m)</td>
<td>city</td>
<td>type</td>
<td>pre_post_conquest</td>
<td>language</td>
<td>int_unint</td>
<td>pop_s_m_l</td>
<td>num_speakers_sng_mult</td>
<td>currency</td>
<td>lat_zone</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>--------</td>
<td>------------</td>
<td>----------</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td>-----------------</td>
<td>-------------</td>
<td>------</td>
<td>------</td>
<td>-----------------</td>
<td>----------</td>
<td>-----------</td>
<td>----------</td>
<td>---------------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>18</td>
<td>Relación que hazen los quipocamayos de las chácaras de coca de Sacaca</td>
<td>Qiupcamayos de las chácaras de coca de Sacaca</td>
<td>33</td>
<td>210</td>
<td>1551</td>
<td>-17.404</td>
<td>-65.737</td>
<td>795.34</td>
<td>3443</td>
<td>Tiraque</td>
<td>pnum</td>
<td>post</td>
<td>aymara</td>
<td>unint</td>
<td>m</td>
<td>mult</td>
<td>no</td>
<td>south</td>
</tr>
<tr>
<td>19</td>
<td>Relación de la declaración que hizieron los dichos caciques e indios de Chayanta encomendados en el dicho Martín de Robles de lo que an dado al dicho Martín Robles</td>
<td>Martín Chiciguancan, Diego Haya, Juan Copaquia y Felipe Cañahire</td>
<td>10</td>
<td>510</td>
<td>1555</td>
<td>-18.7</td>
<td>-66.004</td>
<td>858.05</td>
<td>4175</td>
<td>Chayanta</td>
<td>elicit</td>
<td>post</td>
<td>aymara</td>
<td>int</td>
<td>l</td>
<td>mult</td>
<td>yes</td>
<td>south</td>
</tr>
<tr>
<td>20</td>
<td>Título de la encomienda de Francisco Pizarro a Gómez de Maquelas, Cuzco, 1-VII-1535</td>
<td>NA</td>
<td>162</td>
<td>191</td>
<td>1535</td>
<td>-13.532</td>
<td>-71.968</td>
<td>1.56</td>
<td>3337</td>
<td>Cuzco</td>
<td>summ</td>
<td>pre</td>
<td>quechua</td>
<td>unint</td>
<td>m</td>
<td>na</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>21</td>
<td>Título de la encomienda de Francisco Pizarro a Gonzalo Pizarro, Cuzco, 7-III-1540</td>
<td>NA</td>
<td>79</td>
<td>325</td>
<td>1540</td>
<td>-18.25</td>
<td>-65.951</td>
<td>830.56</td>
<td>3153</td>
<td>Qaraqara</td>
<td>summ</td>
<td>pre</td>
<td>aymara</td>
<td>unint</td>
<td>l</td>
<td>na</td>
<td>no</td>
<td>south</td>
</tr>
<tr>
<td>22</td>
<td>Título de la encomienda de Pedro de la Gasca a Pedro de Hinojosa, Cuzco 29-VIII-1548</td>
<td>NA</td>
<td>124</td>
<td>210</td>
<td>1548</td>
<td>-18.25</td>
<td>-65.951</td>
<td>830.56</td>
<td>3153</td>
<td>Qaraqara</td>
<td>summ</td>
<td>pre</td>
<td>aymara</td>
<td>unint</td>
<td>l</td>
<td>na</td>
<td>no</td>
<td>south</td>
</tr>
<tr>
<td>23</td>
<td>Quipo Inca sobre el número de indios que había en la provincia de Chucuito</td>
<td>Martín Cari</td>
<td>16</td>
<td>88</td>
<td>1567</td>
<td>-15.894</td>
<td>-69.889</td>
<td>345.62</td>
<td>3876</td>
<td>Chucuito</td>
<td>pnum</td>
<td>pre</td>
<td>aymara</td>
<td>unint</td>
<td>l</td>
<td>sng</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>24</td>
<td>Quipo sobre la ropa que toda cabecera paga tributo a su magestad</td>
<td>Martín Cari</td>
<td>125</td>
<td>151</td>
<td>1567</td>
<td>-15.894</td>
<td>-69.889</td>
<td>345.62</td>
<td>3876</td>
<td>Chucuito</td>
<td>pnum</td>
<td>post</td>
<td>aymara</td>
<td>unint</td>
<td>l</td>
<td>sng</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>memoria</td>
<td>title</td>
<td>author</td>
<td>line_start</td>
<td>line_end</td>
<td>year</td>
<td>lat</td>
<td>lon</td>
<td>dist_cuzco (km)</td>
<td>altitude (m)</td>
<td>city</td>
<td>type</td>
<td>pre_post_conquest</td>
<td>language</td>
<td>int_unint</td>
<td>pop_s_m_l</td>
<td>num_speakers_sng_mult</td>
<td>currency</td>
<td>lat_zone</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>--------</td>
<td>------------</td>
<td>----------</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td>----------------</td>
<td>--------------</td>
<td>------</td>
<td>------</td>
<td>----------------</td>
<td>-----------</td>
<td>-----------</td>
<td>----------</td>
<td>-------------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>25</td>
<td>Quipo sobre lo que cada cabecera paga del tributo en plata</td>
<td>Martín Cari</td>
<td>159</td>
<td>203</td>
<td>1567</td>
<td>-15.894</td>
<td>-69.889</td>
<td>345.62</td>
<td>3876</td>
<td>Chacucho</td>
<td>pnum</td>
<td>post</td>
<td>aymara</td>
<td>unint</td>
<td>1</td>
<td>sng</td>
<td>yes</td>
<td>central</td>
</tr>
<tr>
<td>26</td>
<td>Quipo sobre los indios de cada cabecera que van a las minas para pagar el tributo</td>
<td>Martín Cari</td>
<td>208</td>
<td>241</td>
<td>1567</td>
<td>-15.894</td>
<td>-69.889</td>
<td>345.62</td>
<td>3876</td>
<td>Chacucho</td>
<td>pnum</td>
<td>post</td>
<td>aymara</td>
<td>unint</td>
<td>1</td>
<td>sng</td>
<td>yes</td>
<td>central</td>
</tr>
<tr>
<td>27</td>
<td>Quipos sobre el ganado que tienen los religiosos</td>
<td>Quipocamayos</td>
<td>766</td>
<td>796</td>
<td>1567</td>
<td>-15.894</td>
<td>-69.889</td>
<td>345.62</td>
<td>3876</td>
<td>Chacucho</td>
<td>pnum</td>
<td>pre</td>
<td>aymara</td>
<td>unint</td>
<td>1</td>
<td>mult</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>28</td>
<td>Título de la encomienda de Francisco Pizarro a Pedro del Barco, Cuzco 1-VIII-1535</td>
<td>NA</td>
<td>88</td>
<td>104</td>
<td>1535</td>
<td>-17.822</td>
<td>-67.021</td>
<td>713.09</td>
<td>3762</td>
<td>Paria</td>
<td>summ</td>
<td>pre</td>
<td>aymara</td>
<td>unint</td>
<td>m</td>
<td>na</td>
<td>no</td>
<td>south</td>
</tr>
<tr>
<td>29</td>
<td>Título de la encomienda de Francisco Pizarro a Nicolás de Ribera, Los Reyes 13-I-1536</td>
<td>NA</td>
<td>92</td>
<td>116</td>
<td>1536</td>
<td>-11.469</td>
<td>-76.623</td>
<td>553.82</td>
<td>2822</td>
<td>Canta</td>
<td>summ</td>
<td>pre</td>
<td>quechua</td>
<td>unint</td>
<td>1</td>
<td>na</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>30</td>
<td>Información de la encomienda de Francisco Pizarro a Cristóbal Pérez y su hijo, sin fecha</td>
<td>NA</td>
<td>25</td>
<td>62</td>
<td>1550</td>
<td>-15.623</td>
<td>-71.98</td>
<td>232.61</td>
<td>3299</td>
<td>Cabanaconde</td>
<td>summ</td>
<td>pre</td>
<td>quechua</td>
<td>unint</td>
<td>1</td>
<td>na</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>31</td>
<td>Título de la encomienda de Cristóbal Vaca de Castro a Hernán Bueno, Cuzco 24-XI-1542</td>
<td>NA</td>
<td>36</td>
<td>58</td>
<td>1542</td>
<td>-16.811</td>
<td>-70.694</td>
<td>389.34</td>
<td>3072</td>
<td>Caramas</td>
<td>summ</td>
<td>pre</td>
<td>aymara</td>
<td>unint</td>
<td>s</td>
<td>na</td>
<td>no</td>
<td>south</td>
</tr>
<tr>
<td>32</td>
<td>Título de la encomienda de Francisco Pizarro a Lucas Martínez, Cuzco 1-VIII-1535</td>
<td>NA</td>
<td>30</td>
<td>52</td>
<td>1535</td>
<td>-16.409</td>
<td>-71.538</td>
<td>323.09</td>
<td>2320</td>
<td>Arequipa</td>
<td>summ</td>
<td>pre</td>
<td>quechua</td>
<td>unint</td>
<td>1</td>
<td>na</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>33</td>
<td>Título de la encomienda del licenciado Pedro de la Gasca a Hernán Bueno (el hijo), los reyes 20-XI-1549</td>
<td>NA</td>
<td>17</td>
<td>31</td>
<td>1549</td>
<td>-16.811</td>
<td>-70.694</td>
<td>389.34</td>
<td>3072</td>
<td>Caramas</td>
<td>summ</td>
<td>pre</td>
<td>aymara</td>
<td>unint</td>
<td>s</td>
<td>na</td>
<td>no</td>
<td>south</td>
</tr>
<tr>
<td>memoria</td>
<td>title</td>
<td>autor</td>
<td>line_start</td>
<td>line_end</td>
<td>year</td>
<td>lat</td>
<td>lon</td>
<td>dist_cuzco (km)</td>
<td>altitude (m)</td>
<td>ciudad</td>
<td>type</td>
<td>pre_post_conquest</td>
<td>idioma</td>
<td>int_unint</td>
<td>pop_s_m_l</td>
<td>num_speakers_sng_mult</td>
<td>currency</td>
<td>lat_zone</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>-------</td>
<td>------------</td>
<td>----------</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td>----------------</td>
<td>-------------</td>
<td>--------</td>
<td>------</td>
<td>-------------------</td>
<td>--------</td>
<td>-----------</td>
<td>----------</td>
<td>----------------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>34</td>
<td>Título de la encomienda de Francisco Pizarro a Lucas Martínez Vegaso, Cuzco 22-I-1540</td>
<td>NA</td>
<td>71</td>
<td>139</td>
<td>1540</td>
<td>-16.409</td>
<td>-71.538</td>
<td>323.09</td>
<td>2320</td>
<td>Arequipa</td>
<td>sum</td>
<td>pre</td>
<td>quechua</td>
<td>unint</td>
<td>l</td>
<td>na</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>35</td>
<td>Título de la encomienda de Francisco Pizarro a Francisco de Sauzedo, Cuzco 22-I-1540</td>
<td>NA</td>
<td>13</td>
<td>49</td>
<td>1540</td>
<td>-16.409</td>
<td>-71.538</td>
<td>323.09</td>
<td>2320</td>
<td>Cuzco</td>
<td>sum</td>
<td>pre</td>
<td>quechua</td>
<td>unint</td>
<td>m</td>
<td>na</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>36</td>
<td>Título de la encomienda de Francisco Pizarro a Hernando de Silva, Cuzco 22-I-1540</td>
<td>NA</td>
<td>27</td>
<td>36</td>
<td>1540</td>
<td>-16.409</td>
<td>-71.538</td>
<td>323.09</td>
<td>2320</td>
<td>Arequipa</td>
<td>sum</td>
<td>pre</td>
<td>quechua</td>
<td>unint</td>
<td>m</td>
<td>na</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>37</td>
<td>Título de la encomienda de Vaca de Castro a Hernando de Silva, Cuzco 24-XI-1542</td>
<td>NA</td>
<td>35</td>
<td>51</td>
<td>1542</td>
<td>-16.811</td>
<td>-70.694</td>
<td>309.34</td>
<td>3072</td>
<td>Carumas</td>
<td>sum</td>
<td>pre</td>
<td>quechua</td>
<td>unint</td>
<td>s</td>
<td>na</td>
<td>no</td>
<td>south</td>
</tr>
<tr>
<td>38</td>
<td>Título de la encomienda de Francisco Pizarro a Francisco Noguerol de Ulloa, Cuzco 22-I-1540</td>
<td>NA</td>
<td>41</td>
<td>61</td>
<td>1540</td>
<td>-16.388</td>
<td>-70.855</td>
<td>339.43</td>
<td>3372</td>
<td>Ubinas</td>
<td>sum</td>
<td>pre</td>
<td>quechua</td>
<td>unint</td>
<td>m</td>
<td>na</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>39</td>
<td>Capítulo de los indios que el marqués Don Francisco Piçarro encomendó a Antón Ruiz de Guevara</td>
<td>NA</td>
<td>68</td>
<td>78</td>
<td>1535</td>
<td>-16.24</td>
<td>-72.236</td>
<td>302.21</td>
<td>1591</td>
<td>Majes</td>
<td>sum</td>
<td>pre</td>
<td>quechua</td>
<td>unint</td>
<td>m</td>
<td>na</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>40</td>
<td>Título de la encomienda de Francisco Pizarro a Hernando de Aldana, Cuzco 22-I-1540</td>
<td>NA</td>
<td>36</td>
<td>115</td>
<td>1540</td>
<td>-18.25</td>
<td>-65.951</td>
<td>830.56</td>
<td>3153</td>
<td>Qaraqara</td>
<td>sum</td>
<td>pre</td>
<td>aymara</td>
<td>unint</td>
<td>l</td>
<td>na</td>
<td>no</td>
<td>south</td>
</tr>
<tr>
<td>41</td>
<td>Título de la encomienda de Francisco Pizarro a Alonso Manjarres, Cuzco 22-I-1540</td>
<td>NA</td>
<td>65</td>
<td>123</td>
<td>1540</td>
<td>-17.648</td>
<td>-67.207</td>
<td>685.58</td>
<td>3775</td>
<td>Caracollo</td>
<td>sum</td>
<td>pre</td>
<td>aymara</td>
<td>unint</td>
<td>l</td>
<td>na</td>
<td>no</td>
<td>south</td>
</tr>
<tr>
<td>memoria</td>
<td>title</td>
<td>author</td>
<td>line_start</td>
<td>line_end</td>
<td>year</td>
<td>lat</td>
<td>lon</td>
<td>dist_cuzco (km)</td>
<td>altitude (m)</td>
<td>city</td>
<td>type</td>
<td>pre_post_conquest</td>
<td>language</td>
<td>int_unint</td>
<td>pop_s_m_l</td>
<td>num_speakers_sng_mult</td>
<td>currency</td>
<td>lat_zone</td>
</tr>
<tr>
<td>---------</td>
<td>------------------------------------------------------------------------</td>
<td>--------</td>
<td>------------</td>
<td>----------</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td>----------------</td>
<td>--------------</td>
<td>------------</td>
<td>--------------</td>
<td>-----------------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>------------------------</td>
<td>-----------</td>
<td>----------</td>
</tr>
<tr>
<td>42</td>
<td>Título de la encomienda de Francisco Pizarro a Francisco de Retamoso, Cuzco 22-I-1540</td>
<td>NA</td>
<td>41</td>
<td>109</td>
<td>1540</td>
<td>-18.314</td>
<td>-67.857</td>
<td>689.7</td>
<td>3832</td>
<td>Carangas</td>
<td>summ</td>
<td>pre</td>
<td>aymara</td>
<td>unint</td>
<td>m</td>
<td>na</td>
<td>no</td>
<td>south</td>
</tr>
<tr>
<td>43</td>
<td>Título provisional de la encomienda de Francisco Pizarro a Alonso de Camargo, Los Reyes 14-X-1540</td>
<td>NA</td>
<td>74</td>
<td>94</td>
<td>1540</td>
<td>-21.533</td>
<td>-64.733</td>
<td>1173.25</td>
<td>1874</td>
<td>Tarija</td>
<td>summ</td>
<td>pre</td>
<td>quechua</td>
<td>unint</td>
<td>s</td>
<td>na</td>
<td>no</td>
<td>south</td>
</tr>
<tr>
<td>44</td>
<td>Título provisional de la encomienda de Francisco Pizarro a Marcos de Retamoso, Los Reyes 28-II-1541</td>
<td>NA</td>
<td>35</td>
<td>55</td>
<td>1541</td>
<td>-20.203</td>
<td>-69.288</td>
<td>793</td>
<td>2042</td>
<td>Tarapaca</td>
<td>summ</td>
<td>pre</td>
<td>aymara</td>
<td>unint</td>
<td>s</td>
<td>na</td>
<td>no</td>
<td>south</td>
</tr>
<tr>
<td>45</td>
<td>Título provisional de la encomienda de Francisco Pizarro a Gómez de Luna, Cuzco 22-I-1540</td>
<td>NA</td>
<td>48</td>
<td>71</td>
<td>1540</td>
<td>-18.314</td>
<td>-67.857</td>
<td>689.7</td>
<td>3832</td>
<td>Carangas</td>
<td>summ</td>
<td>pre</td>
<td>aymara</td>
<td>unint</td>
<td>m</td>
<td>na</td>
<td>no</td>
<td>south</td>
</tr>
<tr>
<td>46</td>
<td>Título provisional de la encomienda de Francisco Pizarro a Lope de Mendieta, Los Reyes 28-V-1540</td>
<td>NA</td>
<td>22</td>
<td>64</td>
<td>1540</td>
<td>-18.314</td>
<td>-67.857</td>
<td>689.7</td>
<td>3832</td>
<td>Carangas</td>
<td>summ</td>
<td>pre</td>
<td>aymara</td>
<td>unint</td>
<td>l</td>
<td>na</td>
<td>no</td>
<td>south</td>
</tr>
<tr>
<td>47</td>
<td>Título provisional de la encomienda de Francisco Pizarro a Lope de Mendieta, Los Reyes 23-III-1541</td>
<td>NA</td>
<td>5</td>
<td>22</td>
<td>1541</td>
<td>-18.314</td>
<td>-67.857</td>
<td>689.7</td>
<td>3832</td>
<td>Carangas</td>
<td>summ</td>
<td>pre</td>
<td>aymara</td>
<td>unint</td>
<td>s</td>
<td>na</td>
<td>no</td>
<td>south</td>
</tr>
<tr>
<td>48</td>
<td>Visita hecha en el valle de Jayanca</td>
<td>Sebastián de la Gama</td>
<td>316</td>
<td>1001</td>
<td>1540</td>
<td>-6.393</td>
<td>-79.823</td>
<td>1166.56</td>
<td>62</td>
<td>Jayanca</td>
<td>summ</td>
<td>pre</td>
<td>quechua</td>
<td>int</td>
<td>l</td>
<td>na</td>
<td>no</td>
<td>north</td>
</tr>
<tr>
<td>49</td>
<td>Visita hecha en el valle de Jayanca</td>
<td>Sebastián de la Gama</td>
<td>1126</td>
<td>1196</td>
<td>1540</td>
<td>-6.393</td>
<td>-79.823</td>
<td>1166.56</td>
<td>62</td>
<td>Jayanca</td>
<td>elicite</td>
<td>post</td>
<td>quechua</td>
<td>int</td>
<td>l</td>
<td>mult</td>
<td>yes</td>
<td>north</td>
</tr>
<tr>
<td>50</td>
<td>Título de la encomienda de Cristóbal Vaca de Castro a Alexo Rodríguez, Cuzco 17-IX-1543</td>
<td>NA</td>
<td>98</td>
<td>183</td>
<td>1543</td>
<td>-17.025</td>
<td>-68.609</td>
<td>530.16</td>
<td>3951</td>
<td>Caquiaviri</td>
<td>summ</td>
<td>pre</td>
<td>aymara</td>
<td>unint</td>
<td>m</td>
<td>na</td>
<td>no</td>
<td>south</td>
</tr>
<tr>
<td>memoria</td>
<td>title</td>
<td>author</td>
<td>line_start</td>
<td>line_end</td>
<td>year</td>
<td>lat</td>
<td>lon</td>
<td>dist_cuzco (km)</td>
<td>altitude (m)</td>
<td>city</td>
<td>type</td>
<td>pre_post_conquest</td>
<td>language</td>
<td>int_unint</td>
<td>pop_s_m_l</td>
<td>num_speakers</td>
<td>_sng_mult</td>
<td>currency</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------</td>
<td>------------</td>
<td>----------</td>
<td>------</td>
<td>-------</td>
<td>-------</td>
<td>----------------</td>
<td>--------------</td>
<td>-------------</td>
<td>----------</td>
<td>-------------------</td>
<td>----------</td>
<td>-----------</td>
<td>-----------</td>
<td>--------------</td>
<td>-----------</td>
<td>----------</td>
</tr>
<tr>
<td>51</td>
<td>Título de la encomienda de Cristóbal Vaca de Castro a Alonso de Barrionuevo, Cuzco 13-IX-1543</td>
<td>NA</td>
<td>77</td>
<td>125</td>
<td>1543</td>
<td>-17.176</td>
<td>-68.671</td>
<td>537.99</td>
<td>3867</td>
<td>Caquingora</td>
<td>summ</td>
<td>pre</td>
<td>aymara</td>
<td>unint</td>
<td>s</td>
<td>na</td>
<td>no</td>
<td>south</td>
</tr>
<tr>
<td>52</td>
<td>Quipo sobre el tributo dado a Ortún Sánchez</td>
<td>Cacique Guanca</td>
<td>2</td>
<td>7</td>
<td>1550</td>
<td>-17.176</td>
<td>-68.671</td>
<td>537.99</td>
<td>3867</td>
<td>Caquingora</td>
<td>elicita</td>
<td>post</td>
<td>aymara</td>
<td>unint</td>
<td>m</td>
<td>sng</td>
<td>no</td>
<td>south</td>
</tr>
<tr>
<td>53</td>
<td>Título de la encomienda de Pedro de la Gasca al licenciado Garcia de León, Cuzco 14-IX-1548</td>
<td>NA</td>
<td>179</td>
<td>214</td>
<td>1548</td>
<td>-16.166</td>
<td>-69.086</td>
<td>427.06</td>
<td>3863</td>
<td>Copacabana</td>
<td>summ</td>
<td>pre</td>
<td>aymara</td>
<td>unint</td>
<td>m</td>
<td>na</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>54</td>
<td>Título de la encomienda de Cristóbal Vaca de Castro a Francisco de Balboa, San Juan de la Vítoria 4-IV-1544</td>
<td>NA</td>
<td>34</td>
<td>65</td>
<td>1544</td>
<td>-13.558</td>
<td>-74.195</td>
<td>240.35</td>
<td>3321</td>
<td>Cangallo</td>
<td>summ</td>
<td>pre</td>
<td>quechua</td>
<td>unint</td>
<td>m</td>
<td>na</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>55</td>
<td>Visitación del repartimiento del cacique Guanca en la provincia de León de Huamaco</td>
<td>Miguel de la Serna y Juan de Espinosa</td>
<td>19</td>
<td>99</td>
<td>1549</td>
<td>-9.503</td>
<td>-76.885</td>
<td>695.89</td>
<td>3673</td>
<td>Puños</td>
<td>summ</td>
<td>pre</td>
<td>quechua</td>
<td>unint</td>
<td>l</td>
<td>na</td>
<td>no</td>
<td>north</td>
</tr>
<tr>
<td>56</td>
<td>Visitación del repartimiento de Nicolás de Ribera en Canta</td>
<td>Antonio de Ribera y Gerónimo de Silva</td>
<td>214</td>
<td>262</td>
<td>1549</td>
<td>-11.469</td>
<td>-76.623</td>
<td>553.82</td>
<td>2822</td>
<td>Canta</td>
<td>summ</td>
<td>pre</td>
<td>quechua</td>
<td>unint</td>
<td>s</td>
<td>na</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>57</td>
<td>Visitación del repartimiento de Nicolás de Ribera en Canta</td>
<td>Antonio de Ribera y Gerónimo de Silva</td>
<td>316</td>
<td>358</td>
<td>1549</td>
<td>-11.469</td>
<td>-76.623</td>
<td>553.82</td>
<td>2822</td>
<td>Canta</td>
<td>elicita</td>
<td>pre</td>
<td>quechua</td>
<td>unint</td>
<td>s</td>
<td>sng</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>58</td>
<td>Visitación del repartimiento de Nicolás de Ribera en Canta</td>
<td>Antonio de Ribera y Gerónimo de Silva</td>
<td>558</td>
<td>605</td>
<td>1549</td>
<td>-11.469</td>
<td>-76.623</td>
<td>553.82</td>
<td>2822</td>
<td>Canta</td>
<td>summ</td>
<td>pre</td>
<td>quechua</td>
<td>int</td>
<td>s</td>
<td>mult</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>59</td>
<td>Visitación del repartimiento de Nicolás de Ribera en Canta</td>
<td>Antonio de Ribera y Gerónimo de Silva</td>
<td>650</td>
<td>695</td>
<td>1549</td>
<td>-11.469</td>
<td>-76.623</td>
<td>553.82</td>
<td>2822</td>
<td>Canta</td>
<td>pnum</td>
<td>post</td>
<td>quechua</td>
<td>unint</td>
<td>s</td>
<td>mult</td>
<td>no</td>
<td>central</td>
</tr>
<tr>
<td>memoria</td>
<td>title</td>
<td>author</td>
<td>line_start</td>
<td>line_end</td>
<td>year</td>
<td>lat</td>
<td>lon</td>
<td>dist_cuzco(km)</td>
<td>altitude(m)</td>
<td>city</td>
<td>type</td>
<td>pre_post_conquest</td>
<td>language</td>
<td>int_unint</td>
<td>pop_s_m_l</td>
<td>num_speakers</td>
<td>currency</td>
<td>lat_zone</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>--------</td>
<td>------------</td>
<td>----------</td>
<td>------</td>
<td>------</td>
<td>-------</td>
<td>----------------</td>
<td>-------------</td>
<td>--------</td>
<td>--------</td>
<td>------------------</td>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>60</td>
<td>Visita de los repartimientos de Guaraz y Llaguaraz encomendados en Hernando de Torres y Ruy Barba</td>
<td>Licenciado Diego Álvarez</td>
<td>56</td>
<td>676</td>
<td>1558</td>
<td>-9.526</td>
<td>-77.529</td>
<td>749.72</td>
<td>3054</td>
<td>Huaraz</td>
<td>elicit</td>
<td>pre</td>
<td>quechua</td>
<td>unint</td>
<td>l</td>
<td>mult</td>
<td>no</td>
<td>north</td>
</tr>
<tr>
<td>61</td>
<td>Visita de los repartimientos de Guaraz y Llaguaraz encomendados en Hernando de Torres y Ruy Barba</td>
<td>Licenciado Diego Álvarez</td>
<td>679</td>
<td>699</td>
<td>1558</td>
<td>-9.526</td>
<td>-77.529</td>
<td>749.72</td>
<td>3054</td>
<td>Huaraz</td>
<td>summ</td>
<td>pre</td>
<td>quechua</td>
<td>unint</td>
<td>l</td>
<td>mult</td>
<td>no</td>
<td>north</td>
</tr>
<tr>
<td>62</td>
<td>Visita de los repartimientos de Guaraz y Llaguaraz encomendados en Hernando de Torres y Ruy Barba</td>
<td>Licenciado Diego Álvarez</td>
<td>785</td>
<td>817</td>
<td>1558</td>
<td>-9.526</td>
<td>-77.529</td>
<td>749.72</td>
<td>3054</td>
<td>Huaraz</td>
<td>elicit</td>
<td>post</td>
<td>quechua</td>
<td>int</td>
<td>l</td>
<td>mult</td>
<td>yes</td>
<td>north</td>
</tr>
<tr>
<td>63</td>
<td>Visita de los repartimientos de Guaraz y Llaguaraz encomendados en Hernando de Torres y Ruy Barba</td>
<td>Licenciado Diego Álvarez</td>
<td>999</td>
<td>1507</td>
<td>1558</td>
<td>-9.526</td>
<td>-77.529</td>
<td>749.72</td>
<td>3054</td>
<td>Huaraz</td>
<td>elicit</td>
<td>pre</td>
<td>quechua</td>
<td>int</td>
<td>l</td>
<td>mult</td>
<td>no</td>
<td>north</td>
</tr>
<tr>
<td>64</td>
<td>Retasa del repartimiento de los Anparaes que tuvo en encomienda don Bernardino de Mendoza</td>
<td>Andrés Hurtado de Mendoza</td>
<td>32</td>
<td>98</td>
<td>1558</td>
<td>-18.25</td>
<td>-65.951</td>
<td>830.56</td>
<td>3153</td>
<td>Charcas</td>
<td>pnum</td>
<td>post</td>
<td>aymara</td>
<td>unint</td>
<td>l</td>
<td>sng</td>
<td>yes</td>
<td>south</td>
</tr>
<tr>
<td>65</td>
<td>La tasa de Pedro de la Gasca para el repartimiento de Pablo de Meneses, el cual contenía la quipo</td>
<td>Gerónimo de Loayza, Fernando de Santillán y Domingo de Santo Thomas</td>
<td>184</td>
<td>282</td>
<td>1550</td>
<td>-18.25</td>
<td>-65.951</td>
<td>830.56</td>
<td>3153</td>
<td>Charcas</td>
<td>pnum</td>
<td>post</td>
<td>aymara</td>
<td>unint</td>
<td>l</td>
<td>sng</td>
<td>yes</td>
<td>south</td>
</tr>
<tr>
<td>memoria</td>
<td>title</td>
<td>author</td>
<td>line_start</td>
<td>line_end</td>
<td>year</td>
<td>lat</td>
<td>lon</td>
<td>dist_cuzco (km)</td>
<td>altitude (m)</td>
<td>city</td>
<td>type</td>
<td>pre_post_conquest</td>
<td>language</td>
<td>int_unint</td>
<td>pop_s_m_l</td>
<td>num_speakers</td>
<td>sng</td>
<td>mult</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>--------</td>
<td>------------</td>
<td>----------</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td>----------------</td>
<td>-------------</td>
<td>------</td>
<td>------</td>
<td>-----------------</td>
<td>----------</td>
<td>-----------</td>
<td>----------</td>
<td>-------------</td>
<td>-----</td>
<td>-------</td>
</tr>
<tr>
<td>66</td>
<td>Relación de lo que monta la tasa del repartimiento de los Anaparás que tuvo en la encomienda Pablo de Meneses, desde 1565 hasta 1572 por el libro de los indios y por su quipo</td>
<td>Anónimo</td>
<td>355</td>
<td>411</td>
<td>1572</td>
<td>-18.25</td>
<td>-65.951</td>
<td>830.56</td>
<td>3153</td>
<td>Charcas</td>
<td>pnum</td>
<td>post</td>
<td>aymara</td>
<td>unint</td>
<td>1</td>
<td>sng</td>
<td>yes</td>
<td>south</td>
</tr>
<tr>
<td>67</td>
<td>Relación de lo que monta la tasa del repartimiento de los Charcas de Moromoro que tuvieron en encomienda Pablo de Meneses, desde 1565 hasta 1572 por el libro de los indios y por su quipo</td>
<td>Anónimo</td>
<td>422</td>
<td>514</td>
<td>1572</td>
<td>-18.25</td>
<td>-65.951</td>
<td>830.56</td>
<td>3153</td>
<td>Charcas</td>
<td>pnum</td>
<td>post</td>
<td>aymara</td>
<td>unint</td>
<td>1</td>
<td>sng</td>
<td>yes</td>
<td>south</td>
</tr>
<tr>
<td>68</td>
<td>La tasa de Pedro de la Gasca del repartimiento de Machaca que tuvo en encomienda don Gerónimo de Soria</td>
<td>Gerónimo de Loayza, Fernando de Santillán y Domingo de Santo Thomas</td>
<td>18</td>
<td>92</td>
<td>1550</td>
<td>-16.956</td>
<td>-68.969</td>
<td>498.81</td>
<td>3933</td>
<td>Machaca la Grande</td>
<td>pnum</td>
<td>post</td>
<td>aymara</td>
<td>unint</td>
<td>1</td>
<td>sng</td>
<td>yes</td>
<td>south</td>
</tr>
<tr>
<td>69</td>
<td>Relación de lo que valió la mita primera del año de 1552 d'estas cuentas que la cobró Juan López de Alcalá de los indios de Machaca</td>
<td>Bartolomé de Otaça y Francisco de Tapia</td>
<td>631</td>
<td>666</td>
<td>1565</td>
<td>-16.956</td>
<td>-68.969</td>
<td>498.81</td>
<td>3933</td>
<td>Machaca la Grande</td>
<td>pnum</td>
<td>post</td>
<td>aymara</td>
<td>unint</td>
<td>1</td>
<td>mult</td>
<td>yes</td>
<td>south</td>
</tr>
<tr>
<td>70</td>
<td>Información hecha por mandado de su excelencia sobre los daños que se han recaído a los indios del balle de Xauxa en los Pleytos que han tumbado así en los bienes de la comunidad como en los de particulares</td>
<td>Caciques de Lurin Huanca y Hanan Huanca</td>
<td>55</td>
<td>172</td>
<td>1570</td>
<td>-11.781</td>
<td>-75.487</td>
<td>427.27</td>
<td>3364</td>
<td>Jauja</td>
<td>pnum</td>
<td>post</td>
<td>quechua</td>
<td>unint</td>
<td>1</td>
<td>sng</td>
<td>yes</td>
<td>central</td>
</tr>
<tr>
<td>memoria</td>
<td>title</td>
<td>author</td>
<td>line_start</td>
<td>line_end</td>
<td>year</td>
<td>lat</td>
<td>lon</td>
<td>dist_cuzco (km)</td>
<td>altitude (m)</td>
<td>city</td>
<td>type</td>
<td>pre_post_conquest</td>
<td>language</td>
<td>int_unint</td>
<td>pop_s_m_l</td>
<td>num_speakers</td>
<td>sng_mult</td>
<td>currency</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>--------</td>
<td>------------</td>
<td>----------</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td>----------------</td>
<td>--------------</td>
<td>------</td>
<td>------</td>
<td>----------------</td>
<td>----------</td>
<td>-----------</td>
<td>-----------</td>
<td>--------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>71</td>
<td>Ymformación hecha por mandado de su excelencia sobre los daños que se han recibido a los indios del balle de Xauxa en los Pleytos que han tenido así en los bienes de la comunidad como en los de particulares</td>
<td>Caciques de Lurin Huanca y Hanan Huanca</td>
<td>211</td>
<td>231</td>
<td>1570</td>
<td>-11.781</td>
<td>-75.487</td>
<td>427.27</td>
<td>3364</td>
<td>Jauja</td>
<td>pnum</td>
<td>post</td>
<td>quechua</td>
<td>unint</td>
<td>1</td>
<td>sng</td>
<td>yes</td>
<td>central</td>
</tr>
<tr>
<td>72</td>
<td>Ymformación hecha por mandado de su excelencia sobre los daños que se han recibido a los indios del balle de Xauxa en los Pleytos que han tenido así en los bienes de la comunidad como en los de particulares</td>
<td>Caciques de Lurin Huanca y Hanan Huanca</td>
<td>252</td>
<td>359</td>
<td>1570</td>
<td>-11.781</td>
<td>-75.487</td>
<td>427.27</td>
<td>3364</td>
<td>Jauja</td>
<td>pnum</td>
<td>post</td>
<td>quechua</td>
<td>unint</td>
<td>1</td>
<td>sng</td>
<td>no</td>
<td>central</td>
</tr>
</tbody>
</table>
Appendix C

Supplementary Code

C.1 Chapter 3 Supplementary Code

C.1.1 Tokenization, Lemmatization and UPOS-tagging

### Tokenization, lemmatization and POS-tagging using UDPipe ###

# install and load UDPipe wrapper for C++
install.packages("udpipe")
library(udpipe)

# install and load xlsx package for writing data.frame annotation to excel
install.packages("xlsx")
library(xlsx)

# install and load readr package for vectorizing .txt input files
install.packages("readr")
library(readr)

# install and load xml2 package for importing and parsing XML corpus textos files
install.packages("xml2")
library(xml2)

# install and load plyr/dplyr packages for subsetting data frames within corpus list object
install.packages("dplyr")
library(dplyr)
install.packages("plyr")
library(plyr)

# install and load ggplot2 for visualization customization
install.packages("ggplot2")
library(ggplot2)

# install and load Spanish language Universal Dependency model for annotation
langmodel_spanish = upipe_download_model(language="spanish")
langmodel_spanish = upipe_load_model(file = langmodel_spanish$file_model)

## show transcriptions in folder
# set directory path
input_txt.dir = "Textos_TxtFiles"

# create vector of file names. Include $ sign to make sure no hidden files retrieved
textos_txt_files.v = dir(input_txt.dir, "\\.txt$")

# print vector of file names in user-readable format
show.files = function(file.name.v){
  for(i in 1:length(file.name.v)){
    cat(i,file.name.v[i],"\n",sep=" ")
  }
}

# display list of plain-text files in folder
show.files(textos_txt_files.v)

## iterate on path directory to load in transcription corpus

corpus_load = function(files.v,input.dir){
  # set up empty list to hold the transcriptions
textos.word.vector.l = list()
# loop over the files
for(i in 1:length(files.v)){
    texto.v = read_file(file.path(input.dir,files.v[i]))
    # use index id from files.v vector as the 'name' in the list
textos.word.vector.l[[files.v[i]]] = texto.v
}
return(textos.word.vector.l)

# import loaded textos into a master corpus list
textos_txt_corpus.l = corpus_load(textos_txt_files.v,input_txt.dir)

# set output directory for writing out annotated transcription excel files
output_xlsx.dir = "Textos_XLSXFiles"

# annotate each textos file and write to excel file in output folder
for(i in 1:length(textos_txt_corpus.l)){
    annotated_texto = udpipe_annotate(langmodel_spanish,x=textos_txt_corpus.l[[i]])
    annotated_texto = as.data.frame(annotated_texto,detailed=TRUE)
    write.xlsx(annotated_texto,file=file.path(output_xlsx.dir,paste("memoria_",i,".xlsx",sep=""))
}

C.2 Chapter 4 Supplementary Code
C.2.1 XML Corpus Import
### Import XML Markup Files ###
# set input directory for parsing XML files
input_xml.dir = "Textos_XMLFiles"

# retrieve vector of XML file names
textos_xml_files.v = dir(input_xml.dir,"\.xml$")

# form empty list to hold corpus dataframes
textos_corpus.l = list()

## load in XML corpus
# generate a list which contains within it 72 data frames
# each data frame contains basic representation of the document with Doc_ID, UPOS, Lemma & Token
# segmented one token per row
for(i in 1:length(textos_xml_files.v)){
    memoria_tmp = read_xml(file.path(input_xml.dir,textos_xml_files.v[i]),encoding="UTF-8")
    ## make data frame of elements
    # find all word tags
textos_word_tags = xml_find_all(memoria_tmp,"//w")
    # extract the word tokens
textos_tokens = trimws(xml_text(textos_word_tags))
    # extract the lemmas
textos_lemmas = trimws(xml_attr(textos_word_tags,"lemma"))
    # extract the UPOS tags
textos_pos = trimws(xml_attr(textos_word_tags,"pos"))
    # extract the doc_ids
textos_doc_ids = trimws(xml_attr(textos_word_tags,"doc_id"))
    # extract the original (historical) tokens
textos_orig_tokens = trimws(xml_attr(textos_word_tags,"token_orig"))
    # fill data frame with transcription annotation data
    memoria_tmp.df = data.frame("Doc_ID"=textos_doc_ids,"Lemma"=textos_lemmas,
                              "UPOS"=textos_pos,"Orig_Token"=textos_orig_tokens,
                              "Token"=textos_tokens)
C.2.2 Principal Component Analysis
C.2.2.1 Lemma Implementation
### PCA Lemma Implementation ###
## Data preparation
# compile full corpus into single data frame, excluding punctuation
textos_corpus_unified.df = subset(join_all(textos_corpus.l, type="full"), UPOS!="PUNCT")

# find most frequent lemmas in the corpus by sorted count
sort(table(textos_corpus_unified.df$Lemma), decreasing=TRUE)
# final results: y, el, de, él, que, a, dicho, en, dar, uno, indio, ser, pueblo, con, otro

# define frequent lemmas in vector
top_tokens.v = c("y", "el", "de", "él", "que", "a", "dicho", "en", "dar", "uno", "indio", "ser", "pueblo", "con", "otro")

# form empty list to hold the relative frequency dataframes
memoria_freqs.l = list()

# populate list with 72 data frames, where each data frame is a scaled frequency measure with Doc_ID, tokens, relative frequency
for (i in 1:length(textos_xml_files.v)){
  mem.object = textos_corpus.l[[i]]
  memoria_lemmadata = count(filter(mem.object, UPOS!="PUNCT"), c("Doc_ID", "Lemma"))
  memoria_lemmadata$Doc_ID = as.numeric(as.character(memoria_lemmadata$Doc_ID))
  memoria_lemmadata$freq = memoria_lemmadata$freq/sum(memoria_lemmadata$freq)
  memoria_freqs.l[[textos_xml_files.v[i]]] = memoria_lemmadata
}

# form empty list to hold the frequent lemma relative frequencies
top_memoria_tokens.l = list()

# extract matching stop words from each frequency list and populate top lemma list
for (i in 1:length(textos_xml_files.v)){
  mem.object = memoria_freqs.l[[i]]
  mem.object.parsed = filter(mem.object, Lemma %in% top_tokens.v)
  # name each data frame with proper memoria label
  top_memoria_tokens.l[[substring(textos_xml_files.v[i], 1, 10)]] = mem.object.parsed
}

# full join frequency data frames into single data frame
# with 0s for non-appearing values
tokens_freqs.df = join_all(top_memoria_tokens.l, by="Doc_ID", type="full")

# remove missing factors from compiled data frame
tokens_freqs.df$Lemma = factor(tokens_freqs.df$Lemma)

# generate contingency table and populate with relative frequencies of top 15 lemmas
tokens_freqs_lftable = xtabs(freq ~ Doc_ID+Lemma, data=tokens_freqs.df)

# generate final frequency matrix from long-form table
tokens_freqs.m = apply(tokens_freqs_lftable, 2, as.numeric)

## Test appropriateness of PCA on the data

# install and load psych package for Bartlett's Test of Sphericity
install.packages("psych")
library(psych)
# run Bartlett test on correlation matrix
cor.test.bartlett(cor(tokens_freqs.m), n=nrow(tokens_freqs.m))
#$chisq
# [1] 445.0085
#$p.value
# [1] 2.096686e-43
#$df
# [1] 105

# run Kaiser-Meyer-Olkin test on correlation matrix
KMO(cor(tokens_freqs.m))
# Kaiser-Meyer-Olkin factor adequacy
# Call: KMO(r = cor(tokens_freqs.m))
# Overall MSA = 0.7
# MSA for each item =
#  el  a  dar  de  dicho  él  en  indio  otro  que  ser  uno  y
#  0.59  0.75  0.43  0.77  0.58  0.66  0.67  0.78  0.76  0.63  0.47  0.71
#  con  pueblo
#  0.82  0.83

# check correlation matrix determinant
det(cor(tokens_freqs.m))
# 0.001082183

## Load metadata for PCA implementation
# define input directory
memoria_input.dir = "Textos_Metadata/textos_metadata_final.xlsx"

# load metadata for PCA as data frame and select relevant columns (variables)
# with which to generate the subset data frame
memoria_metadata.df = read.xlsx(memoria_input.dir, header = TRUE, encoding = "UTF-8", 1)
metadata_subset.df = select(memoria_metadata.df, year, lat, lon, dist_cuzco, altitude, type, pre_post_conquest, language, int_unint, pop_s_m_l, num_speakers_sng_mult, currency, lat_zone)

# join normalized frequencies with the metadata
# set row names to memoria names
memoria_token_compiled.df = cbind(data.frame(tokens_freqs.m), metadata_subset.df)
rownames(memoria_token_compiled.df) = names(top_memoria_tokens.l)

## PCA lemma implementation
# install and load FactoMineR package for PCA
install.packages("FactoMineR")
library(FactoMineR)

# install and load factoextra package for added visualizations
install.packages("factoextra")
library(factoextra)

# perform PCA with designated qualitative and quantitative supplementary variables
mem_tok.pca = PCA(memoria_token_compiled.df, quanti.sup = 16:20, quali.sup = 21:28, graph = FALSE)
# display cumulative percentage of total variance (table 4.1)
head(mem_tok.pca$eig)

# generate and export scree plot (figure 4.1)
tiff("figures/4.1_lemma_scree_plot.tiff", units="in", width=6.5, height=5.25, res=500)
fviz_eig(mem_tok.pca, choice="variance", addlabels = TRUE,
         main="", xlab = "Component",
         ylab = "Percentage of Explained Variance",
         linecolor = "red")
dev.off()

# generate and export variable factor map along dimensions 1 and 2 (figure 4.2)
tiff("figures/4.2_lemma_var_factor_1_2.tiff", units="in", width=6.5, height=5.25, res=500)
plot(mem_tok.pca, choix = "var", axes = c(1,2), cex=0.8, title="")
dev.off()

# generate and export variable factor map along dimensions 3 and 4 (figure 4.3)
tiff("figures/4.3_lemma_var_factor_3_4.tiff", units="in", width=6.5, height=5.25, res=500)
plot(mem_tok.pca, choix = "var", axes = c(3,4), cex=0.8, title="")
dev.off()

# generate and export plot of individuals (figure 4.4)
# scale and color by contribution of individuals
# generate and export plot of 99% confidence ellipses (figure 4.5)
tiff("figures/4.5_lemma_conf_ellipses.tiff", units="in", width=11.25, height=7.6875, res=600)
plotellipses(mem_tok.pca, level=0.99)
dev.off()

# generate and export biplot (figure 4.6)
# include only highly-contributing variables
# generate and export pre/postconquest distinction (figure 4.7)
tiff("figures/4.7_lemma_pre_post.tiff", units="in", width=6.5, height=4.2, res=600)
plot(mem_tok.pca, cex=0.8, col.ind = "grey", select = "contrib 20",
     habillage = "pre_post_conquest", invisible="quali", title="",
     xlim = c(-5,7))
dev.off()

# generate and export currency distinction (figure 4.8)
tiff("figures/4.8_lemma_currency.tiff", units="in", width=6.5, height=4.2, res=600)
```r
plot(mem_tok.pca,cex=1.8,
     habillage = "currency",invisible="quali",title="",
     xlim = c(-4,6),label="none")
dev.off()

# generate and export parallel analysis (figures appendix)
tiff("figures/App1_lemma_par_analysis.tiff", units="in", width=6.5, height=5.25, res=600)
fa.parallel(cor(tokens_freqs.m),n.obs=72, fa = "pc", n.iter = 10000,
            quant = 0.95, error.bars = TRUE, se.bars = FALSE,
            main = "Parallel Analysis with 10,000 Simulations",
            ylab = "Eigenvalues of Principal Components")
dev.off()

# print factor loadings using dimdesc
dimdesc(mem_tok.pca,axes=c(1:4),proba = 0.01)

# print variable contributions
write.xlsx(data.frame(mem_tok.pca$var$contrib),file="testers/lemma_contras.xls")

# print individual contributions
write.xlsx(data.frame(mem_tok.pca$ind$contrib),file="testers/lemma_ind_contras.xls")

# print correlation matrix
write.xlsx(cor(memoria_token_compiled.df[,1:15]),file="testers/lemma_cor.xls")

# using installed and loaded PerformanceAnalytics package
install.packages("PerformanceAnalytics")
library(PerformanceAnalytics)

# generate correlation plot (figures appendix)
tiff("figures/App1_lemma_var_corplot.tiff", units="in", width=6.5, height=5.25, res=600)
chart.Correlation(memoria_token_compiled.df[,1:15],
                  histogram = TRUE,pch=19)
dev.off()

C.2.2.2 UPOS Implementation
### PCA UPOS Implementation ###
#### Data preparation

#### Form empty list for UPOS tags and Doc_IDs

```
# generate contingency table and populate with relative frequencies of UPOS categories
upos_freqs_lftable = xtabs(freq ~ Doc_ID + UPOS, data=upos_freqs.df)

# generate final frequency matrix from long-form table
upos_freqs.m = apply(upos_freqs_lftable,2,as.numeric)

## Test appropriateness of PCA on the data
# install and load psych package for Bartlett's Test of Sphericity
# (in case doing analysis before lemma implementation)
install.packages("psych")
library(psych)

# run Bartlett test on correlation matrix and exclude multicollinear columns
# (see implementation)
cortest.bartlett(cor(upos_freqs.m[,c(1,3:4,6,8:12)]),n=nrow(upos_freqs.m))
#$chisq
# [1] 305.1013
#$p.value
# [1] 2.322673e-44
#$df
# [1] 36

# run Kaiser-Meyer-Olkin test on correlation matrix
KMO(cor(upos_freqs.m[,c(1,3:4,6,8:12)]))
# Kaiser-Meyer-Olkin factor adequacy
# Call: KMO(r = cor(upos_freqs.m[, c(1, 3:4, 6, 8:12)]))
# Overall MSA = 0.72
# MSA for each item =
# DET   ADJ   ADP   AUX  NOUN   NUM  PRON SCONJ  VERB
#0.61  0.67  0.68  0.53  0.69  0.72  0.75  0.74  0.87

# check correlation matrix determinant
det(cor(upos_freqs.m[,c(1,3:4,6,8:12)]))
# 0.01064728

## Load metadata for PCA implementation
# define input directory
memoria_input.dir = "Textos_Metadata/textos_metadata_final.xlsx"

# load metadata for PCA as data frame and select relevant columns (variables)
# with which to generate the subset data frame
memoria_Metadata.df = read.xlsx(memoria_input.dir,header = TRUE,encoding = "UTF-8",1)
metadata_subset.df = select(memoria_Metadata.df,year,lat,lon,dist_cuzco,altitude,
type,pre_post_conquest,language,int_unint,
pop_s_m_l,num_speakers_sng_mult,currency, lat_zone)

# join normalized UPOS frequencies with the metadata
# set row names to memoria names
memoria.POS_compiled.df = cbind(data.frame(upos_freqs.m[,c(1,3:4,6,8:12)]),metadata_subset.df)
rownames(memoria.POS_compiled.df) = names(memoria.POS_freqs.l)

## PCA UPOS implementation
# install and load FactoMineR package for PCA
install.packages("FactoMineR")
library(FactoMineR)

# install and load factoextra package for added visualizations
install.packages("factoextra")
library(factoextra)

# perform PCA with designated qualitative and quantitative supplementary variables
mem_POS.pca = PCA(memoria_POSCompiled.df,quanti.sup = 10:14,
                    quali.sup = 15:22,graph = FALSE)

# display cumulative percentage of total variance (table 4.2)
head(mem_POS.pca$eig)

# generate and export scree plot (figure 4.9)
tiff("figures/4.9_upos_scree_plot.tiff", units="in", width=6.5, height=5.25, res=500)
fviz_eig(mem_POS.pca,choice="variance",addlabels = TRUE,
         main="",xlab = "Component",
         ylab = "Percentage of Explained Variance",
         linecolor = "red")
dev.off()

# generate and export variable factor map along dimensions 1 and 2 (figure 4.10)
tiff("figures/4.10_upos_var_factor_1_2.tiff", units="in", width=6.5, height=5.25, res=500)
plot(mem_POS.pca,choix = "var",cex=0.8,title="")
dev.off()

# generate and export plot of individuals (figure 4.11)
# scale and color by contribution of individuals
# generate and export 95% confidence ellipses (figure 4.12)
# generate and export biplot (figure 4.13)
# include only highly-contributing variables
# generate and export pre/postconquest distinction (figure 4.14)
# generate and export currency distinction (figure 4.15)
plot(mem_POS.pca, cex = 2, 
  habillage = "currency", invisible = "quali", select = "contrib 20",
  title = "", label = "none")
dev.off()

# generate and export parallel analysis (figures appendix)
tiff("figures/App1_upos_par_analysis.tiff", units = "in", width = 6.5, height = 5.25, res = 600)
fa.parallel(cor(upos_freqs.m[, c(1, 3:4, 6, 8:12)]), n.obs = 72, fa = "pc", n.iter = 10000,
  quant = 0.95, error.bars = TRUE, se.bars = FALSE,
  main = "Parallel Analysis with 10,000 Simulations",
  ylab = "Eigenvalues of Principal Components")
dev.off()

# print factor loadings using dimdesc
dimdesc(mem_POS.pca, axes = c(1, 2), proba = 0.01)

# generate and export variable contributions to dimension 1 (figure 4.21)
tiff("figures/4.21_upos_dim1_var_contribs.tiff", units = "in", width = 6.5, height = 5.25,
  res = 600)
fviz_contrib(mem_POS.pca, choice = "var",
  axes = 1, title = "")
dev.off()

# print variable contributions
write.xlsx(data.frame(mem_POS.pca$var$contrib), file = "testers/upos_contribs.xlsx")

# print correlation matrix
write.xlsx(cor(memoria_POS_compiled.df[, 1:9]), file = "testers/upos_cor.xlsx")

# using installed and loaded PerformanceAnalytics package
install.packages("PerformanceAnalytics")
library(PerformanceAnalytics)

# generate and export correlation plot (figures appendix)
tiff("figures/App1_upos_var_corrplot.tiff", units = "in", width = 6.5, height = 5.25, res = 600)
chart.Correlation(memoria_POS_compiled.df[, 1:9],
  histogram = TRUE, pch = 19)
dev.off()

# print individual contributions
write.xlsx(data.frame(mem_POS.pca$ind$contrib), file = "testers/upos_ind_contribs.xlsx")

C.2.3 Hierarchical Agglomerative Clustering
C.2.3.1 Lemma Implementation

### Hierarchical Clustering Implementation ###
# install and load dendextend package for analysis and visualization of dendrograms
install.packages("dendextend")
library(dendextend)

# install and load colorspace package for coloring legend and dendrograms
install.packages("colorspace")
library(colorspace)

# Lemma implementation
# create distance object using distance method: Euclidean
dm_tok.m = dist(tokens_freqs.m, method = "eucl"

# complete hierarchical clustering on distance object
# using Ward's Method for amalgamation rule
tok.hclust = hclust(dm_tok.m, method="ward.D")

# use transcription names as labels
tok.hclust$labels = substring(names(memoria_freqs.l),1,10)

# store the hierarchical lemma-based clustering as a dendrogram for subsequent analysis
tok.dend = as.dendrogram(tok.hclust)

# generate pre and postconquest chronological legend
pre_post.legend = rev(levels(memoria_token_compiled.df$pre_post_conquest))

# color lemma hierarchical clustering according to bipartite theoretical ideal
tok.hclust = color_branches(tok.hclust, k=2, col = diverge_hcl(2))

# color leaves (labels) according to actual chronological (pre/postconquest) value
labels_colors(tok.hclust) =
   diverge_hcl(2)[sort_levels_values{
       as.numeric(memoria_token_compiled.df$pre_post_conquest)
       [order.dendrogram(tok.hclust)]
   }]

# generate and export the hierarchical clustering with corrected legend (figure 4.17)
tiff("figures/4.17_lemma_dendrogram_pre_post.tiff", units="in", width=11.25, height=7.6875, res=600)

plot(tok.hclust,cex.lab=0.3)
legend("topleft", legend = pre_post.legend, fill = diverge_hcl((2)))

device()

C.2.3.2 UPOS Implementation

## UPOS implementation

# create distance object using distance method: Euclidean
dm_pos.m = dist(upos_freqs.m, method = "euclidean")

# complete hierarchical clustering on distance object
# using Ward's Method for amalgamation rule
pos.hclust = hclust(dm_pos.m, method="ward.D")

# use transcription names as labels
pos.hclust$labels = substring(names(memoria_freqs.l),1,10)

# store the hierarchical UPOS-based clustering as a dendrogram for subsequent analysis
pos.dend = as.dendrogram(pos.hclust)

# generate pre and postconquest chronological legend
pre_post.legend = rev(levels(memoria_token_compiled.df$pre_post_conquest))

# color UPOS hierarchical clustering according to bipartite theoretical ideal
pos.hclust = color_branches(pos.hclust, k=2, col = diverge_hcl(2))

# color leaves (labels) according to actual chronological (pre/postconquest) value
labels_colors(pos.hclust) =
   diverge_hcl(2)[sort_levels_values{
       as.numeric(memoria_token_compiled.df$pre_post_conquest)
       [order.dendrogram(pos.hclust)]
   }]

# generate and export the hierarchical clustering with corrected legend (figure 4.18)
tiff("figures/4.18_upos_dendrogram_pre_post.tiff", units="in", width=11.25, height=7.6875, res=600)
C.2.4 Hierarchical Clustering Comparison

```r
plot(pos.hclust, cex=0.3)
legend("topleft", legend = pre_post.legend, fill = diverge_hcl((2)))
dev.off()
```

### Dendrogram/hierarchical clustering comparison ###

- Load and install profdpm package for Polkes-Mallow index analysis:
  ```r
  install.packages("profdpm")
  library(profdpm)
  ```

- Print entanglement of original trees:
  ```r
  entanglement(tok.dend, pos.dend)
  # 0.4265654
  ```

- Plot and export tanglegram (figure 4.19):
  ```r
  tiff("figures/4.19_tanglegram.tiff", units="in", width=11.25, height=7.6875, res=600)
  plot(untangle(tok.dend, pos.dend, method = "step2side"),
       common_subtrees_color_branches = TRUE)
  dev.off()
  ```

- Print Fowlkes-Mallows Index under expectation:
  ```r
  FM_index_R(cutree(pos.hclust, k = 2), cutree(tok.hclust, k=2))
  # [1] 0.6843131
  # attr(,"E_FM")
  # [1] 0.5031147
  # attr(,"V_FM")
  # [1] 9.637995e-05
  ```

- Generate and export (k,Bk) plot:
  ```r
  Bk_plot(pos.dend, tok.dend, rejection_line_permutation = TRUE, conf.level = 0.99, R = 10000,
          main="(k,Bk) Plot under H0 with 10,000 Replications")
  dev.off()
  ```

- Baker's Gamma plots (in appendix A):
  ```r
  pos_tok.l = dendlist(pos.dend, tok.dend)
  self_cor = cor_bakers_gamma(tok.dend, tok.dend)
  between_cor = cor_bakers_gamma(pos_tok.l)
  # between_cor = 0.371
  ```

- Permutation test implementation:
  ```r
  R=1000
  ```

- Form empty vector to hold Baker's gamma indices:
  ```r
  bakers_gamma_samples.v = numeric(R)
  ```

- Designate dendrogram to be shuffled:
  ```r
  dend_shuffle = tok.dend
  ```

- Perform R permutations on the leaves of the shuffled dendrograms:
  ```r
  for(i in 1:R){
    dend_shuffle = sample.dendrogram(dend_shuffle,
    ...}
  ```
replace = FALSE)
bakers_gamma_samples.v[i] = cor_bakers_gamma(tok.dend,
    dend_shuffle)
}

# visualize permutation test results
# plot and export density of Baker's gamma indices under H0 (R permutations)
tiff("figures/App1_permtest_baker.tiff", units="in", width=6.5, height=5.25, res=600)
plot(density(bakers_gamma_samples.v),
    main = "Baker's Gamma Distribution under H0 with 1,000 Permutations",
    xlim = c(-1,1.1))
# plot zero correlation, theoretical self-correlation
# and observed Baker's Gamma
abline(v = 0, lty = 2)
abline(v = self_cor, col=2)
abline(v = between_cor, lty = 2, col = 4)
# include legend in plot
legend("topleft", legend = c("self_cor", "between_cor"), fill = c(2,4))
# generate plot subtitle to display permutation test results
title(sub = paste("One-sided p-value:",
    "self_cor =",sum(self_cor < bakers_gamma_samples.v)/R,
    "; between_cor =",sum(between_cor < bakers_gamma_samples.v)/R))
dev.off()

## Bootstrap sampling for Baker's gamma confidence interval
# store labels from dendrograms for shuffling
pos.dend_labels = labels(pos.dend)
tok.dend_labels = labels(tok.dend)

# define bootstrap sample size
N = 1000
# form vector for bootstrap-sampled indices
bakers_gamma_bootstrap.v = numeric(N)

# perform Baker's gamma bootstrap sampling on shuffled trees
for(i in 1:N)
    label_samples = sample(pos.dend_labels,replace=TRUE)
pos.dend_mixed = sample.dendrogram(pos.dend,
    dend_labels = pos.dend_labels,
    fix_members = TRUE,
    fix_order = TRUE,
    fix_midpoint = FALSE,
    replace = TRUE,
    sampled_labels = label_samples)

tok.dend_mixed = sample.dendrogram(tok.dend,
    dend_labels = tok.dend_labels,
    fix_members = TRUE,
    fix_order = TRUE,
    fix_midpoint = FALSE,
    replace = TRUE,
    sampled_labels = label_samples)
bakers_gamma_bootstrap.v[i] = cor_bakers_gamma(pos.dend_mixed,
    tok.dend_mixed,
    warn = FALSE)

# extract 95% confidence interval
ci95 = quantile(bakers_gamma_bootstrap.v, probs=c(.025,.975))

# ci95
# 2.5% 97.5%
# 0.3247357 1.0000000

plot and export bootstrap sample density with observed index value
tiff("figures/App1_bootstrap_baker.tiff", units="in", width=6.5, height=5.25, res=600)
plot(density(bakers_gamma_bootstrap.v),
     main = "Baker's Gamma Bootstrap Distribution with 1,000 Replications",
     xlim = c(-0.5,1))
abline(v = ci95, lty = 2, col = 3)
abline(v = max(bakers_gamma_samples.v), lty = 2, col = 4)
abline(v = cor_bakers_gamma(pos_tok.l), lty = 2, col = 2)
legend("topleft", legend = c("95% CI", "Baker's Gamma Index"), fill = c(3,2))
dev.off()

## Mantel test on cophenetic correlations implementation
# install and load vegan package for Mantel test
install.packages("vegan")
library(vegan)
# implement Mantel test on cophenetic correlations
cor(cophenetic(pos.hclust), cophenetic(tok.hclust))
mantel(cophenetic(pos.hclust), cophenetic(tok.hclust))

C.3 Metadata Appendix, Summary Statistics and Transcription Map Supplementary Code

C.3.1 Transcription Metadata Appendix
# import and load sjPlot package for table output
install.packages("sjPlot")
library(sjPlot)
# output metadata table with alternating-shaded rows
tab_df(memoria_metadata.df, title = "Memoria Index",
       alternate.rows = TRUE,
       file = "testers/memoria_metadata.doc")

C.3.2 Summary Statistics
### Stats for summary statistics (table 3.1)
lemma_table_stats.df = textos_corpus_unified.df
lemma_table_stats.df$Doc_ID = as.numeric(lemma_table_stats.df$Doc_ID)

# total word tokens
lemma_table_stats.df$Token = factor(lemma_table_stats.df$Token)
length(sort(table(lemma_table_stats.df$Token), decreasing = TRUE))

# total lemmas
lemma_table_stats.df$Lemma = factor(lemma_table_stats.df$Lemma)
length(sort(table(lemma_table_stats.df$Lemma), decreasing = TRUE))

# mean (distinct) lemmas per text
lemma_distinct.v = numeric(length(textos_xml_files.v))
lemma_totals.v = numeric(length(textos_xml_files.v))
for(i in 1:length(textos_xml_files.v)){
    memoria.tmp.df = subset(lemma_table_stats.df, Doc_ID==i)
    memoria.tmp.df$Lemma = factor(memoria.tmp.df$Lemma)
    # total lemmas for transcriptions
    lemma_totals.v[i] = length(memoria.tmp.df$Lemma)
    # distinct lemmas for transcriptions
    lemma_distinct.v[i] = length(table(memoria.tmp.df$Lemma))
}
# generate summary statistics
mean(lemma_totals.v)
median(lemma_totals.v)
SD(lemma_totals.v)
mean(lemma_distinct.v)
median(lemma_distinct.v)
SD(lemma_distinct.v)

C.3.3 Transcription Map
### Map implementation ###
# install and load ggplot2 package for visualizations
install.packages("ggplot2")
library(ggplot2)
# install and load ggmap package for map retrieval
install.packages("ggmap")
library(ggmap)
# install and load ggsn package for additional map symbols
install.packages("ggsn")
library(ggsn)
# load ggrepel package for text label placements
library(ggrepel)

# define input directory
memoria_input.dir = "Textos_Metadata/textos_metadata_final.xlsx"

# load metadata as data frame
memoria_metadata.df = read.xlsx(memoria_input.dir,header = TRUE,encoding = "UTF-8",1)

## Create map with memoria markers
# generate transcription labels
memoria_numbers.v = 1:72
# isolate longitudes and latitudes from metadata data frame
lon_memoria.v = as.numeric(memoria_metadata.df$lon)
lat_memoria.v = as.numeric(memoria_metadata.df$lat)

# bind coordinates for each point
cords_memoria.df = as.data.frame(cbind(memoria_numbers.v,lon_memoria.v,lat_memoria.v))

# register google API key to initialize queries
register_google(key = ***suppressed***)

# retrieve transcription map plot using center of coordinates
memoria_map = get_map(location = c(lon = mean(cords_memoria.df$lon_memoria.v)-2,
                                   lat = mean(cords_memoria.df$lat_memoria.v)+1.5),
                        zoom = 5, maptype = "terrain", scale = 2)

# retrieve map of south america for inset
south_am_map = get_map(location = c(lon = -60.4915,
                                    lat = -15.7832),source="stamen",
                        zoom = 4, maptype = "terrain", scale = 2)

# plot and export the compiled maps
tiff("figures/2.1_memoria_locations.tiff", units="in", width=6.5, height=5.25, res=900)
# generate south america inset map
south_am_zoom = ggmap(south_am_map,exten="device")+
  theme(panel.border = element_rect(colour = "black", fill=NA, size=1))+
  guides(fill=FALSE, alpha=FALSE, size=FALSE, colour=FALSE)
south_am_grob = ggplotGrob(south_am_zoom)
# plot memoria map with inset south america map and repelled labels

```r
ggmap(memoria_map) +
  geom_point(data = coords_memoria.df,aes(x = lon_memoria.v, y = lat_memoria.v, fill = "red"), size = 3, shape = 21, alpha = 0.8, stroke = 0.3) +
  geom_text_repel(data = coords_memoria.df,aes(x = lon_memoria.v, y = lat_memoria.v, label = memoria_numbers.v, hjust = 0, vjust = 0),
                  segment.size = 0.1, segment.alpha = 0.8, force = 4, size = 2.5) +
  guides(fill = FALSE, alpha = FALSE, size = FALSE) +
  xlab("Longitude") + ylab("Latitude") +
  inset(south_am_grob, xmin = -87, xmax = -76, ymin = -23.4, ymax = -15.6)

# include north arrow
north2(memoria_map, x = 0.25, y = 0.9, scale = 0.09, symbol = 1)
dev.off()
```