UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

UNIDAD ACADÉMICA PROFESIONAL TIANGUISTENCO

"Segmentación de líneas de texto en documentos manuscritos antiguos independiente del lenguaje"

Tesis
Para Obtener el Grado de
Maestro en Ciencias de la Computación

Que presenta
Ing. Miguel Ángel García Calderón

Tutor académico:
Dr. René Arnulfo García Hernández

Tutores adjuntos:
Dra. Yulia Ledeneva
Dr. José Luis Tapia Fabela

TIANGUISTENCO, MÉX. NOVIEMBRE 2017

UAEM | Universidad Autónoma del Estado de México

DICTÁMEN DE AUTORIZACIÓN Y OBTENCIÓN DE GRADO DE MAESTRÍA

Tlanguistenco, Méx., a 28 de noviembre de 2017

Título del proyecto:

Segmentación de líneas de texto en documentos manuscritos antiguos independiente del lenguaje.

Tesista:

Ing. Miguel Ángel García Calderón

Dictamen: **No. de revisión:** 6

- ☐ **Rechazado**
- ☐ **Sujeto a modificaciones**
- ☐ **Aceptado, condicionado**
- ■ **Aceptado**

Observaciones generales:

Aceptado para la impresión
Aceptado para la defensa de grado

Tutor Adjunto	Tutor Académico	Tutor Adjunto
Dra. Yulia Nikolaevna Ledeneva	Dr. René Arnulfo García Hernández	Dr. José Luis Tapia Fabela

Agradecimientos

Gracias por datarme con sabiduría e inteligencia, por regalarme una familia que me quiere y apoya en todos los aspectos sin importar si son éxitos o fracasos. Gracias por mostrarme el camino el camino a seguir cuando todo parece imposible para mi. Gracias Dios.

Agradezco a las únicas personas que han estado toda mi vida conmigo, por darme la mejor herencia que alguien puede recibir, por educarme con el ejemplo del respeto, la dedicación, la fe y la esperanza. Papá y mamá.

Por compartir la mayor parte de su conocimiento, porque gracias a usted tuve el primer acercamiento a la investigación, gracias por compartir sus ideas. Siempre será un gran maestro. Dr. René.

Por compartir anécdotas y experiencias del cómo ser investigador, por cada uno de sus consejos y sugerencias durante mis estudios. Por sus correciones, por el apoyo recibido en la escritura y traducción del artículo resultante de este trabajo. Dra. Yulia

Por todo el amor, apoyo y comprensión, por creer en mi, por hacerme reír tanto. Mi inspiración y motivación. Diana Natalie.

Por estar al pendiente de las fechas de conciertos, por encargarse de comprar o preparar la cena mientras yo trabajaba. Mi hermano Ulises.

Por permitirme ser un ejemplo para él, por las tardes de videojuegos. Mi hermano Diego

Gracias a mis compañeros de posgrado, por su ayuda en la realización del material necesario para realizar experimentación y concluir este trabajo. Sin tanto apoyo no habría terminado este trabajo a tiempo.

Gracias por la plantilla proporcionada por el Dr. René para la escritura de esta tesis.

Agradecimiento especial al Consejo Nacional de Ciencia y Tecnología (CONACyT) por el apoyo económico otorgado durante mis estudios de posgrado en el programa Maestría en Ciencias de la Computación a través de la beca con número CVU: 712165. Y a la red temática de tecnologías del lenguaje por el apoyo recibido para la elaboración del corpus resultante.

Resumen

Hasta el momento no se ha utilizado todo el conocimiento que hay en los manuscritos antiguos debido a que reconocimiento de texto manuscrito aún no cuenta con métodos robustos para esta tarea.

El primer problema de los métodos para el reconocimiento de texto manuscrito es que requieren que el texto se encuentre dividido en líneas. Los métodos actuales para la segmentación de líneas de texto manuscrito no han sido optimizados para trabajar con manuscritos antiguos.

La primera etapa de la Segmentación de Líneas de Texto (SLT) manuscrito consiste en la Localización de Líneas de Texto (LLT). Para la SLT se han propuesto métodos que buscan los valores máximos locales en un histograma. El problema para estos métodos es que existen demasiados máximos locales, lo cual no permite localizar las líneas que hay.

La segunda etapa de la SLT en manuscritos antiguos consiste en la búsqueda de una ruta que permita separar las líneas de texto, el problema de los métodos actuales es que algunos realizan una búsqueda local de la ruta y los otros métodos buscan la ruta evitando pasar por la mayor cantidad de caracteres.

En este trabajo se presenta un sistema compuesto por dos nuevos métodos para la LLT manuscrito y otro método para la Búsqueda de una Ruta que permita Segmentar Líneas de Texto en documentos manuscritos (BRSLT) que supera a los métodos analizados en el estado del arte en las dos etapas. En el primer método propuesto se presenta la extracción de un mapa de energía que incrementa las diferencias entre los máximos y mínimos locales en un histograma. El segundo método propuesto consiste en buscar la mejor ruta para segmentar líneas de texto manuscrito antiguo usando un algoritmo genético.

Para evaluar la exactitud de los métodos propuestos se han realizado experimentos con dos colecciones de documentos. Se ha realizado una evaluación independiente de los dos métodos propuesto. Las colecciones de documentos incluyen los idiomas: español, chino, árabe, inglés, árabe-español con escritura moderna y escritura antigua.

Con los resultados de la experimentación se ha demostrado que es posible mejorar la LLT implementando un mapa de energía que incremente las diferencias entre máximos y mínimos locales. Los experimentos de la segunda sección demuestran que es necesario realizar una optimización global de la ruta para segmentar líneas de texto.

Contenido

Página

Lista de figuras

Lista de tablas

CAPÍTULO 1.

Introducción

1.1 Antecedentes

Desde hace tiempo el hombre ha tenido la necesidad de comunicar, transmitir y almacenar sus necesidades, pensamientos y conocimiento. El conocimiento es almacenado para que perdure por varias generaciones (Rendón Rojas, n.d.). El primer medio para transmitir el conocimiento de una generación a otra fue a través del lenguaje natural.

Desde que el hombre vivía en cavernas comenzó a almacenar el conocimiento mediante un lenguaje basado en dibujos. Con el paso de los siglos, las pinturas rupestres fueron evolucionaron hasta la creación de las primeras formas de escritura en el año 3200 A.C (Baines, Bennet, & Houston, 2008). Al igual que las pinturas rupestres, la escritura también cambia con el paso del tiempo; algunos lenguajes y estilos de escritura desaparecen y son tomados como base para nuevas formas de escritura.

Con la invención de la escritura se incrementó la cantidad de información almacenada. Además, se crearon plantillas de trazos para definir un conjunto de caracteres (Bagley, 2004). Los primeros materiales usados para almacenar información de manera manuscrita fueron piedras, papiros, pergaminos, tablillas, cuero y papel.

Cabe resaltar que la escritura antigua se diferencia de la escritura moderna debido a que en el primer caso no se encuentran algunos símbolos especiales que ahora se conocen, por ejemplo: @, #, etc. Algunos de estos documentos tienen faltas de ortografía debido a que no existían esas reglas. Además, algunos de los documentos antiguos no contienen signos de puntuación debido a que la escritura y el lenguaje cambian con el paso del tiempo (Muñoz y Rivero, 1880; Sumano, 2002).

En la escritura de los documentos de los siglos XVI al XVIII se puede distinguir que la ortografía de esa época difiere de la ortografía actual. En cada época se tienen diferentes patrones de caligrafía y ortografía. En algunos casos es posible notar que existe una irregularidad en el uso de mayúsculas y minúsculas o la ausencia del acento (Sumano, 2002).

En algunos documentos se añadía algún tipo de ornamentación que permitiera identificar al autor, la autenticidad, la fecha o el área geográfica en donde se escribió el documento. Algunos métodos de ornamentación son dibujos, sellos, firmas, etc. (Sumano, 2002). En la Figura 1.1 se proporciona un ejemplo de documento manuscrito que contiene un sello distintivo de la biblioteca en donde se almacenó el documento.

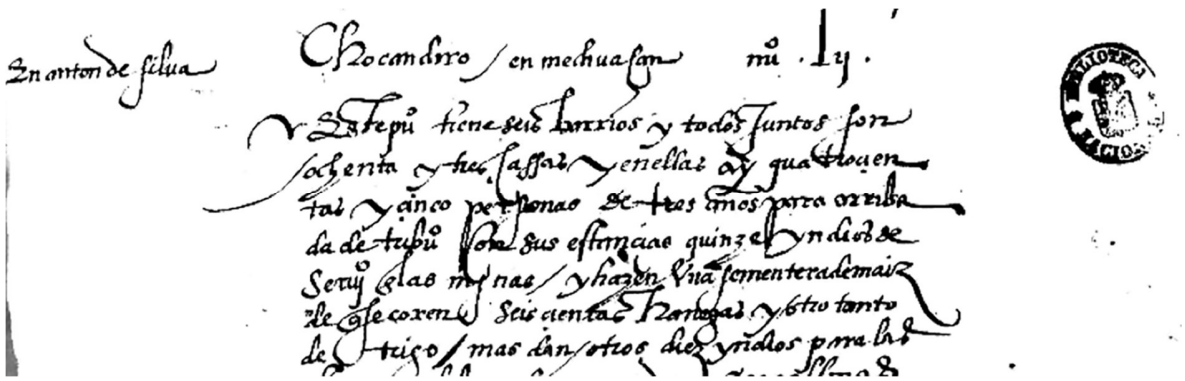


Figura 1.1. Ejemplo de documento manuscrito en español que no contiene signos de puntuación y como ornamentación contiene un sello en la parte superior derecha.

Otra de las variaciones presentes en los documentos manuscritos antiguos es el color de tinta usado trazar los caracteres. En la antigüedad se utilizaron pigmentos orgánicos para escribir documentos, posteriormente comenzaron a usar distintas mezclas de minerales, cada escribano o institución utilizaba diferentes pigmentos que permitieran identificar al autor del documento (Muñoz y Rivero, 1880). En la Figura 1.2 se muestra un ejemplo de documento que contiene mezclas de diferentes tintas y materiales.

Figura 1.2. Ejemplo de documento manuscrito escrito con diferentes materiales de tinta (*Voynich Manuscript, Beinecke MS 408, General Collection*, 1912).

Aún entre los hombres de letras existieron diferencias en la caligrafía. Además, estas diferencias se incrementaron cuando personas con escasa preparación redactaban los documentos como en el caso de los escribanos de la Nueva España del siglo XVI (Sumano, 2002).

En los documentos manuscritos antiguos es natural el deterioro debido al paso del tiempo, por lo que es necesario preservar el conocimiento histórico y cultural de obras antiguas que están bajo resguardo de diferentes bibliotecas y universidades del mundo.

Cada obra manuscrita antigua puede contener conocimiento teológico, cultural, histórico, literario y científico de una civilización o región (Gray, 1948; Gy\Hory, 2008). Por ello, es necesario desarrollar herramientas que faciliten el proceso de búsqueda y localización de información en documentos manuscritos antiguos y así, evitar dañar las obras durante su estudio (Mauricio, Alejandro, Joan-Andreu, & Enrique, n.d.).

Actualmente, muchas de las bibliotecas y universidades del mundo están interesadas en proporcionar el acceso a documentos manuscritos mediante la indexación y creación de plataformas semi-automáticas para la *recuperación de información* (Mauricio et al., n.d.).

Se han creado proyectos y plataformas para facilitar la transcripción manual de documentos manuscritos (Causer & Wallace, 2012). Un ejemplo de estas plataformas es el proyecto *Transcribed Bentham* en el que 1,009 manuscritos fueron transcritos por 1,207 personas en un periodo de 6 meses (Causer & Wallace, 2012).

Uno de los problemas de estas plataformas se presenta cuando un documento contiene más de un lenguaje, por lo que es necesario que el humano conozca todos los lenguajes presentes en el documento. Un ejemplo de documento con más de un lenguaje es la piedra roseta que

permitió la traducción de los jeroglíficos egipcios en el siglo XVII (Medina Morán, 2011). En la Figura 1.3 se muestra un segmento de la imagen de la piedra roseta. Las primeras dos líneas contienen jeroglíficos egipcios y en la parte media escritura demótica y en la parte inferior griego antiguo.

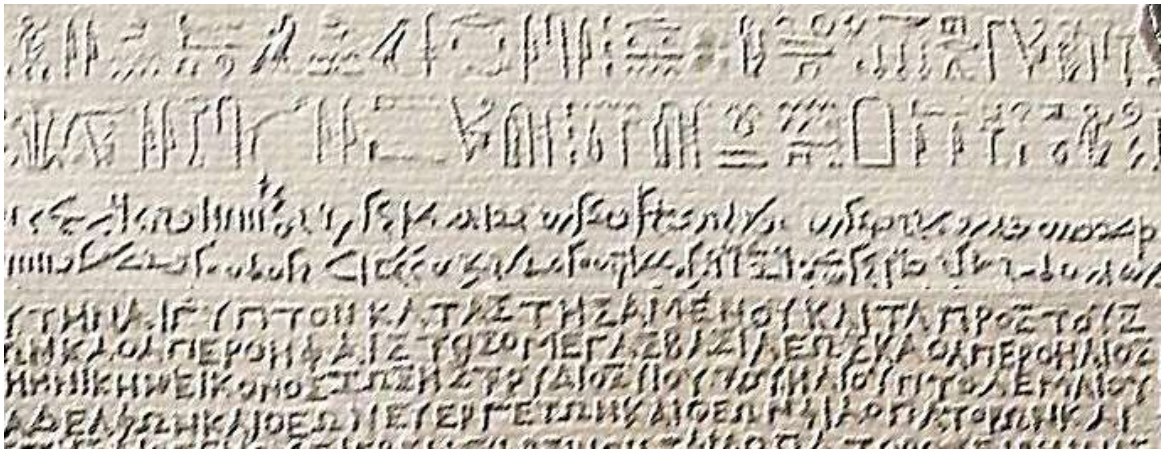


Figura 1.3. Segmento de seis líneas de texto de la piedra roseta.

Es por eso que existe la necesidad de generar sistemas que permitan analizar imágenes de documentos con el objetivo de Reconocer Texto Manuscrito (RTM). Los sistemas existentes para la transcripción de documentos manuscritos necesitan como entrada la imagen de una línea de texto a transcribir.

Por otro lado, los métodos actuales para la Segmentación de Líneas de Texto (SLT) necesitan localizar y posteriormente extraer las líneas de texto a partir de la imagen de una página. La primera etapa de la tarea SLT es la Localización de las Líneas de Texto (LLT) donde se determina el punto de inicio y fin de cada línea de texto. A partir de los puntos de la etapa LLT es necesaria la Búsqueda de una Ruta que Separe Líneas (BRSL) de texto manuscrito vecinas (Steinherz, Rivlin, & Intrator, 1999).

El problema de RTM y SLT es que la escritura contenida en los documentos es acorde a la región, tiempo, lenguaje y estilo del autor (Khandelwal et al., 2009). Dependiendo del autor se puede ver que existen muchas variaciones como la forma del carácter, el tamaño de los caracteres, caracteres que se intersectan, espacio entre caracteres, espacio entre líneas y líneas que se solapan, ornamentación y texto rayado (véase Figura 1.4, 1.5) (Gray, 1948; Muñoz y Rivero, 1880; Sumano, 2002). En el anexo 1 se muestran documentos con todas las variaciones descritas anteriormente.

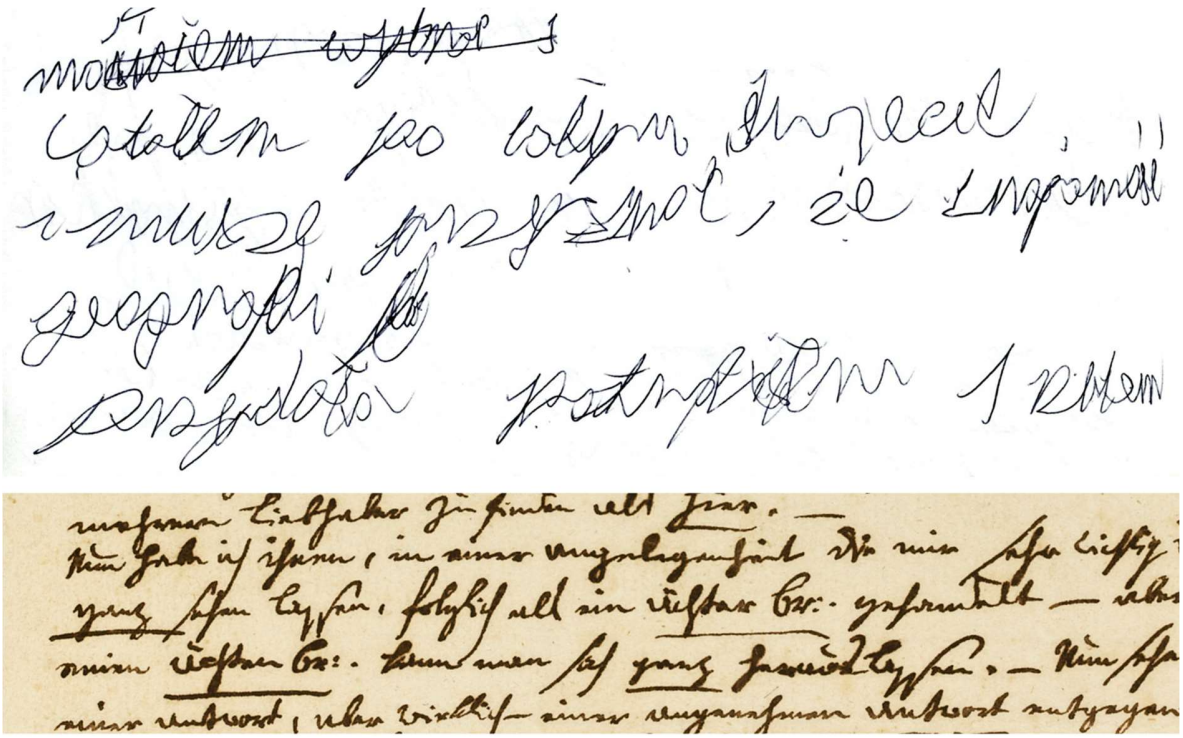


Figura 1.4. Ejemplos de documentos con las diferentes variaciones que se presentan en los documentos manuscritos.

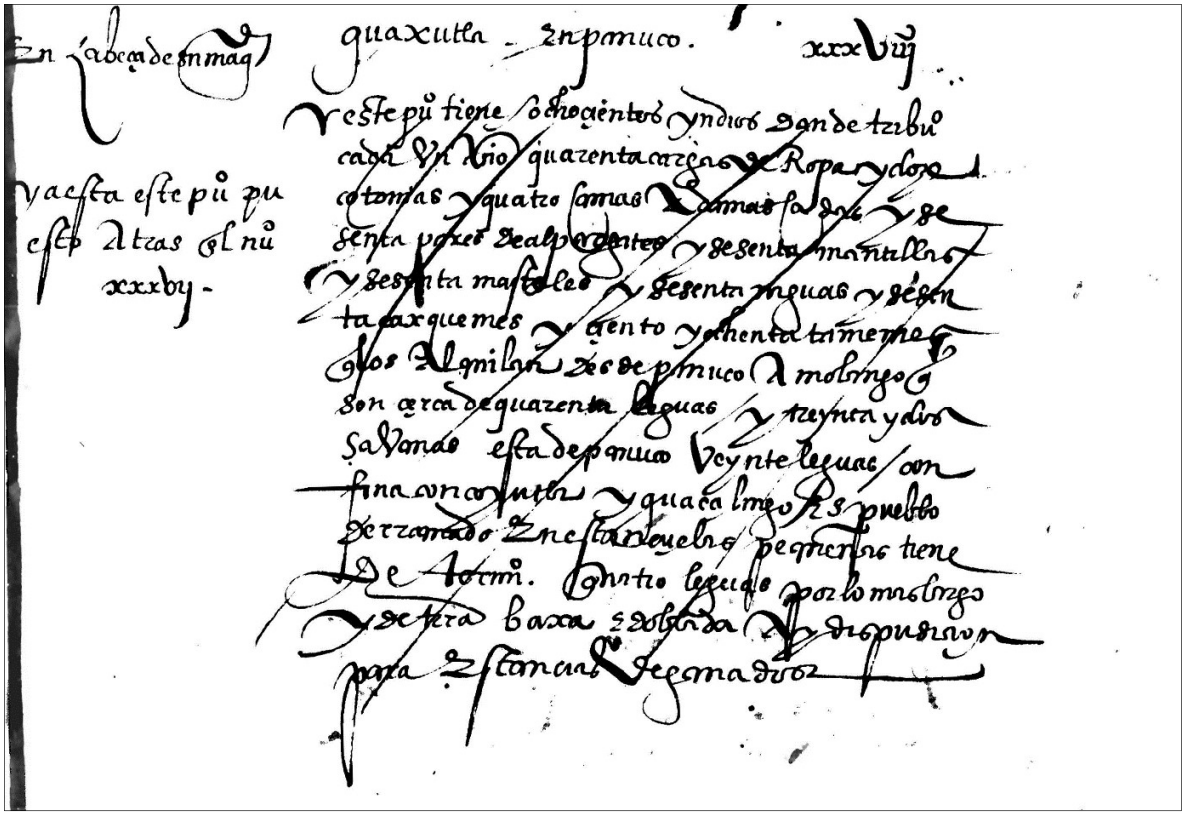


Figura 1.5. Ejemplos de documento con palabras rayadas, párrafos rayados, distancias diferentes entre líneas, notas de página (en el borde izquierdo del documento) y líneas que se intersectan verticalmente.

Antes de localizar las líneas de texto es necesario preprocesar la imagen con el objetivo de eliminar las variaciones relacionadas a la calidad del documento y digitalización, como el tipo de material, ruido, resolución, inclinación, etc. (U. V. Marti & H. Bunke, 2001). Todas las variaciones mencionadas incrementan la complejidad de la SLT por lo que se han propuesto métodos que trabajan en lenguajes y estilos de escritura específicos (Peng, Yu, Li, & He, 2016), (Kesiman, Burie, & Ogier, 2016a).

Existen colecciones estándar para la SLT en documentos manuscritos, como el presentado en (B. Gatos, K. Ntirogiannis, & I. Pratikakis, 2009), sin embargo, esta colección no se encuentra disponible públicamente. En el trabajo de (Saabni, Asi, & El-Sana, 2014) se presenta una nueva colección de documentos manuscritos para la SLT. La colección de (Saabni et al., 2014) contiene una muestra de los documentos de la colección ICDAR 2009. Además, contiene documentos en diferentes lenguajes como el español, inglés, árabe, chino y documentos con dos lenguajes por página (árabe-español). Por esta razón es necesario el desarrollo de métodos para la SLT que puedan trabajar con documentos con lenguajes mezclados.

Existen muchos métodos para la LLT: métodos basados en la extracción del perfil de proyección horizontal (PPH) (Arvanitopoulos & Süsstrunk, 2014; Ptak, Zygadlo, & Unold, 2017), basados en agrupamiento (Mauricio et al., n.d.; Medina Morán, 2011) , basados en aprendizaje (Khandelwal et al., 2009; Steinherz et al., 1999) y métodos estocásticos (U. V. Marti & H. Bunke, 2001). Estos métodos no concluyen la tarea de la BRSL por lo que para separar dos líneas de texto se calcula el punto medio de cada línea a partir de los puntos previamente encontrados.

Los métodos para la LLT basados en PPH determinan el número de líneas localizando los picos en el histograma, véase Figura 1.6. Los trabajos basados en PPH están más enfocados en encontrar la localización de cada línea de texto que en buscar la ruta que mejor divida las líneas de texto.

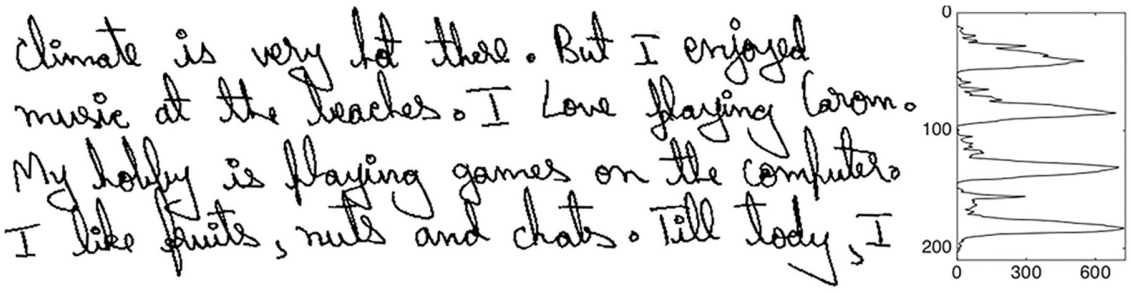


Figura 1.6. Extracción del perfil de proyección horizontal de un documento manuscrito.

En este tipo de trabajos se buscan los máximos locales (picos) en el PPH para después determinar un umbral promedio entre los picos y así, establecer un punto de corte. Uno de los problemas más importantes de estos métodos es que en el PPH existe más de un máximo local para una sola línea de texto, por lo que en algunos trabajos se resuelve este problema suavizando el PPH (Peng et al., 2016) o determinando un promedio en los máximos locales (Ptak et al., 2017).

Recientemente en (Arvanitopoulos & Süsstrunk, 2014) se propone obtener un mapa de energía de la imagen que contiene el documento para resaltar las diferencias entre los mínimos locales (valles) y los máximos locales (picos). Véase Figura 2.10.

Por otro lado, para el problema de la BRSLT algunos trabajos separan líneas de texto buscando una ruta de derecha a izquierda que permita separar las líneas de texto. La mayoría de estos métodos realizan una búsqueda local de la ruta para evitar pasar sobre los caracteres. La mayoría de estos trabajos hacen una optimización global de la ruta para separar las líneas de texto que se diferencian de los anteriores debido a que en los últimos es necesario una función que debe ser minimizada por el algoritmo (Arvanitopoulos & Süsstrunk, 2014; Du, Pan, & Bui, 2009; Koppula & Negi, 2014; Peng et al., 2016; Saabni et al., 2014).

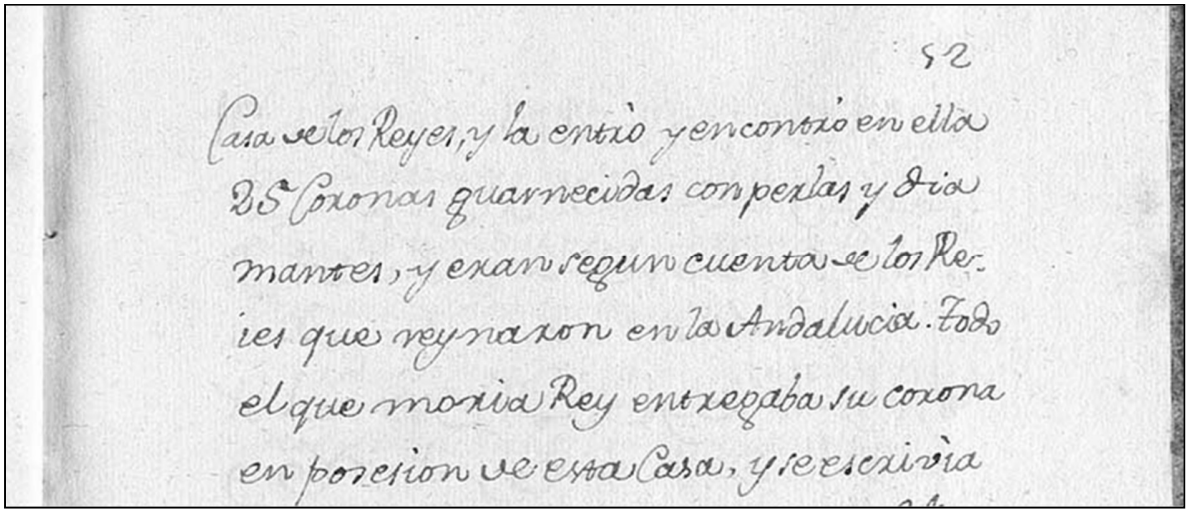


Figura 1.7. Ejemplo de documento usado por los trabajos del estado del arte para la etapa de segmentación. Visualmente se ve que es posible separar cada una de las líneas de texto manuscrito.

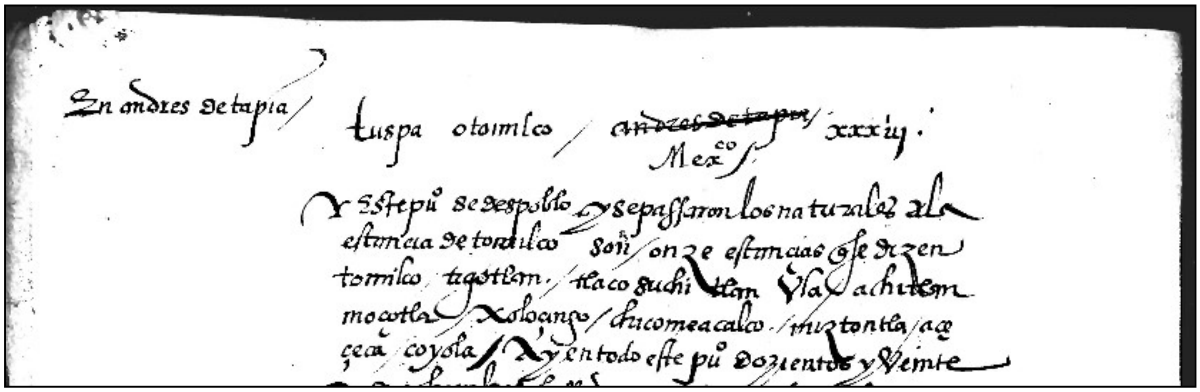


Figura 1.8. A diferencia del documento de la Figura 1.7 se puede ver que no es posible separar las líneas de texto manuscrito sin cortar algunos caracteres.

1.2 Planteamiento del problema

Después de analizar las características de los documentos manuscritos, los enfoques del estado del arte para la tarea de la (SLT) se sabe que para segmentar líneas de texto manuscrito es necesario generar un trazo que permita separar las líneas de texto tocando la menor cantidad de caracteres. Por lo que en este trabajo se resuelven dos problemas computacionales, el primero consiste en: ¿Cómo desarrollar un método computacional para la LLT en documentos

con líneas de texto que se intersectan verticalmente? Y a partir de las líneas localizadas es necesario saber ¿Cómo desarrollar un método automático e independiente del lenguaje que genere una ruta que permitan segmentar líneas de texto manuscrito que además sea un método no supervisado y no dependa de los materiales del documento, la época del documento y la caligrafía del autor?

1.3 Justificación o motivación

Es importante automatizar las etapas que permiten aplicar procesos de recuperación de información en documentos manuscritos para disminuir la cantidad de recursos invertidos en este proceso. Si se desarrolla una herramienta que facilite la segmentación de líneas en documentos manuscritos entonces los historiadores y paleógrafos pueden enfocarse en el estudio de la escritura y así, facilitar el proceso de interpretación de los documentos.

Peng y Villegas afirman que la tarea abordada en este trabajo aún no está concluida (Mauricio et al., n.d.; Peng et al., 2016). Además, concuerdan en que los métodos actuales no son robustos, por lo que es necesario continuar con la investigación para desarrollar métodos que puedan aplicarse a diferentes idiomas y estilos de escritura.

1.4 Objetivo General

Desarrollar un método automático e independiente del lenguaje para segmentar líneas de texto en documentos manuscritos.

1.4.1 Objetivos particulares

- Desarrollar un método independiente del lenguaje.
- Realizar un análisis de los métodos del estado del arte para identificar las categorías que se tienen hasta el momento
- Implementar los métodos del estado del arte que hasta el momento tienen mejores resultados para la LLT y BRSLT.
- Desarrollar un método para la LLT buscando los valles en un HPP.
- Mejorar los resultados del estado del arte.
- Desarrollar un método no supervisado.
- Probar el método propuesto con el corpus propuesto en (Saabni et al., 2014).
- Crear un corpus para el reconocimiento de escritura y SLT.
- Mejorar todo el proceso de reconocimiento de escritura manuscrita mejorando la etapa de segmentación de líneas de texto en documentos manuscritos.
- Desarrollar todo este trabajo mediante la metodología de desarrollo de software orientado a componentes.

1.5 Hipótesis

Cuando el humano segmenta líneas de texto manuscrito no necesita comprender el contenido de todo el documento, porque la separación entre líneas de texto se hace generando un camino que mejor separe el espacio entre líneas de texto manuscrito. Por lo que si se implementa un método de optimización para buscar la ruta óptima para llegar del lado izquierdo al lado derecho de la hoja tocando la menor cantidad de caracteres entonces será posible segmentar líneas de texto en documentos manuscritos antiguos.

En este trabajo se plantean dos hipótesis para la SLT, la primera hipótesis consiste en que es posible mejorar el rendimiento si el proceso de localización de líneas de texto se centra en los mínimos locales (espacios blancos entre las líneas) en un perfil de proyección horizontal de un mapa de energía alfa de un documento manuscrito; el cual reducirá el valor de los mínimos locales incluso cuando existan líneas que se tocan verticalmente o se solapan.

La segunda hipótesis consiste en que es posible mejorar la exactitud en la BRSLT si se calcula una ruta no lineal con una función que trate de cruzar por la menor cantidad de letras minimizada globalmente desde el punto inicial hasta el punto final del documento. De esta forma es posible mejorar el proceso de SLT.

1.6 Estructura de la tesis

En el capítulo uno, correspondiente a la introducción, se presentan las características que anteceden al problema abordado en este trabajo. También se presentan los conceptos fundamentales necesarios para poder comprender el problema abordado en este trabajo. Además, se realiza una descripción del proceso actual que se sigue para poder transcribir el texto de documento manuscritos antiguos. Se presenta la hipótesis que se comprobó en este trabajo. Además, en esta sección se describe el objetivo general y los objetivos particulares que se completaron con este trabajo.

En el capítulo dos, correspondiente al marco teórico, se presentan todos los conceptos que se necesitan para poder comprender el método propuesto.

El capítulo tres presenta el estado del arte actual para el problema abordado en este trabajo. En esta sección también se presenta una comparación de los métodos actuales dependiendo de su enfoque (métodos supervisados y métodos no supervisados). Además, se presenta una descripción de los corpus que se han usado actualmente para la tarea de segmentación de líneas de texto manuscrito

En el capítulo cuatro está descrito el método propuesto basado en la hipótesis planteada en el capítulo uno. En este capítulo se presentan los algoritmos de cada una de las etapas del

método propuesto en este trabajo. Además, se presenta el diagrama de componentes para describir el funcionamiento general del método propuesto.

En el capítulo cinco, correspondiente a la etapa de experimentación, se pone aprueba el método propuesto en este trabajo. Se realizan comparaciones con diferentes configuraciones en nuestro método. En este capítulo también se describe la configuración del algoritmo genético y la configuración de cada uno de los operadores usados. Para complementar lo anterior, también se hace una comparación del rendimiento del método propuesto usando un corpus del estado del arte. También se presenta una comparación de la complejidad que hay en la escritura entre dos corpus para la segmentación de líneas de texto manuscrito.

En el capítulo seis se presenta la validación de nuestra hipótesis tomando como referencia los resultados de la etapa de experimentación. También se presentan nuevos objetivos para trabajo futuro.

CAPÍTULO 2.

Marco Teórico

En este capítulo se presentan los conceptos necesarios para comprender el estado del arte y el método propuesto. Este capítulo comienza con los conceptos básicos de procesamiento de imágenes. Posteriormente son descritos los conceptos y métodos de preprocesamiento que se han usado para la tarea el *análisis de imágenes de documentos*. Al final, se concluye describiendo cada uno de los elementos de un algoritmo genético.

2.1 Imagen digital

Es una representación de un objeto o escena real o imaginaria. La representación visual de una escena se realiza mediante función bidimensional $f(x, y)$, donde (x) y (y) representan una coordenada en la función bidimensional, en cada coordenada se almacena un valor de brillo durante la captura de la imagen. El proceso de creación de una imagen digital finaliza cuando se almacena en el ordenador.

Una imagen digital puede generarse de dos formas: mediante un dispositivo digitalizador como un escáner o cámara digital o mediante un programa informático denominado editor de mapa de bits (Gonzalez, Woods, Davue Rodríguez, & Rosso, 1996).

2.2 Píxel

Es el elemento fundamental que compone a una *imagen digital*. Cada píxel tiene un valor de brillo (Burger & Burge, 2008; Gonzalez et al., 1996), el número total de píxeles puede calcularse multiplicando la anchura por altura de la imagen (x, y) donde (x) es el número de columnas y (y) el número de filas.

Con el proceso de digitalización se asigna un valor y una posición a cada píxel para formar una imagen digital. En la Figura 2.1 se muestra un ejemplo de la representación de los píxeles en una imagen digital.

A partir de la matriz es posible establecer un sistema de coordenadas que tiene por origen la esquina superior izquierda (Martinsanz, PAJARES, & de la Cruz García, 2007). Todas las coordenadas de una imagen tienen valores positivos, la posición en el eje (x) es de izquierda a derecha y la posición en el eje (y) es de arriba hacia abajo. En la Figura 2.1 se proporciona una imagen con el sistema de coordenadas en una imagen digital.

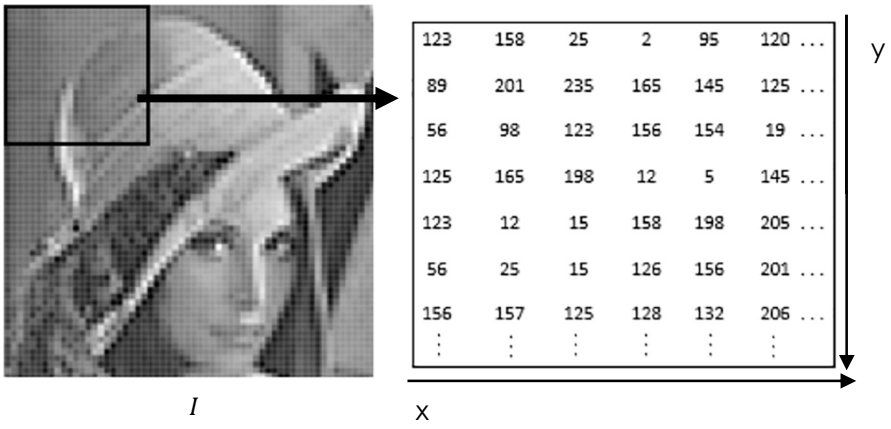


Figura 2.1. Representación de una imagen digital. Cada elemento de la matriz es un pixel.

2.3 Ruido

Las variaciones generadas durante el proceso de impresión, digitalización, variaciones debido a la antigüedad de un documento, manipulación de los documentos y captura de las imágenes generan variaciones aleatorias. Dentro del procesamiento digital de imágenes estas variaciones son llamadas ruido. El ruido está definido como un conjunto de alteraciones aleatorias que no corresponde con la realidad.

El ruido es generado por el dispositivo de captura de la imagen, el ruido aumenta la complejidad para el tratamiento de las imágenes digitales (Burger & Burge, 2008).

Cada dispositivo de captura añade distintos patrones de ruido, un ejemplo del ruido generado por un dispositivo de captura se muestra en la Figura 2.2.

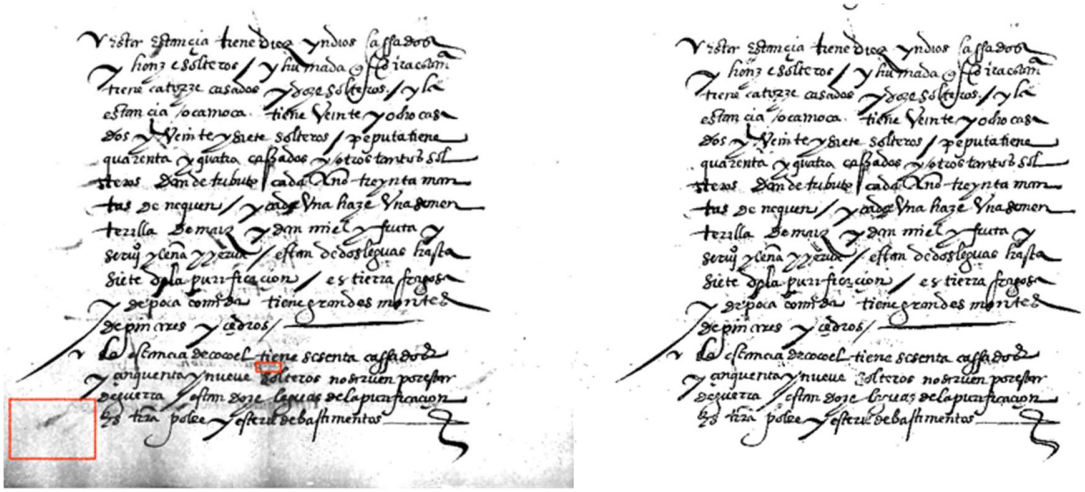


Figura 2.2. Ejemplo de ruido en una imagen digital. La imagen de la izquierda fue capturada con un escáner con la lampara desgastada, esto agregó y modificó las características originales del documento sin modificar su estructura. En la imagen de la derecha se muestra el mismo documento, digitalizado con un escáner en óptimas condiciones.

2.4 Procesamiento digital de imágenes

Son un conjunto de técnicas que permiten adquirir, representar, mejorar y manipular imágenes digitales. El objetivo de este conjunto de técnicas es procesar las imágenes digitales para disminuir la cantidad de información y facilitar el proceso de búsqueda de información (Gonzalez et al., 1996).

2.5 Imágenes en escala de grises

Una imagen digital en escala de grises es un tipo de imagen digital en el cuál los valores de los tres canales de color tienen el mismo valor. Este tipo de imágenes está compuesto sólo por diferentes valores de grises (Gonzalez et al., 1996). Este tipo de imágenes se diferencian de las imágenes binarias ya que en las imágenes binarias los píxeles sólo pueden tomar un valor, blanco o negro (0 o 1) (Burger & Burge, 2008).

Las imágenes en escala de grises tienen diferentes valores entre el color blanco y el negro en un rango [0-255] (Burger & Burge, 2008). En la Figura 2.3 se proporciona un ejemplo de comparación de una imagen en color y una imagen digital en escala de grises.

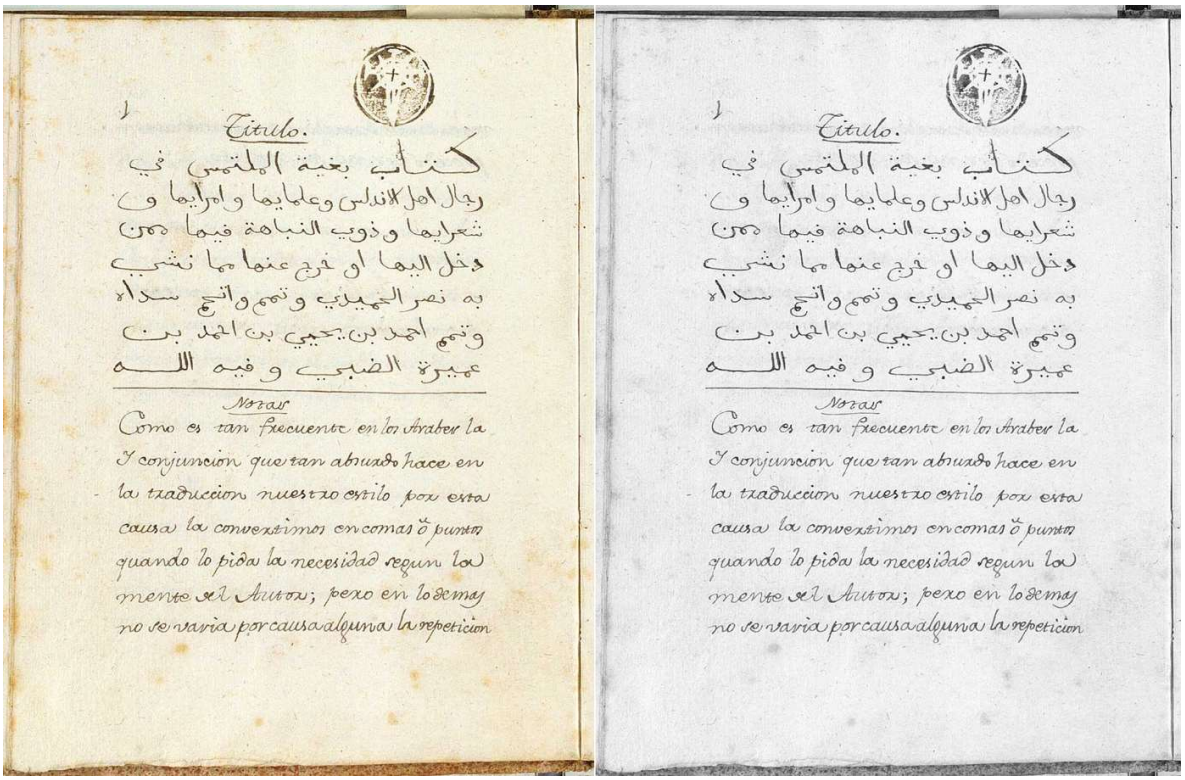
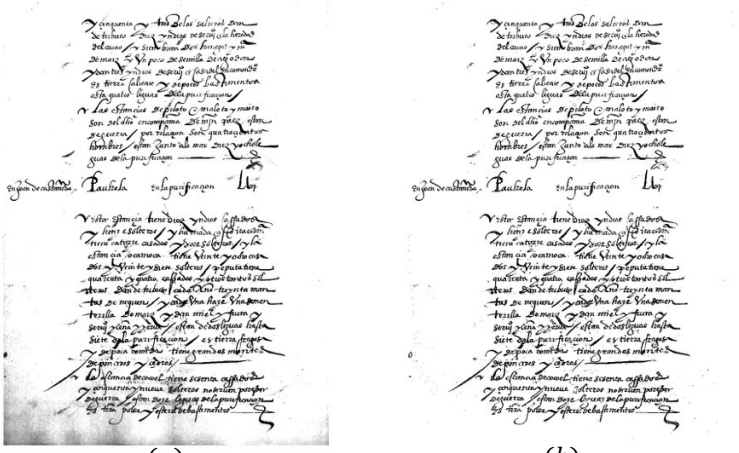


Figura 2.3. Comparación de color entre una imagen en escala de grises y una imagen en color. En la Imagen del lado izquierdo se muestra un documento que ha sido digitalizado conservando los valores de color del documento. En la imagen del lado derecho se muestra un documento que ha sido procesado para generar una imagen en escala de grises. La imagen del lado derecho contiene 3 veces menos información que el documento digitalizado en color.

2.6 Imágenes binarias

Las imágenes binarias son un tipo de imagen digital con solo 2 niveles de color normalmente blanco y negro. En la Figura 2.4 se muestra un ejemplo de transformación de una imagen (a) en escala de grises a una imagen (b) en escala binaria. El proceso de binarización permite separar objetos con respecto al fondo de la imagen (O'Gorman, Sammon, & Seul, 2008).

Con el proceso de binarización se reducen las variaciones de los píxeles a sólo dos valores, normalmente 0 y 1 (blanco y negro). Son codificados usando un solo bit por píxel para el almacenamiento del color. Normalmente, uno de los colores se emplea como fondo y el otro para los objetos que aparecen en la imagen. Las imágenes binarias son usadas para representar líneas, documentos digitalizados, etcétera (Burger & Burge, 2008). Los métodos de binarización pueden usarse para reducir la cantidad de en imágenes de documentos antiguos (Peng et al., 2016).



(a) (b)

Figura 2.4. Conversión de (a) una imagen en escala de grises a (b) una imagen binarizada para reducir la cantidad de información a procesar.

2.7 Operaciones geométricas

La geometría es una rama de las matemáticas y es útil en el procesamiento digital de imágenes para realizar modificaciones sobre una imagen, por ejemplo, ampliación, reducción, traslación, rotación (Gonzalez et al., 1996).

2.7.1 Traslación

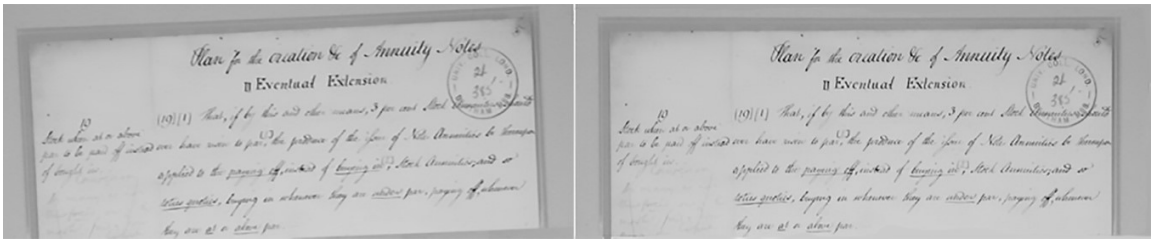
El operador de traslación realiza una transformación geométrica sobre una imagen. El proceso de traslación de imágenes digitales vuelve a dibujar una imagen (a) en una nueva posición sobre una imagen de salida (b). Es el proceso de cambiar la ubicación de una figura en un plano sin modificar su rotación o tamaño (Burger & Burge, 2008). En la Figura 2.5 se puede observar que la imagen (a) fue trasladada hacia la derecha y hacia arriba produciendo una imagen (b).

(a) (b)

Figura 2.5. Traslación de una imagen digital 100 pixeles sobre el eje x y 50 píxeles sobre el eje y.

2.7.2 Rotación

Es el proceso de girar una imagen digital alrededor de un centro previamente definido y con un ángulo establecido, donde cada punto sigue un círculo alrededor del centro (Burger & Burge, 2008). Un ejemplo de rotación de una imagen es mostrado en la Figura 2.6 donde se muestra una imagen (a) de un documento manuscrito que fue digitalizado con una inclinación, en la imagen (b) se muestra la imagen (a) con el ángulo de inclinación corregido.



a b

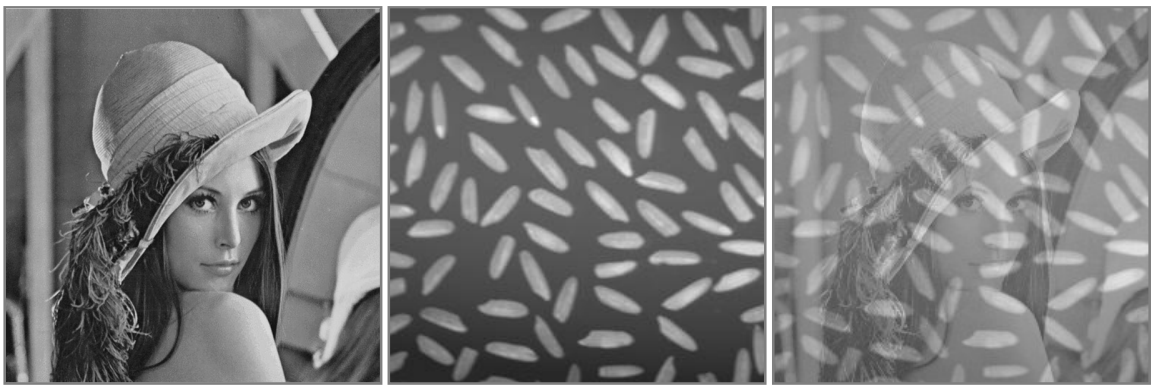
Figura 2.6. Rotación de una imagen digital. La imagen (a) muestra la imagen original y la imagen (b) muestra el resultado después de rotar la imagen (a) 3 grados.

2.8 Mezcla alfa

El operador de mezcla alfa es un método que permite mezclar dos imágenes con una transparencia (Burger & Burge, 2008), $\mathrm{I_{BG}}$ y $\mathrm{I_{FG}}$. La imagen original I es cubierta por una ventana de la imagen I, cuya transparencia es controlada por el valor $\propto$ en la forma:

$$\mathrm{Alpha}(I, w, \alpha)^r = I_{BG}(u + w) + (1-\propto) \cdot I_{FG}(u + w)$$

donde $0 \leq \propto \leq 1$, $\propto = .5$, u es la posición en el eje x y r es el número de veces que la mezcla alfa se aplica.



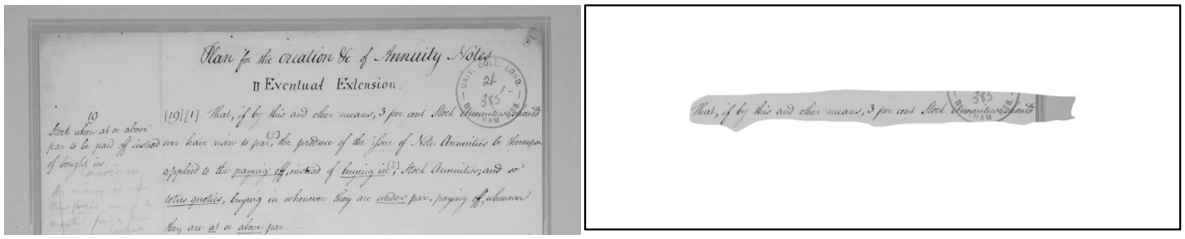
(a) (b) (c)

Figura 2.7. Ejemplo de aplicación de una marca de agua sobre una imagen digital, este operador del procesamiento digital de imágenes permite combinar dos o más imágenes.

2.9 Segmentación

Es el proceso mediante el cual se extraen objetos individuales desde una imagen digital. El proceso de segmentación puede llegar a ser extremadamente difícil y por regla general la etapa más delicada y difícil de todo sistema de procesamiento de imágenes (Gomez-Allende & Gómez-Allende, 1993; Gonzalez et al., 1996).

Es el proceso que consiste en dividir una imagen en diferentes partes. El objetivo de la segmentación es simplificar la representación de una imagen para facilitar su análisis. En la imagen (a) de la Figura 2.9 se proporciona un ejemplo de segmentación de imágenes. El proceso de segmentación tiene diferentes objetivos, en la imagen (b) de la Figura 2.8 se tiene proporciona un ejemplo de segmentación de una línea de texto manuscrito.



a *b*

Figura 2.8. Ejemplo de segmentación de imágenes. La imagen a es la imagen fuente y la imagen b muestra el resultado de la segmentación.

2.10 Transformada de Radon

Es una herramienta fundamental en muchas disciplinas incluyendo el mapeo de radares, imágenes geofísicas, testeo no destructivo e imágenes médicas (Helgason, 1999). En el trabajo presentado en (Helgason, 1999) se afirma que la transformada de Radon es el mapeo de una función $f(x,y)$ con $(x,y) \in R2$ a función $Rf(t,\theta)$ con $t \in R$ y $\theta \in (0,\pi)$ y está definida por la siguiente ecuación:

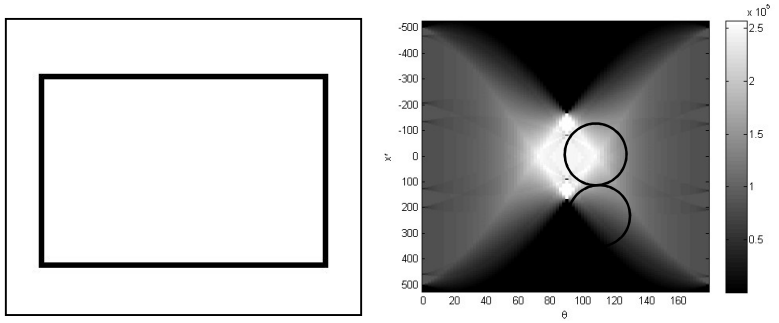
$$Rf(t,\theta) = f(x,y)\delta(t - xcos(\theta) + y\,sin(\theta))dxdy$$

Ecuación 2.2. Ecuación para aplicar la transformada de Radon a un conjunto de puntos $(\boldsymbol{x},\boldsymbol{y})$.

Donde δ denota la función Dirac δ. Rf es una integral de f sobre una línea Lt,θ definida por: $\mathrm{t} = \mathrm{xcos}(\theta) + \mathrm{ysin}(\theta)$.

Con lo anterior se dice que la transformada de Radon es una función que asigna un valor numérico a cada elemento de un grupo de líneas para permitir identificar el nivel de inclinación de las imágenes.

En la Figura 2.9.a se muestra un rectángulo con un ángulo de inclinación igual a 0. En la Figura 2.9.b se muestra la trasformada de Radon del rectángulo donde las intersecciones entre líneas permiten identificar el ángulo de cada imagen.



(a) (b)

Figura 2.9. En la imagen mostrada en el lado izquierdo (a) se muestra un rectángulo con una inclinación de 0 grados. En el lado derecho (b) de esta Figura se muestra la transformada de Radon del rectángulo. Al analizar el resultado de la transformada de Radon se puede apreciar que solo hay dos intersecciones entre las líneas (puntos negros) y están ubicadas al centro de la figura, esto indica que la imagen no tiene inclinación.

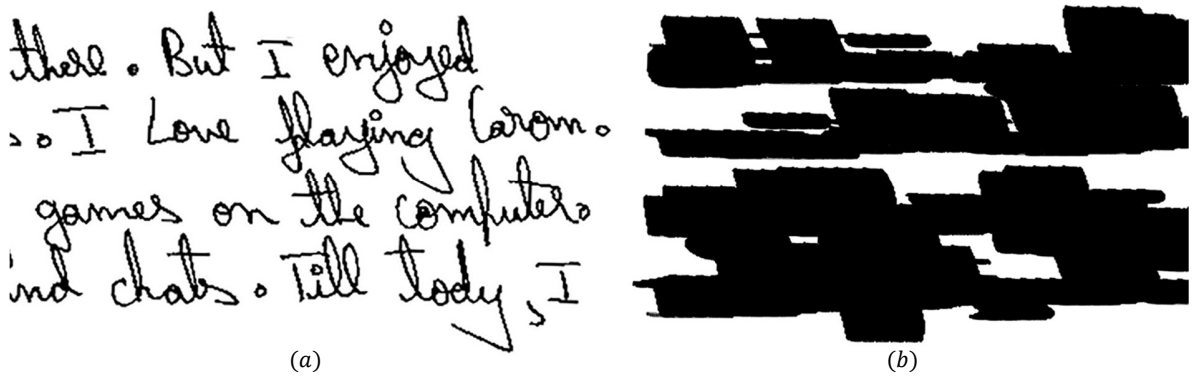
2.11 Mapa de energía

Un mapa de energía es un método de caracterización de imágenes que permite analizar la distribución de la información almacenada en una imagen digital que contiene texto (Du et al., 2009).

Los mapas de energía se utilizan como una etapa de preprocesamiento de documentos manuscritos (Kesiman et al., 2016a; Koppula & Negi, 2014; Nicolaou & Gatos, 2009).

Se han propuesto diversos métodos para extraer un mapa de energía de un documento manuscrito, el más utilizado consiste en analizar la imagen digital de manera horizontal.

El objetivo de la extracción de un mapa de energía es eliminar los espacios horizontales vacíos entre caracteres (Kesiman et al., 2016a). En la Figura 2.10 se muestra un ejemplo de la extracción del mapa de energía de una imagen digital que contiene texto manuscrito.

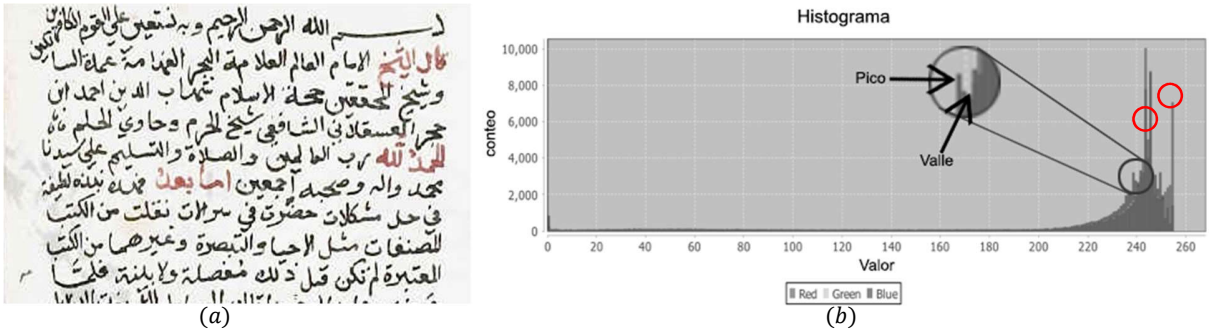


(a) (b)

Figura 2.10. Ejemplo de extracción del mapa de energía de una imagen que contiene texto.

2.12 Histograma

Un histograma de una imagen digital es una representación de la frecuencia de los valores de intensidad presentes en una imagen (Gonzalez et al., 1996). En la Figura 2.11 se muestra el histograma de una imagen digital. Los elementos por los que está compuesto un histograma son los picos y valles. Un pico es un valor que está por encima de sus vecinos. Un valle es un valor que está por debajo de sus vecinos. En la imagen (b) se muestra encerrado en un círculo un pico y un valle del histograma.



(a) (b)

Figura 2.11. Histograma de color de una imagen. En la imagen (a) se muestra un ejemplo de documento manuscrito en idioma Árabe. En la imagen (b) se muestra el histograma del documento de ejemplo.

2.13 Histograma de proyección horizontal

Es una presentación unidimensional de una imagen bidimensional. Los valores mostrados en el histograma de proyección vertical representan la densidad de la distribución de la información contenida en la imagen digital (O'Gorman et al., 2008; Ptak et al., 2017).

Los histogramas de proyección vertical son usados para calcular la inclinación de una imagen (Bagdanov & Kanai, 1998), reducción de ruido en imágenes de documentos digitales (Prachanucroa & Phongsuphap, 2013), identificación de autores en documentos manuscritos

(Biswas & Das, 2012), entre otros. Para una imagen de un documento con (x) filas y (y) columnas el histograma de proyección vertical se puede extraer con la Ecuación 2.3 (Likforman-Sulem, Zahour, & Taconet, 2006).

$$VPP(p) = \sum_{1 \le x \le m} f(x, y)$$

Ecuación 2.3 Ecuación que permite calcular el histograma de proyección vertical en un documento de 'm' filas y 'n' columnas.

En la Figura 2.12 se muestra el histograma vertical de una imagen digital de un documento manuscrito. Este histograma es usado en el estado del arte para localizar la posición de cada línea buscando los picos del histograma.

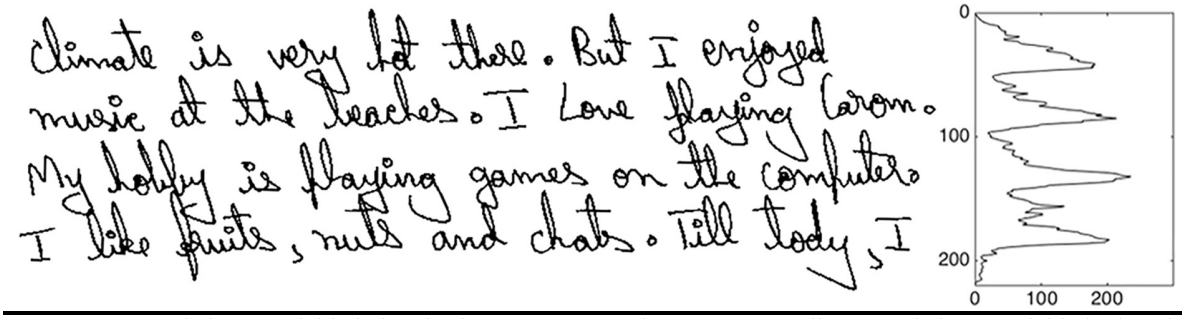


Figura 2.12. En la imagen del lado izquierdo se muestra un documento a analizar y en la imagen del lado derecho se proporciona el histograma de proyección horizontal del documento.

2.14 Algoritmos genéticos

Un algoritmo genético es un algoritmo basados en los principios de la teoría de la evolución de Charles Darwin. Un algoritmo genético está compuesto por: un operador de selección, un operador de cruza, un operador de mutación, una o más funciones de aptitud y la población de inicial.

Para crear la población inicial se genera un conjunto de individuos con valores aleatorios, cada individuo es evaluado por medio de la función de aptitud. En un algoritmo genético se aplica el proceso de selección natural para mejorar las soluciones mejor adaptadas.

Los individuos mejor adaptados tendrán una mayor probabilidad para ser padres de nuevos individuos. El proceso evolutivo que ocurre dentro de un algoritmo genético permite encontrar la mejor solución a un problema dado entre un conjunto de soluciones posibles (García-Hernández & Ledeneva, 2013).

El objetivo de cualquier algoritmo genético es la búsqueda de un conjunto de parámetros para aproximar una o más funciones.

Los algoritmos genéticos trabajan con valores codificados. Así, después de identificar la función que se desea aproximar, el siguiente paso es definir una codificación para representar las variables que se usan para aproximar la función. La codificación de las variables es llamada genotipo (Yarushkina, 2002).

Para crear el genotipo es necesario definir una cadena de tamaño finito en un alfabeto finito. La forma más común de crear un genotipo es utilizando cadenas binarias, pero también se pueden codificar valores reales (Yarushkina, 2002). En el genotipo se codifica la estructura de los datos que se aplicarán a la función que se desea maximizar o minimizar.

2.15 Resumen del capítulo

A lo largo de este capítulo se han descrito los elementos básicos que conforman una imagen digital. Además, se han descrito los elementos que intervienen en la representación de las imágenes digitales. Se hace mención de las variaciones que modifican el contenido de las imágenes.

Al final de este capítulo se ha expuesto el concepto de algoritmo genético. Además, se encuentran descritos todos los elementos necesarios para la implementación de un algoritmo genético para optimizar el proceso de búsqueda de un conjunto de parámetros para optimizar una función.

CAPÍTULO 3.

Estado del Arte

En este capítulo se presenta un análisis de los métodos del estado del arte para la segmentación de líneas en documentos manuscritos.

3.1 Preprocesamiento

En el estado del arte se hace énfasis en la etapa de binarización (B. Gatos et al., 2009; I. Pratikakis, K. Zagoris, G. Barlas, & B. Gatos, 2016; Mauricio et al., n.d.), corrección de la inclinación (Arica & Yarman-Vural, 2001; Bagdanov & Kanai, 1998; Mauricio et al., n.d.) y la reducción de ruido (A. Prachanucroa & S. Phongsuphap, 2013) son procesos fundamentales en todas las tareas del análisis de imágenes de documentos y especialmente en la tarea de la SLT.

Los métodos propuestos en (Kesiman, Burie, & Ogier, 2016b; Likforman-Sulem et al., 2006; U. V. Marti & H. Bunke, 2001; Z. Shi & Venu Govindaraju, 2004) tienen una etapa de preprocesamiento antes de realizar la etapa de la LLT o la BRSLT. Por otro lado, en los trabajos que se proponen en (Koppula & Negi, 2014; Nicolaou & Gatos, 2009; Peng et al., 2016; Q. N. Vo & G. Lee, 2016) se asume que la entrada de los métodos son imágenes de documentos

previamente binarizadas y que además se les ha corregido el ángulo de inclinación. En los trabajos en donde no se realiza una etapa de preprocesamiento se afirma que sus resultados pueden mejorar cuando se aplique una etapa de preprocesamiento (Saabni et al., 2014).

3.2 Métodos de abajo hacia arriba para la LLT

Los métodos que se encuentran en esta categoría agrupan elementos básicos de la imagen como píxeles, caracteres o componentes conectados (Koppula & Negi, 2014; Valy, Verleysen, & Sok, 2016) para formar patrones de líneas (Saabni et al., 2014). Estos métodos tienen buen rendimiento en documentos que contienen grupos de líneas de texto con diferentes longitudes e inclinación para cada párrafo, En la Figura 4 se muestra un ejemplo de documento en los cuales estos métodos han mostrado un mejor rendimiento.

El problema de los métodos de esta categoría es que no pueden segmentar líneas de texto en documentos con líneas de texto que se intersectan verticalmente como el mostrado en la Figura 1.5.

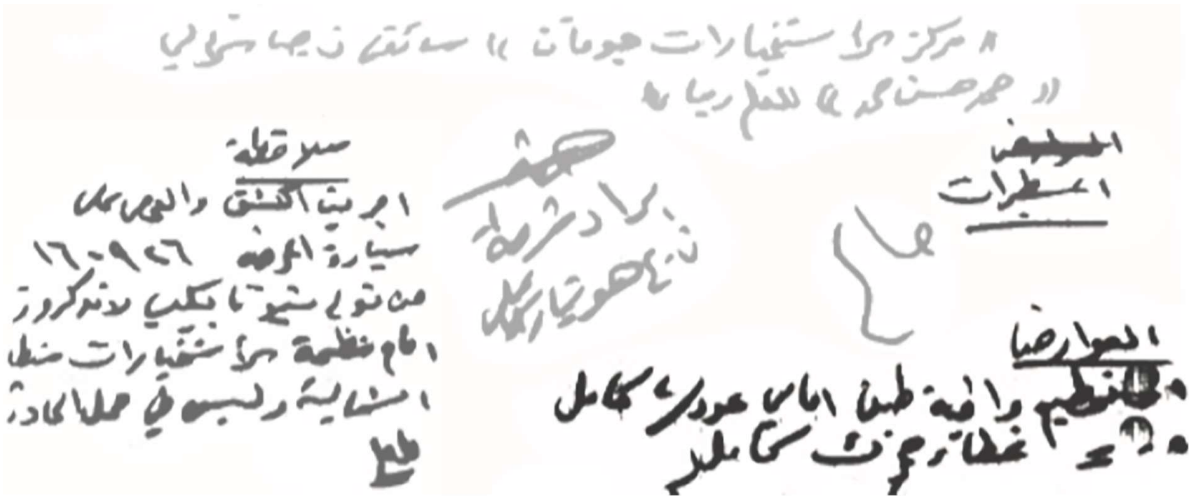


Figura 3.1. Ejemplo de documentos en donde los métodos de arriba hacia abajo muestran un mejor rendimiento (Du et al., 2009). En esta figura se muestra cada grupo de texto con un tono diferente.

3.3 Métodos de arriba hacia abajo para la LLT

Los métodos de esta categoría están basados en aprendizaje, extracción de PPH y extracción de mapa de energía.

3.3.1 Métodos basados en aprendizaje

Los métodos basados en aprendizaje necesitan una muestra de entrenamiento para la LLT (Q. N. Vo & G. Lee, 2016) y una muestra de entrenamiento con trazos para generar un modelo que permita aprender a realizar el proceso del SLT (Q. N. Vo & G. Lee, 2016). Estos métodos tienen la desventaja de ser dependientes del lenguaje.

3.3.2 Métodos basados en el perfil de proyección horizontal

Estos métodos son los más comunes para la LLT en imágenes de documentos con caracteres impresos (Ha, Haralick, & Phillips, 1995). Algunos de estos métodos no pueden ser aplicados directamente a textos manuscritos porque necesitan una separación clara entre cada una de las líneas de texto vecinas. La inclinación de los documentos y caracteres que se tocan afectan el rendimiento de estos métodos.

Los métodos de esta sub-categoría están enfocados en localizar los picos en un PPH con el objetivo de identificar la separación entre las líneas de texto. Sin embargo, cuando esta técnica se aplica directamente a documentos con líneas de texto que se intersectan verticalmente (Figura 3.2) no es posible encontrar los picos porque todos tienen diferente anchura y longitudes. Por lo que en los métodos basados en este enfoque se propone ajustar un conjunto de umbrales que son definidos empíricamente para cada colección de documentos (Arvanitopoulos & Süsstrunk, 2014; Kesiman et al., 2016b; Ptak et al., 2017). Además, el problema principal de los métodos de esta categoría es que están basados en localizar los picos en el PPH. Sin embargo, el documento de la Figura1.6 contiene 4 líneas de texto, por lo que buscando solo los picos en el PPH se puede estimar que existen cinco líneas de texto manuscrito.

Los trabajos presentados en (Arvanitopoulos & Süsstrunk, 2014; Kesiman et al., 2016b; Ptak et al., 2017) presentan un método basado en la extracción del PPH para estimar la posición de cada línea de texto manuscrito localizando los valores de los máximos locales. Sin embargo, estos métodos presentan problemas para estimar la separación entre líneas de texto vecinas (valores de los mínimos locales) cuando las líneas de texto se solapan o se tocan. En el trabajo de (Peng et al., 2016) solo se realiza la etapa de la LLT. Para aplicar un método basado en el PPH es necesario que el texto se represente sobre líneas horizontales, por lo tanto, no es posible aplicarlo directamente a documentos como el que se presenta en la Figura 3.1.

3.3.3 Métodos basados en la extracción de un mapa de energía

Un mapa de energía es el proceso mediante el cual un documento es procesado con el objetivo de eliminar los espacios en blanco entre caracteres y palabras (Du et al., 2009; Liwicki,

Indermuhle, & Bunke, 2007; Saabni et al., 2014); provocando una gran diferencia entre los máximos y mínimos locales en el PPH.

Los trabajos que se presentan en (Du et al., 2009; Kesiman et al., 2016b; Liwicki et al., 2007; Nicolaou & Gatos, 2009; Saabni et al., 2014) reducen los espacios en blanco entre cada carácter y cada palabra aplicando un método de extracción de mapa de energía basado en un operador gradiente (ME-gradiente) o una función específica (ME-F) (Du et al., 2009), etc. Para hacer esto, la imagen del documento es suavizada o trasladada sobre la imagen original para obtener un mapa de energía como el que se muestra en la Figura 3.2.

Después de aplicar este proceso en algunos métodos se propone agrupar las regiones con más información (Du et al., 2009). En otros trabajos se propone extraer el PPH del mapa de energía (Arvanitopoulos & Süsstrunk, 2014; Ptak et al., 2017). Sin embargo, al aplicar los métodos actuales para la extracción de mapa de energía en documentos con líneas de texto que se intersectan verticalmente no es posible diferenciar las líneas buscando los máximos locales. (ver Figura 3.3)

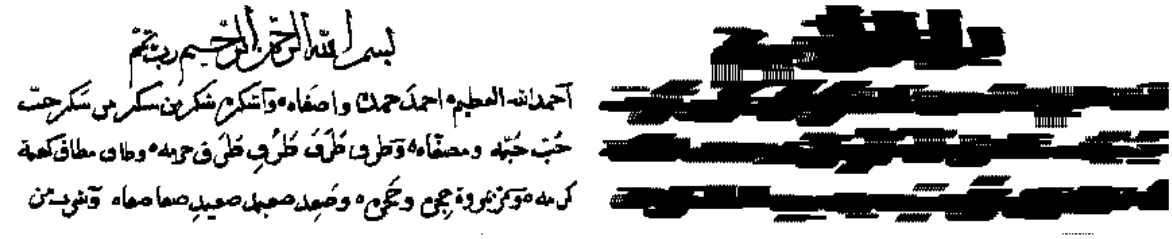


Figura 3.2. Mapa de energía generado usando el operador morfológico dilatación para escritura sin líneas conectadas método propuesto en (Kesiman et al., 2016b). En esta imagen es posible observar que los espacios vacíos entre cada letra se cubren, esto facilita la identificación de las líneas candidatas.

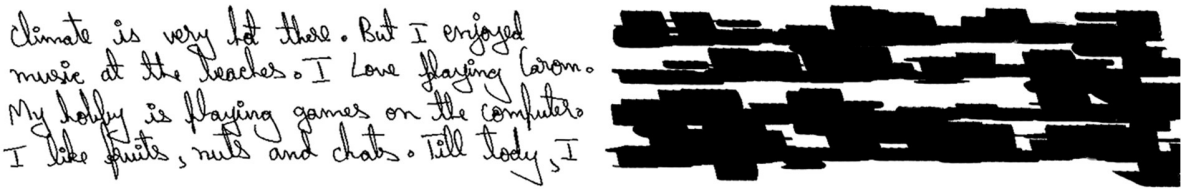


Figure 3.3. Ejemplo de mapa de energía para eliminar los espacios en blanco entre caracteres y palabras usando el método propuesto por (Du et al., 2009).

En la Figura 3.3 se puede observar que han desaparecido los espacios en blanco de cada carácter y palabra, pero además han desaparecido los espacios en blanco entre líneas de texto vecinas. Es importante preservar los espacios en blanco entre las líneas de texto vecinas para facilitar la búsqueda de ruta que permita segmentar las líneas de texto manuscrito. Por lo tanto, estos métodos tienen problemas separando documentos en donde líneas de texto vecinas se intersectan verticalmente.

3.4 Métodos para la BRSLT

Algunos trabajos para buscar la ruta con la mayor cantidad de píxeles de color blanco realizan una búsqueda local, pero esto no garantiza que se encuentre la ruta óptima (Koppula & Negi, 2014; Nicolaou & Gatos, 2009; Peng et al., 2016). En (Kesiman et al., 2016b) se realiza una búsqueda local de la mejor ruta tomando el valor de una función de costo que considera la menor cantidad de píxeles negros dentro de la ruta.

En el método que se propone en (Arvanitopoulos & Süsstrunk, 2014) se usa una adaptación del método Seam Carving (Avidan & Shamir, 2007) para encontrar la ruta óptima. Seam Carving que ha sido usado para cambiar el tamaño de una imagen. El objetivo de esta modificación es encontrar el Seam Carving o ruta que mejor separe dos líneas de texto manuscrito. Eventualmente, la ruta con el menor error o costo es la ruta deseada. Para evitar que este método se desvíe en un mínimo local es necesario usar un método de optimización global, esta técnica es discutida en (Liwicki et al., 2007; Saabni et al., 2014).

3.5 Corpus

En el estado del arte se han presentado corpus de documentos escritos en tablillas de palma en (Kesiman et al., 2016a; Peng et al., 2016; Valy et al., 2016). En algunos corpus se tienen valores repetidos en todos los documentos, en algunos se tiene el mismo número de líneas (Peng et al., 2016; Valy et al., 2016). Algunos corpus sólo están compuestos por documentos del mismo idioma (Peng et al., 2016; Valy et al., 2016). Debido a todas las combinaciones posibles de escritura e idioma hasta el momento no se realizado una comparación entre los métodos del estado del arte.

El corpus usado en (Ptak et al., 2017) se compone por 1,514 líneas de texto agrupadas en 60 páginas manuscritas en idioma polaco. La anchura promedio de cada línea es de 6.4 píxeles. En este corpus no se tienen líneas de texto con caracteres conectados verticalmente. En la Figura 3.5 se muestran dos ejemplos de documentos que pertenecen a ese corpus.

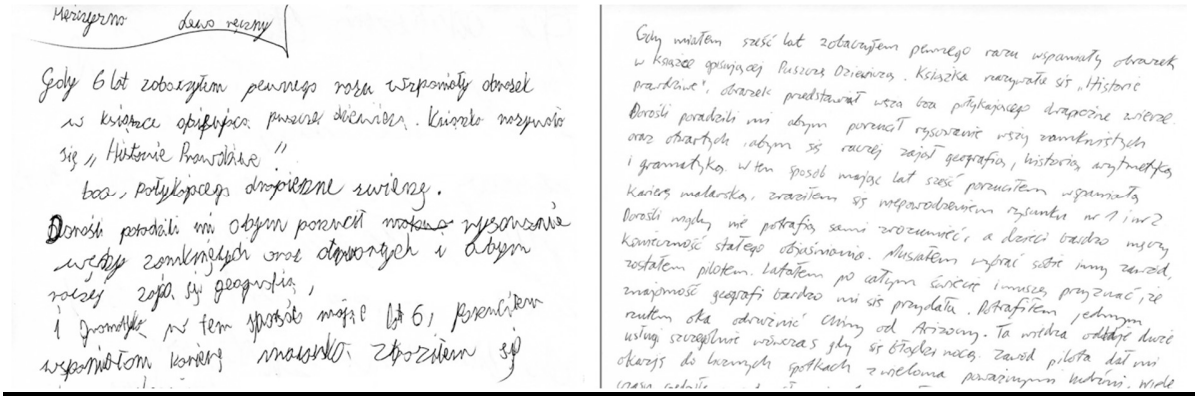


Figura 3.4. Ejemplo de documentos del corpus del corpus utilizado en (Ptak et al., 2017). En ninguna línea de los documentos se tienen caracteres que intersectan otras líneas verticalmente.

En (Du et al., 2009) se hace uso de un corpus compuesto por documentos escritos en chino, hindú y coreano con un total de 296 páginas. Este trabajo está enfocado en encontrar párrafos de texto manuscrito.

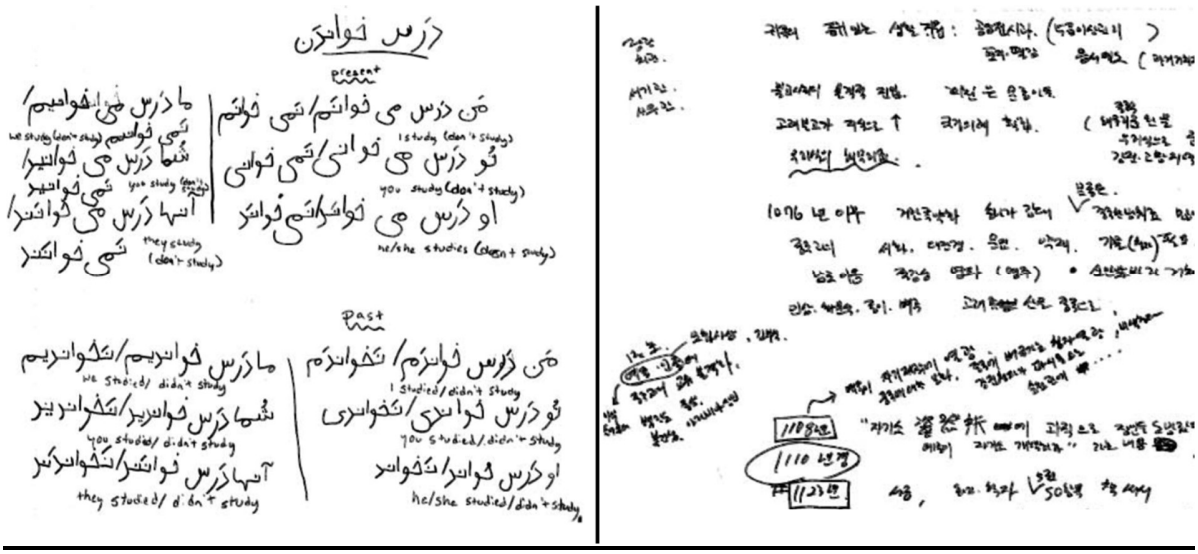


Figura 3.5. Ejemplo de documentos del corpus utilizado en (Du et al., 2009). El método diseñado para este corpus no considera documentos con líneas de texto que se intersectan verticalmente.

En el trabajo presentado en (Valy et al., 2016) se presenta un corpus compuesto por 224 líneas de texto agrupadas en 50 páginas de texto manuscrito antiguo en idioma Khmer. Este corpus no contiene líneas de texto conectadas verticalmente.

En la Figura 3.7 se muestra un ejemplo de documentos para los que se ha diseñado el método para segmentar líneas de texto manuscrito presentado en (Valy et al., 2016).

Figura 3.6. Ejemplo de documentos del corpus utilizado en (Valy et al., 2016). El método diseñado para este corpus no considera documentos con líneas de texto que se intersectan verticalmente.

En (Peng et al., 2016) se presenta un corpus en idioma Dai que contiene 1,050 líneas de texto manuscrito distribuidas en 290 documentos. Este corpus no contiene líneas de texto conectadas verticalmente. Todos los documentos del corpus contienen el mismo número de líneas. En la Figura 3.8 se proporcionan documentos del corpus usado para desarrollar el método propuesto en (Peng et al., 2016).

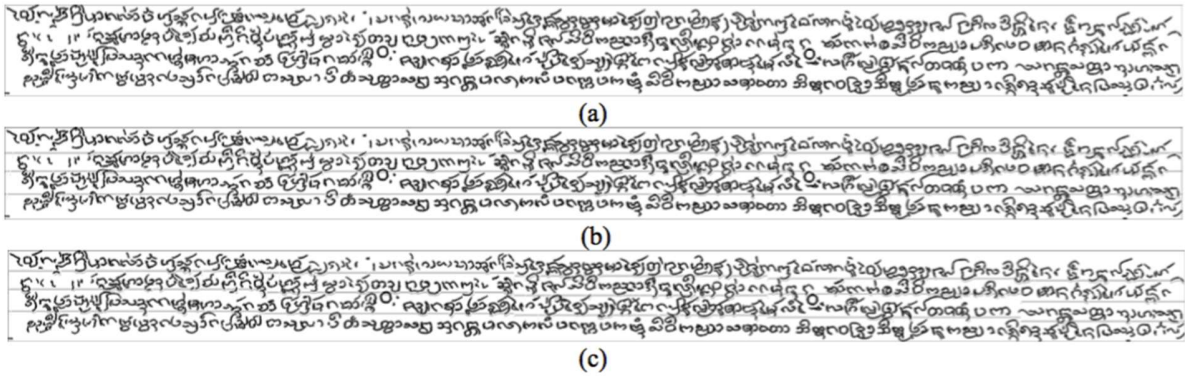


Figura 3.7. Ejemplo de documentos del corpus utilizado en (Peng et al., 2016). El método diseñado para este corpus no considera documentos con líneas de texto que se intersectan verticalmente.

El método presentado en (Kesiman et al., 2016a) se ha probado en un corpus compuesto por 140 líneas de texto manuscrito agrupados en 35 páginas. En la Figura 3.9 se proporciona un ejemplo de documento del corpus usado en este trabajo.

Figura 3.8. Ejemplo de documentos del corpus utilizado en (Peng et al., 2016). El método diseñado para este corpus no considera documentos con líneas de texto que se intersectan verticalmente.

En (Koppula & Negi, 2014) se menciona que se usa el corpus de la competencia "Handwritten text line segmentation" del congreso ICDAR 2009. En la Figura 3.7 se muestran ejemplo de documentos usado para probar este método.

En el trabajo presentado en (Nicolaou & Gatos, 2009) también se usaron dos corpus para probar el método propuesto, el primer está compuesto por el 20% del corpus de la competencia "Handwritten text line segmentation" del congreso ICDAR 2009 y un 70% del corpus que se presenta en (Saabni et al., 2014). En la figura 3.10 se muestran ejemplos de documentos del corpus usado en este trabajo.

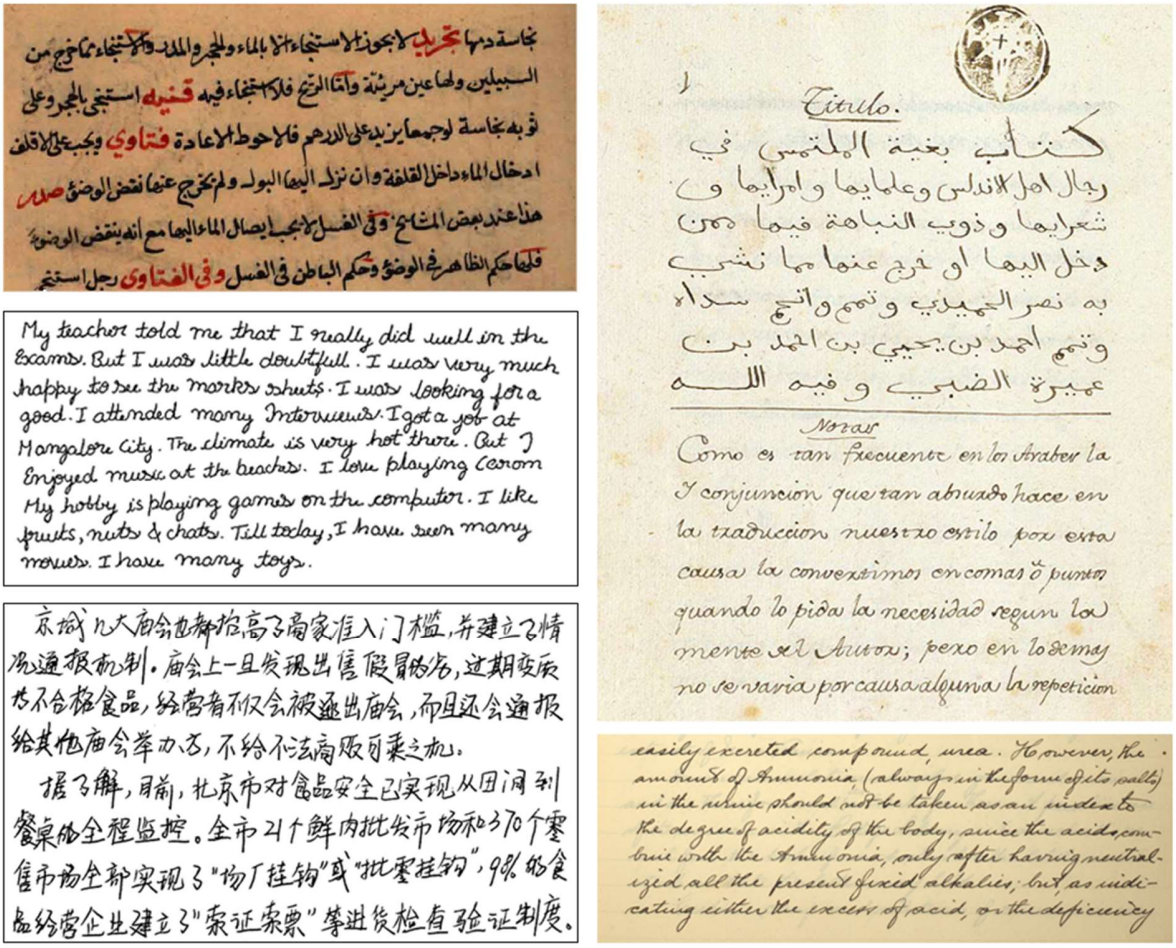


Figura 3.9. Ejemplo de documentos que pertenecen al corpus usado en los trabajos presentados en (Arvanitopoulos & Süsstrunk, 2014; Saabni et al., 2014). Los documentos de este corpus se han usado para evaluar el método propuesto en este trabajo.

Los corpus disponibles para evaluar el rendimiento de los métodos para la segmentación de líneas de texto en documento manuscritos fueron creados usando documentos en los que siempre es posible encontrar una ruta que sólo pase por píxeles en color blanco.

Hasta el momento no se tiene una evaluación de los trabajos actuales usando un mismo corpus. En cada uno de los trabajos del estado del arte se presentan corpus de documentos manuscritos en diferentes idiomas.

De los ocho trabajos analizados durante esta investigación sólo está disponible el corpus propuesto en el trabajo de (Saabni et al., 2014) ya que todos los trabajos analizados han sido

probados con colecciones de documentos que son privados y no están disponibles para realizar investigación.

Se ha establecido contacto con los autores de los trabajos en el caso de las colecciones privadas, pero no fue posible que compartieran el corpus debido a restricciones de derechos de autor.

Resumen del capítulo

A lo largo de este capítulo se han presentado una clasificación de los métodos para la segmentación de líneas de texto en documento manuscritos.

Al analizar cada uno de los trabajos del estado del arte se concluye que estimar las posiciones de cada línea de texto con un histograma de proyección vertical sólo es un paso de todo el proceso para la segmentación automática de líneas de texto manuscrito.

Todos los trabajos usan una función de costo que permite optimizar los valores que permiten dibujar una separación entre cada línea de texto manuscrito.

CAPÍTULO 4.

Método Propuesto

En este capítulo se presenta el método propuesto en esta tesis para la segmentación de líneas de texto en documentos manuscritos independiente del lenguaje mediante un algoritmo genético. Retomando los problemas planteados en este trabajo que consisten en: ¿Cómo desarrollar un método computacional para la LLT en documentos con líneas de texto que se intersectan verticalmente? Y ¿Cómo desarrollar un método automático e independiente del lenguaje que genere una ruta que permitan segmentar líneas de texto manuscrito que además sea un método no supervisado y no dependa de los materiales del documento, la época del documento y la caligrafía del autor?

La hipótesis de este trabajo para el primer problema consiste en que es posible mejorar el rendimiento si la búsqueda de las líneas se centra en los mínimos locales (espacios blancos entre las líneas) en un perfil de proyección horizontal de un mapa de energía alfa de un documento manuscrito; el cual reducirá el valor de los mínimos locales incluso cuando existan líneas que se tocan verticalmente o se solapan.

La segunda hipótesis consiste en que es posible mejorar la exactitud en la BRSLT si se calcular una ruta no lineal con una función que trate de cruzar por la menor cantidad de letras

minimizada globalmente desde el punto inicial hasta el punto final del documento. De esta forma, la tarea de la SLT puede mejorarse.

En la primera sección se presenta una descripción general de la metodología de solución. En las siguientes subsecciones se describen detalladamente el método propuesto para generar el mapa de energía basado en el operador *alpha blending*. Posteriormente se describe cómo es que los parámetros son calculados automáticamente para el método del mapa de energía propuesto. Al final de este capítulo se describe el método propuesto para la BRSLT.

4.1 Metodología general para la SLT

La entrada de nuestro método es una imagen con el ángulo de inclinación corregido. Los pasos de nuestro método propuesto son los siguientes:
1.Determinar automáticamente el valor de los parámetros para la ME-Alfa con una submuestra aleatoria de la colección completa (Ver sección 4.3)
2. Para cada imagen en la colección, la ME-Alfa con los parámetros del paso 1 (Paso 3.2) es generado y, después, se extrae el perfil de proyección horizontal (HPP) de la imagen del ME-Alfa.
3. Remover los mínimos locales en el HPP que son menores a un umbral (en porcentaje) basado en el promedio de los máximos locales encontrados.
4. Localizar los mínimos locales en el PPH para determinar los puntos iniciales y finales para trazar la ruta.
5. Buscar una ruta entre el punto inicial y final que mejor divida líneas de texto vecinas.

4.2 Mapa de energía propuesto

A diferencia de los trabajos del estado del arte en este trabajo se propone un nuevo mapa de energía basado en el operador de mezcla alfa (Burger & Burge, 2008). El objetivo de generar un mapa de energía con el operador de mezcla alfa es generar un perfil de proyección horizontal en el que todos los mínimos locales bajen al valor cero. El resultado de esta etapa es la localización de las líneas de texto encontradas.
Para generar un mapa de energía alfa se proponen dos elementos: una etapa de binarización y el operador de mezcla alfa.
Cuando la imagen de entrada para este método es una imagen binaria, el resultado del método ME-Alfa es una imagen en escala de grises.
En la imagen que contiene la ME-Alfa las regiones con mayor energía (regiones más oscuras) corresponden con el centro de las líneas de texto y las regiones con menos energía

corresponden con los bordes superiores e inferiores de cada línea de texto. Al Figura 4.1 muestra un ejemplo de regiones con mayor energía y menor energía.

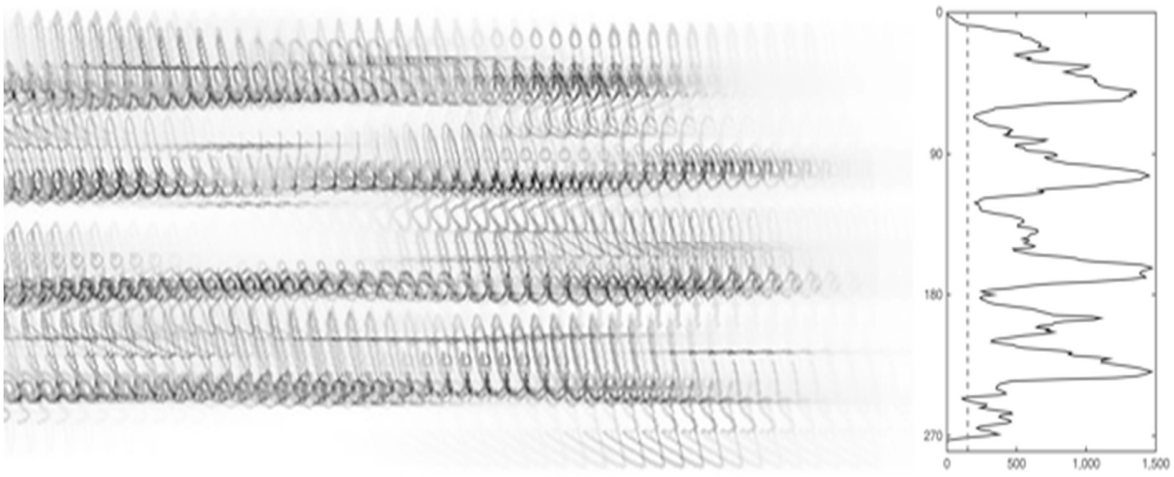


Figura 4.1. Ejemplo del ME-Alfa propuesto y el HPP.

La binarización del mapa de energía alfa (ME-Alfa) permite remover los píxeles con menos energía en comparación de la extracción directa del perfil de proyección horizontal (ver Figura 4.2). A partir de la extracción del PPH podemos mejorar la etapa de la LLT como puede verse en la Figura 4.2.

Figura 4.2. Ejemplo de PPH del ME-Alfa binarizado.

Por medio de este método, es posible generar un PPH donde la distancia entre los picos y valles sea mayor (véase Fig. 4.3) a diferencia del PPH mostrado en la Figura 4.1, dejando un salto muy grande entre los mínimos locales con el valor cero.

Para determinar los puntos de origen para trazar la ruta necesaria para localizar los valles con una longitud mayor a un pixel en el PPH. La Figura 4.3 muestra el inicio y el fin de cada valle en el PPH. En la Figura 4.4 el valle inicial (P0) y el valle final (PF) encontrado por medio del ME-Alfa está dibujado sobre cada línea de texto correspondiente.

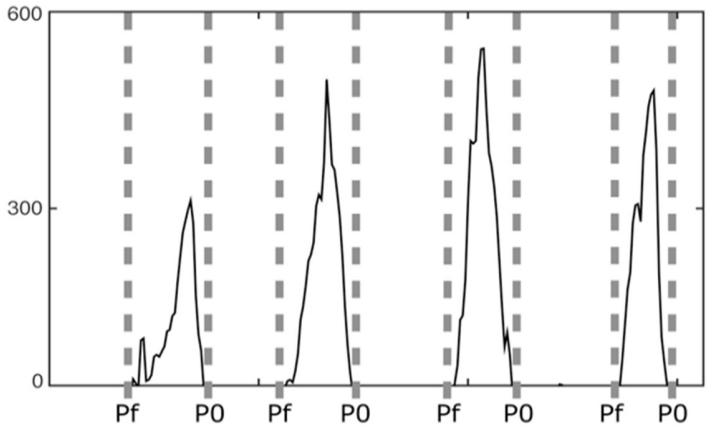


Figura 4.3. Perfil de proyección horizontal del ME-Alfa del documento E18 de la colección presentada en (Saabni et al., 2014).

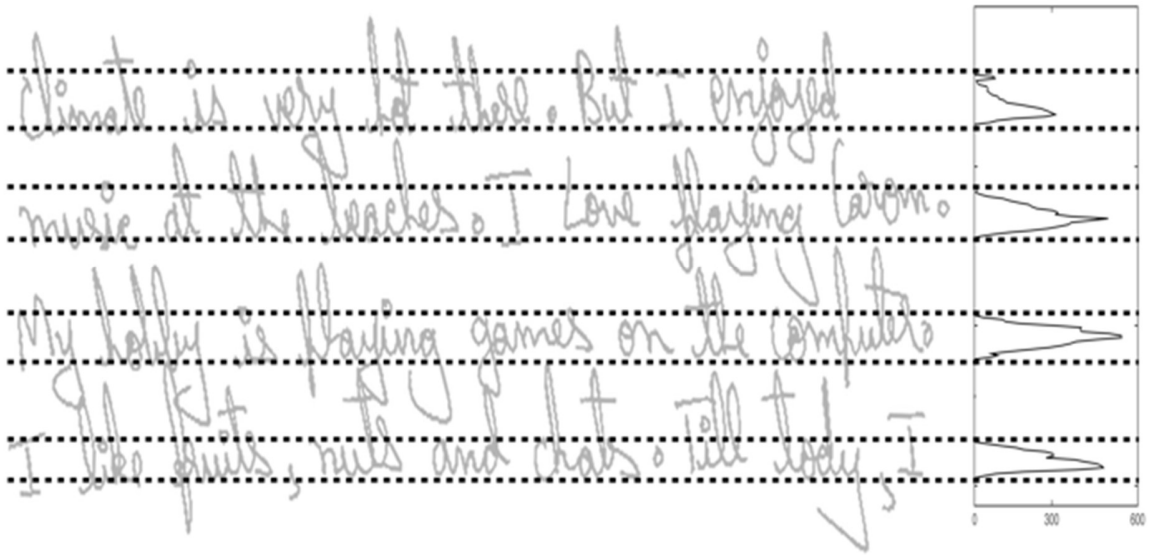


Figura 4.4. Líneas de texto encontradas mediante la búsqueda de los valles en el PPH mostrado en la Figura 4.3. Imagen del documento E18 de la colección presentada en (Saabni et al., 2014).

4.3 Buscando los mejores parámetros para el PEM-Alfa

Para aplicar el operador alfa es necesario definir un número de repeticiones (r) para computar los medial seams y un tamaño de ventana (w) para trasladar la imagen.
El problema con estos parámetros es que dependen del autor, lenguaje y estilo de la escritura. Por ejemplo, una ventana de tamaño muy pequeño no completará el objetivo de llenar los espacios en blanco entre caracteres. Por otro lado, un tamaño de venta muy grande no conservará la información que se necesita.
Dado un mínimo y máximo rango para la ventana $[w_{min}, w_{max}]$; y un mínimo y máximo valor para el rango de la variable *slice* $[r_{min}, r_{max}]$ en una imagen binaria I donde los bits 1 corresponden a píxeles negros y los bits 0 al fondo, el tamaño apropiado para r y w es aquel que produzca la mayor cantidad de bits 1 aplicando el operador de mezcla alfa con $\propto = 0.5$ a todas las combinaciones de r y w con $w_{min} \leq w \leq w_{max}$ and $r_{min} \leq r \leq r_{max}$. Representado de la siguiente forma:

$$Find(I, r, w) = \max\left(\sum_{\substack{w_{min} \leq w \leq w_{max} \\ r_{min} \leq r \leq r_{max}}} Bits1\big(Bin(Alpha(I, w, \propto)^r)\big) \right)$$

4.4 Búsqueda de una Ruta para Segmentar Líneas de Texto basado en un Algoritmo Genético (BRSLT-AG)

En esta etapa proponemos realizar una búsqueda global para un conjunto intermedio de puntos entre los puntos iniciales y finales (encontrados en la etapa anterior) para generar una ruta óptima que permita segmentar un par de líneas vecinas con texto manuscrito. Particularmente proponemos usar un Algoritmo Genético (AG) porque ha probado ser una de las mejores técnicas para resolver problemas de optimización.

En la primera etapa del algoritmo genético se genera un conjunto de soluciones aleatorias (etapa de generación de la población) que son evaluadas con una métrica que cuantifique la calidad de las soluciones (Etapa de cálculo de función de aptitud) con el objetivo de maximizar o minimizar la aptitud de cada individuo.

La solución al problema no es absoluta, por el contrario, es posible que existe un conjunto de soluciones donde algunas de ellas sean mejores que otras. En la siguiente etapa, es la selección y extracción de las mejores soluciones (etapa de selección de los padres), por lo que el AG propone una nueva población mezclando (etapa de cruza) algunos fragmentos de los genes que contienen las mejores soluciones con el objetivo de generar mejores soluciones (principio de evolución).

Después de varias iteraciones mezclando los genes de los mejores individuos, comienzan a aparecer soluciones repetidas. Para resolver este problema el AG aplica una pequeña variación (etapa de mutación) a los genes de cada individuo de la nueva población para explorar nuevas soluciones.

Al final de la etapa de mutación la nueva población es evaluada y el proceso se repite hasta que se encuentre una solución satisfactoria o hasta que se cumpla un criterio arbitrario que detenga la búsqueda (condición de paro).

4.4.1 Etapa de preprocesamiento

Antes de codificar a los individuos es necesario conocer le posición de cada punto inicial y final de la ruta.

4.4.2 Codificación de los cromosomas

El AG necesita codificar cada solución ($cromosoma$) usando una representación canonica. Para la SLT se propone representar los genes del cromosoma (C) con un vector de tamaño *n* (número de puntos intermedios entre el punto inicial y final de la ruta) con valores en base-k. La base de la representación de los genes ($base - k$) se determina como $k = |P_f - P_0| + 1$.

4.4.3 Población inicial

Después de conocer la codificación de los individuos, la primera generación es creada aleatoriamente, donde cada gen puede tomar un valor entre –k y k. ($C_{i=1\ldots n} = Random[-k, k]$).

4.4.4 Función de aptitud

El elemento más importante de un algoritmo genético es la función de aptitud, Para este problema se plantea como una función de minimización en donde es necesario buscar una ruta que cruce por la menor cantidad de píxeles oscuros en una imagen en escala de grises, pero que a su vez también sea la ruta más corta para que se acerque lo mayor posible a la línea inicial entre el punto inicial y el punto final como el humano normalmente lo hace. Por lo tanto, si existe una división clara entre dos líneas de texto entonces será posible trazar una línea recta desde el punto de inicio hasta el punto final. Por lo tanto, la función de aptitud consiste en agregar todos los píxeles <120 encontrados en la ruta generada más, la suma de los valores absolutos de los genes C_i.

$$FA(C_n) = C_1 + \sum_{i=2}^{n} \left(Bits < 120\left(P_{C_{i-1}}, P_{C_i}\right)\right) + C_i$$

4.4.5 Selección de padres

En esta etapa cada cromosoma es asociado con su valor de aptitud para permitir la selección de los mejores cromosomas. El principio de evolución establece que es natural mejorar el valor de aptitud de los individuos cuando dos buenas soluciones son cruzadas. En esta etapa el AG aplica el operador de selección por torneo.

4.4.6 Etapa de cruza

Un nuevo cromosoma es creado a partir de un par de padres seleccionados por el operador clásico de cruza uniforma en n −puntos.

4.4.7 Etapa de mutación

De acuerdo a la teoría de la evolución, la mutación ocurre con una probabilidad muy baja (cerca de 0.1%), sin embargo, es importante que ocurra para garantizar la evolución de la población. Para esta etapa, se propone mutar aleatoriamente (con una muy baja probabilidad por gen) los genes de acuerdo a la base$-k$ del cromosoma codificado. Para este paso, se propone usar mutación aleatoria.

$$Mutación(C_i) = Random[-k, k].$$

Resumen del capítulo

En este capítulo se han presentado dos nuevos métodos para la tarea SLT. El primer método permite localizar las líneas de texto en un documento manuscrito a partir de los valles de un HPP. El segundo método propuesto realiza una búsqueda global de la ruta que permita segmentar líneas de texto manuscrito con un algoritmo genético.

CAPÍTULO 5.

Experimentación y resultados

A lo largo de este capítulo se describe la metodología y configuración utilizada para validar la hipótesis con el método que se propone en este trabajo. También se describen los corpus utilizados. Además, se muestra el resultado obtenido con cada banco de pruebas.

En este trabajo se han planteado dos hipótesis para el problema de la SLT, la primera hipótesis consiste en que es posible mejorar el rendimiento si la búsqueda de las líneas se centra en los mínimos locales (espacios blancos entre las líneas) en un perfil de proyección horizontal de un mapa de energía alfa de un documento manuscrito; el cual reducirá el valor de los mínimos locales incluso cuando haya líneas que se tocan verticalmente o se solapan.

La segunda hipótesis consiste en que es posible mejorar la exactitud en la BRSLT si se calcula una ruta no lineal con una función que trate de cruzar por la menor cantidad de letras minimizada globalmente desde el punto inicial hasta el punto final del documento. De esta forma la tarea de la SLT puede mejorarse.

La sección de evaluación está dividida en tres secciones, en la primera sección se describen las colecciones de documentos utilizadas, la segunda sección contiene la evaluación del

rendimiento de tres métodos para buscar el número de líneas de cada documento. La tercera sección contiene la evaluación del rendimiento de los métodos que generan una ruta para segmentar líneas de texto manuscrito.

5.1 Colecciones de documentos

Después de realizar una búsqueda exhaustiva encontramos muy pocos documentos públicos para la segmentación de líneas de texto manuscrito, la mayoría de los trabajos utilizan corpus privados para realizar la etapa de experimentación (Kesiman et al., 2016b; Koppula & Negi, 2014; Ptak et al., 2017).

Otro problema encontrado es que algunos trabajos relacionados sólo utilizan un subconjunto de documentos de un corpus público (Saabni et al., 2014; Tseng & Lee, 1999). Dado este problema, en (Saabni et al., 2014) se presenta una colección de imágenes de documentos de cuatro lenguajes (árabe, chino, inglés y español) y una combinación de árabe-español.

5.1.1 Descripción del primer corpus

Para evaluar el método propuesto en este trabajo se han usado dos corpus, el primer corpus se ha usado previamente en algunos trabajos del estado del arte (Saabni et al., 2014),(Arvanitopoulos & Süsstrunk, 2014). Este corpus está compuesto de 1,645 líneas de texto y 215 páginas en cuatro idiomas diferentes: español, chino, inglés y árabe. El segundo corpus está compuesto por un total de 444 páginas con un total de 14,948 líneas de texto manuscrito en español antiguo (1548-1550). Además, contiene 10,583 palabras diferentes.

5.1.2 Colección de español antiguo

Para validar nuestra hipostesis en documentos con todas las variaciones de la escritura manuscrita creamos un corpus para el reconocimiento de texto y la SLT. Este interés surge debido a que existe una gran riqueza en dichos documentos, actualmente se están digitalizando los documentos antiguos y se ponen a disposición para su consulta. Hasta el momento no se cuenta con métodos robusto para recuperar la información contenida en estos medios. Un ejemplo de ello es el archivo general de indias ("Archivo General de Indias," n.d.).

El proceso de transcribir y recuperar información en documentos manuscritos es una tarea vigente y una prueba de ello es la creación de la tarea Handwritten Recognition en el ImageCLEF (Mauricio et al., n.d.) del año 2016 y los concursos que se han realizado en los últimos años (B. Gatos et al., 2009; J. A. Sánchez, A. H. Toselli, V. Romero, & E. Vidal, 2015). Sin embargo, el problema con los corpus anteriores es que algunos no están disponibles,

contienen escritura moderna o están en inglés, lo cual, dificulta el desarrollo de método para documentos en español.

De acuerdo con Villegas (Mauricio et al., n.d.) el interés por digitalizar libros y documentos manuscritos se ha incrementado a nivel mundial. La mayoría de los métodos actuales no son lo suficientemente robustos e independientes del lenguaje para realizar procesos de recuperación de información en documentos manuscritos. El objetivo principal de tener documentos manuscritos digitalizados es la posibilidad de hacer públicos documentos históricos y permitir la búsqueda de información y así prevenir el deterioro de los documentos durante su consulta.

Para el desarrollo del corpus se tomó como documento fuente el libro "Suma de visitas de pueblos de la Nueva España". Esta obra es muy importante ya que es una colección inédita de documentos acerca de la historia de México, esta obra fue escrita por más de un autor. En la obra se muestra la información recolectada sobre los recursos y acontecimientos de los pueblos que visitaban en ese entonces.

Es un documento que merece el calificativo de "fuente" para la historia de los pueblos nativos de México (García Castro, 2013) y fue publicado por la Universidad Autónoma del Estado de México. Esta obra comprende el período de 1548-1550 y consta de 255 fojas (hojas) microfilmadas. La transcripción de dicho documento se encuentra en el documento (García Castro, 2013) el cual ha sido transcrito por 4 diferentes paleógrafos en un período de 3 años.

El libro usado para la creación de este corpus está disponible como un conjunto de imágenes en internet, pero se necesitó de una mayor resolución de las imágenes digitales. Contamos con una copia microfilmada de estos documentos por lo cual pudimos digitalizarlos para tener una mayor cantidad de información.

El proceso de transcripción de las 444 páginas ya había sido realizado por 4 diferentes paleógrafos. En comparación con el corpus presentado en (Saabni et al., 2014) el corpus que se generó es más complejo, ya que en el corpus presentado en (Saabni et al., 2014) se puede identificar claramente la separación de cada palabra y línea (ver Figura 5.1). En la obra usada para la creación del corpus en español no se tiene una separación clara de cada una de las palabras y líneas. En la figura 5.1 se presenta una muestra de documento del corpus presentado en (Saabni et al., 2014) y en la figura 5.2 se presenta una muestra de documento del corpus generado durate esta investigación. Al comparar los documentos se puede ver que tienen una complejidad visual diferente. En el anexo 1 de este documento se presentan las muestras de documentos del corpus generado durante esta investigación.

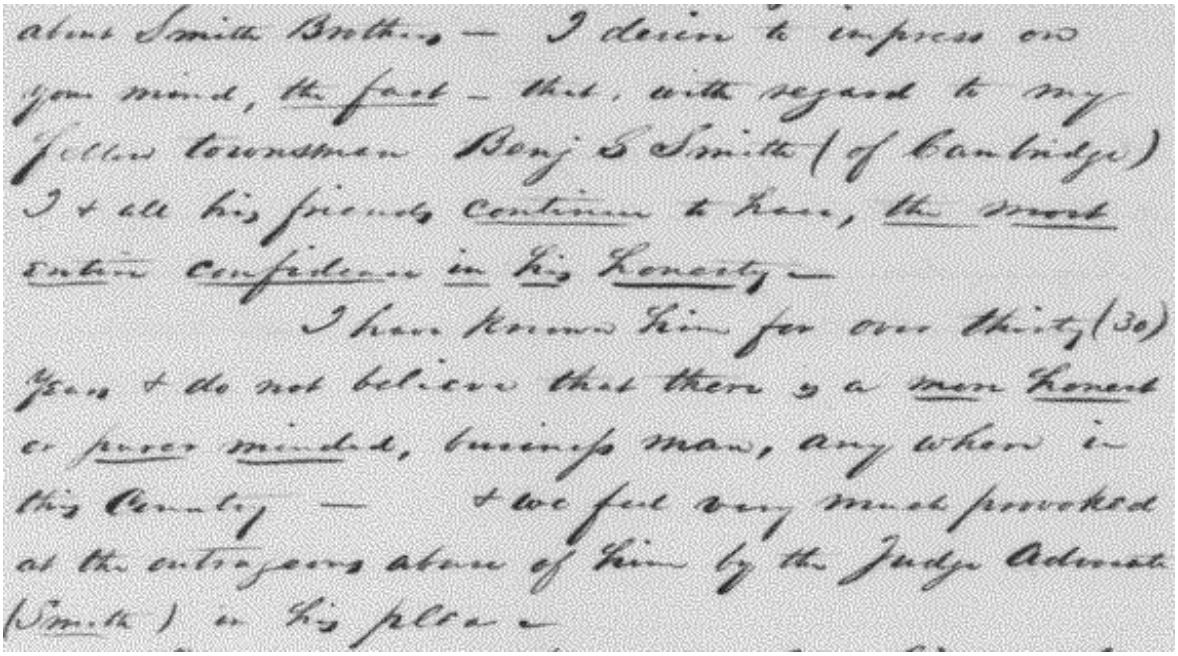


Figura 5.1. Muestra del corpus presentado en (Nicolaou & Gatos, 2009). El texto de este documento está en inglés.

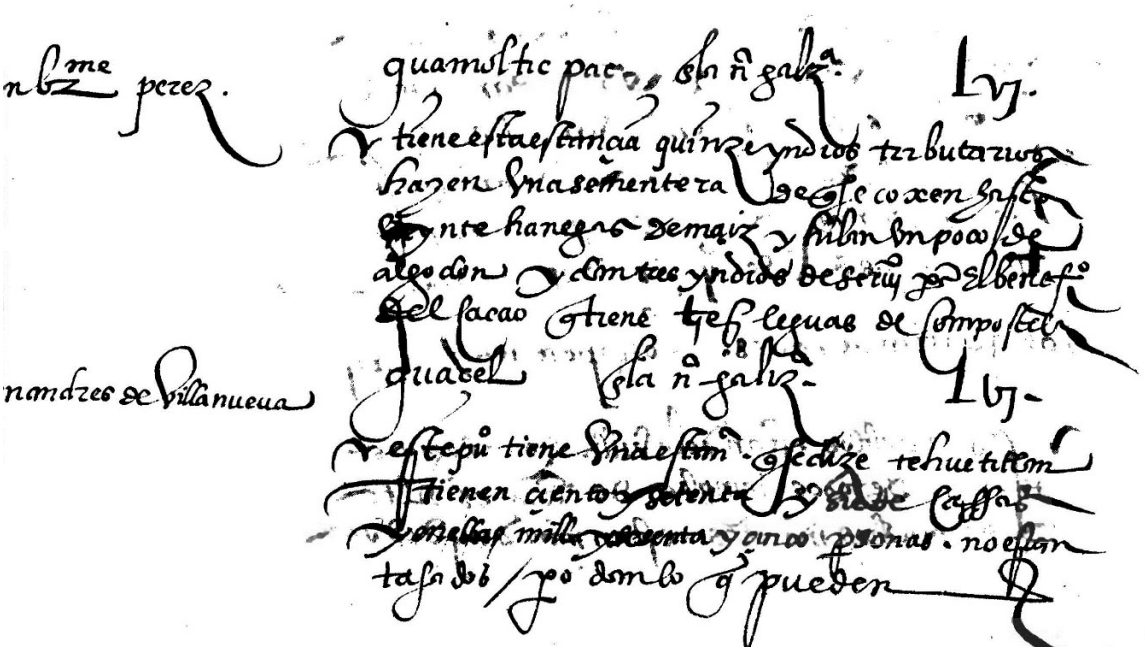


Figura 5.2. Muestra de un documento del corpus en español. Foja 70, página 155 en español mexicano del año 1538.

La Figura 5.2 presenta texto manuscrito que tiene una separación diferente entre las palabras, el texto fue escrito por escribanos entre los años 1548-1550. Hasta el momento no se cuenta con un corpus en español mexicano para esta tarea, los corpus disponibles están en inglés, hindú, árabe y chino (Du et al., 2009; Koppula & Negi, 2014; Mauricio et al., n.d.; Nicolaou & Gatos, 2009; Peng et al., 2016; Ptak et al., 2017).

Para el proceso de etiquetado de líneas de texto y palabras fueron necesarias 518 horas hombre. Además de esto se tuvieron que desarrollar las herramientas para realizar el proceso de etiquetado, ya que no se encontraron herramientas disponibles para este proceso. El desarrollo de las herramientas para el etiquetado y extracción de palabras y líneas llevó 120 horas hombre.

El recurso resultante está compuesto por 444 imágenes que contienen texto manuscrito, cada página fue etiquetada tomando como referencia la transcripción de la obra "Suma de visitas de los pueblos de México" (García Castro, 2013). El proceso de etiquetado de las imágenes fue realizado por tres personas.

Además, se han extraído las líneas de texto para equipos de investigación que estén interesados únicamente en el proceso de transcripción del texto. El corpus contiene 14,948 líneas de texto manuscrito. Además, contiene 10,583 palabras diferentes.

5.2 Preprocesamiento

Con el objetivo de hacer comparables los resultados de Ptak (Ptak et al., 2017) con el método que aquí se propone, se aplicaron los mismos métodos de preprocesamiento a todas las colecciones. En este caso a todas las imágenes de los documentos en las colecciones se les ha corregido el ángulo de inclinación con la transformada de Radon (Helgason, 1999).

Para esta tarea se han presentado corpus que fueron segmentados por expertos y contienen un conjunto de puntos que permite generar un conjunto de trazos para segmentar las líneas de texto manuscrito. Para este trabajo se utilizaron dos corpus, el primero está presentado en (Saabni et al., 2014), este corpus está compuesto por 122 páginas en cuatro diferentes idiomas: inglés, chino, hindú y español.

El corpus propuesto en (Saabni et al., 2014) contiene un total de 500 líneas de texto manuscrito. En la Figura 1.1 se muestra una página del primer corpus usado en este trabajo. El segundo corpus usado durante esta investigación está compuesto por 30 páginas de la obra "Suma de Visitas a los Pueblos de México" (García Castro, 2013).

5.3 Método de evaluación

Con el objetivo de mostrar la relevancia de nuestros métodos propuestos para la SLT se han dividido las etapas LLT y la BRSLT para mostrar su evaluación.

Para evaluar la etapa de la LLT se usó una métrica basada en la presentada en (Ptak et al., 2017), no se usó la misma métrica de evaluación porque solamente evalúa el número de separadores mal identificados y el número de falsos positivos está limitado a uno por página.

De manera similar a la métrica presentada en (Ptak et al., 2017), evaluamos el rendimiento de los métodos para la LLT de acuerdo al número de líneas de texto manuscrito correctamente identificadas (verdaderos positivos) menos el número de separadores incorrectos (falsos positivos). Un separador se considera correcta, si se localiza entre el centro de dos líneas vecinas adyacentes.

En la métrica de evaluación propuesta se consideran dos tipos de error. El primer tipo de error ocurre cuando dos líneas de texto vecinas son identificadas como una sola línea. El segundo tipo de error ocurre cuando se coloca más de un separador que segmenta una línea de texto como dos o más líneas de texto. Es importante considerar los dos tipos de error porque afectan la tarea del reconocimiento de texto manuscrito (RTM)(Arica & Yarman-Vural, 2001; V. Romero, J. A. Sánchez, V. Bosch, K. Depuydt, & J. de Does, 2015).

Existen diferentes métodos de evaluación para la BRSLT, sin embargo, muchos de los métodos de evaluación recientes están basados en la métrica *Match-Score*. La métrica *Match-Score* fue presentada por Yanikoglu et. al. (Yanikoglu & Vincent, 1998) y está definida como el porcentaje de los píxeles en el fondo G_u que están cubiertos por R_u menos el porcentaje de los píxeles en el frente de R_u que se encuentran fuera de G_v.

Dado G_u el conjunto de todos los puntos de la región i *groundtruth*, R_v el conjunto de todos los puntos de la región j resultante, $T(s)$ es una función que cuenta los elementos del conjunto s. La table $MatchScore(u, v)$ representa el número de coincidencia de la región i del *groundtruth* y la región j resultante de la siguiente forma:

$$MatchScore(u, v) = \frac{T(G_u \cap R_v)}{T(G_u \cup R_v)}$$

Es importante mencionar que esta métrica de evaluación fue usada para medir el rendimiento de los métodos para la SLT en ICDAR 2007, ICDAR2010, ICDAR2013 en la competencia Handwriting segmentation.

Existen pocos métodos multilenguaje para la SLT, por lo que la comparación que aquí se presenta es con los métodos del mismo tipo de Ptak (Ptak et al., 2017) y Arvanitopoulos (Arvanitopoulos & Süsstrunk, 2014).

El sistema de Arvanitopoulos (Arvanitopoulos & Süsstrunk, 2014) realiza la tarea de la LLT y la BRSLT en documentos históricos y demostró que es mejor que el sistema propuesto en (Saabni et al., 2014). Sin embargo, el sistema de Ptak (Ptak et al., 2017) si solamente para la tarea de la LLT por lo que no es capaz de realizar la tarea completa de la SLT. Para los sistemas con los que

se compara nuestro método es necesario definir manualmente un conjunto de parámetros, por lo que fue necesario realizar una búsqueda exhaustiva en todas las colecciones probando todas las combinaciones posibles. Los resultados que a continuación se reportan son los mejores obtenidos. Es importante recalcar que los siguientes resultados de la evaluación se realizaron con sobre las mismas colecciones de documentos presentados en (Saabni et al., 2014).

La tabla 5.1 y 5.2 muestran la evaluación de acuerdo al lenguaje, los parámetros con los que se obtienen mejores resultados en la etapa de experimentación para los sistemas de Ptak (Ptak et al., 2017) y Arvanitopoulos (Arvanitopoulos & Süsstrunk, 2014).

Tabla 5.1. Mejor configuración encontrada para el método de Arvanitopoulos (Arvanitopoulos & Süsstrunk, 2014).

Lenguaje	smooth	slides	sigma	offset
Árabe	0.0001	6	2	1
Chino	0.00007	7	2	4
Inglés	0.001	3	2	1
Español	0.001	3	2	1
Árabe-Español	0.00001	3	10	2
Español Antiguo	0.03	7	3	6

Tabla 5.2. Mejor configuración encontrada para el método de Ptak (Ptak et al., 2017)

Lenguaje	umbral	slides
Árabe	0.98	12
Chino	0.997	12
Inglés	0.99	35
Español	0.99	35
Árabe-Español	0.99	35
Español Antiguo	0.97	15

5.4 Experimentación para LTL

En esta sección se presentan los experimentos realizados con nuestro método propuesto para LLT en cuatro lenguajes (árabe, chino, inglés y español) y la colección de documentos combinada de árabe y español.

Para nuestro método propuesto se ajustaron automáticamente los parámetros $\mathrm{Wmin} = 7$, $\mathrm{Wmax} = 30$, $\mathrm{Rmin} = 10$ and $\mathrm{Rmax} = 50$; usando cinco documentos aleatorios para cada colección. Para el Árabe, Inglés, Español y la colección Árabe-Español se obtuvo $\mathrm{r} = 30$ and $\mathrm{w} = 7$, solo para la colección en lenguaje Chino cambiaron los parámetros con $\mathrm{r} = 30$ y $\mathrm{w} =$

15. Después de realizar algunos experimentos el porcentaje de umbral para eliminar los mínimos locales irrelevantes fue de 10% para todas las colecciones.

La Tabla 5.3 muestra los resultados obtenidos con nuestro método para las cinco colecciones y una comparación con los sistemas de Arvanitopoulos (Arvanitopoulos & Süsstrunk, 2014) y Ptak (Ptak et al., 2017). Además, en la Tabla 5.3 se incluye el promedio de las evaluaciones por sistema.

Tabla 5.3. Resultados para la identificación de líneas en toda la colección.

Método	Arvanitopoulos (Arvanitopoulos & Süsstrunk, 2014)	Ptak (Ptak et al., 2017)	PEM-Alpha propuesto
Árabe	94.45%	96.33%	98.98%
Chino	97.18%	95.77%	98.59%
Inglés	94.13%	94.84%	99.20%
Español	98.30%	92.38%	98.72%
Árabe-Español	75.35%	87.09%	97.18%
Español Antiguo	65.3%	91.2%	97.32%
Promedio	76.56%	92.93%	98.33%

Analizando los resultados de la Tabla 5.3 se concluye que el método PME-Alfa tiene la mejor exactitud comparado con otros sistemas. Es importante recalcar que otros métodos se ven más afectados con los documentos de la colección combinada.

5.5 Experimentación para la tarea completa de SLT

En esta sección comparamos nuestro método propuesto PEM-Alfa + BRSLT con el sistema de Arvanitopoulos (Arvanitopoulos & Süsstrunk, 2014). Con el objetivo de ajustar los parámetros del algoritmo genético, los siguientes parámetros se plantearon pata toda la etapa de experimentación: número de cromosomas 10; número de genes por cromosoma $n = 100$; tamaño del torneo 3; puntos de cruza 2; probabilidad de mutación 0.2% y condición de paro 70 generaciones.

La Tabla 5.4 muestra los resultados obtenidos por nuestro método propuesto. Como puede verse nuestro método supera en todas las colecciones al sistema de Arvanitopoulos (Arvanitopoulos & Süsstrunk, 2014). Además, en la Tabla 5.4 se incluye el promedio de la evaluación por sistema.

Tabla 5.4. Comparación de exactitud del método propuesto contra el método de Arvanitopoulos (Arvanitopoulos & Süsstrunk, 2014).

Método	Arvanitopoulos (Arvanitopoulos & Süsstrunk, 2014)	PEM-Alpha+BRL-GA propuesto
Árabe	93.68%	97.83%
Chino	99.35%	99.68%
Inglés	99.35%	99.68%
Español	97.30%	98.72%
Árabe-Español	82.90%	90.27%
Español Antiguo	58.6%	96.34%
Promedio	**88.53%**	**97.08%**

Las Figura 5.3, 5.4, 5.5 y 5.6 muestran en la parte superior un ejemplo de separación de líneas generado por el humano, en la parte media se muestran las rutas generadas generadas por nuestro método y en la parte baja se muestran las rutas generadas por el sistema de Arvanitopoulos; para el inglés, árabe y la colección combinada, respectivamente.

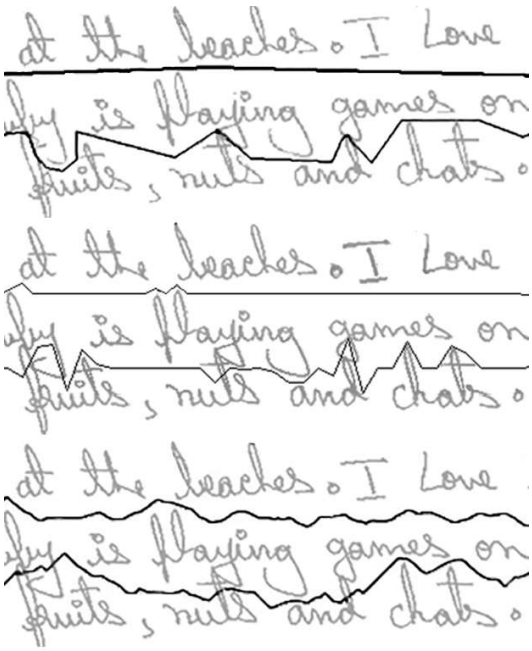


Figura 5.3. Comparación visual de los métodos evaluados para el lenguaje inglés.

Figura 5.4. Comparación visual del método evaluados en el lenguaje árabe.

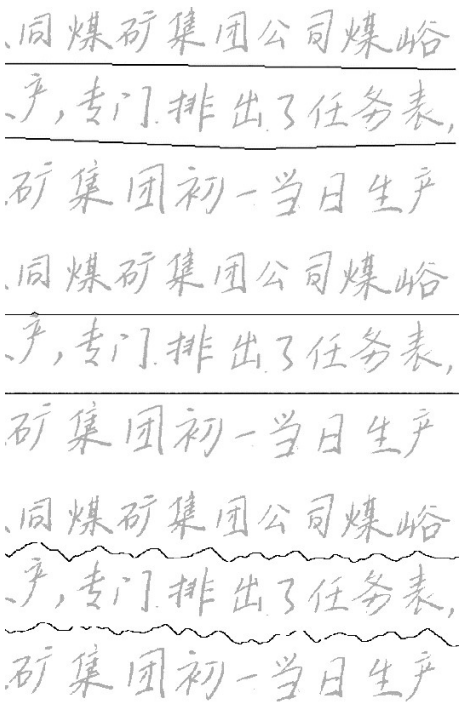


Figura 5.5. Comparación visual del método evaluados en el lenguaje árabe.

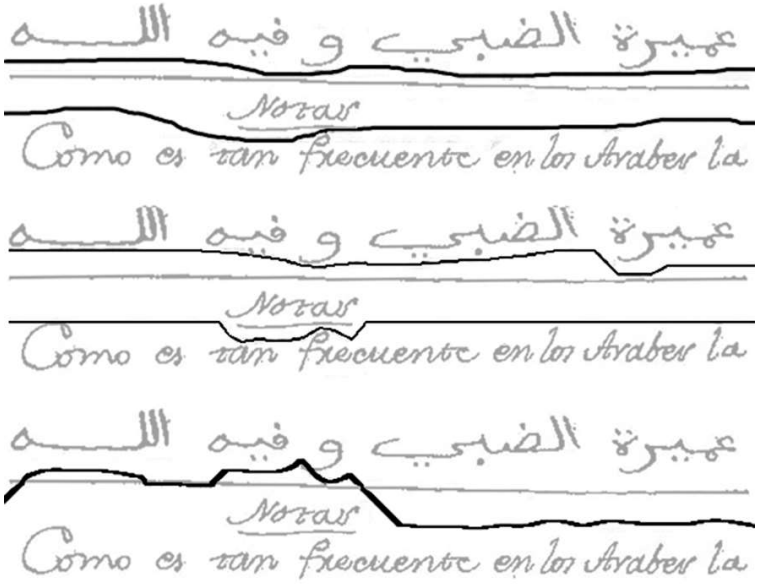


Figura 5.6. Comparación visual del método evaluados en el lenguaje árabe.

CAPÍTULO 6.

Conclusiones y Trabajo Futuro

En este capítulo se presentan las conclusiones generadas a partir de la etapa de experimentación realizada en este trabajo. Este capítulo se encuentra dividido en tres subsecciones: conclusiones, aportaciones y trabajo futuro.

6.1. Conclusiones

Con los experimentos que se presentan en la tabla 5.1 y específicamente con la fila correspondiente a la colección *español antiguo* se sabe que para el primer sub-problema abordado en esta tesis que consiste en: ¿Cómo desarrollar un método computacional para la LLT en documentos con líneas de texto que se intersectan verticalmente? Se planteó mejorar el rendimiento de la localización de líneas de texto mediante la búsqueda de los mínimos locales (espacios blancos entre las líneas) en un perfil de proyección horizontal de un mapa de energía alfa de un documento manuscrito; el cual redujo el valor de los mínimos locales incluso cuando existan líneas que se tocan verticalmente o se solapan.

Para el segundo sub-problema de esta tesis que consiste en ¿Cómo desarrollar un método automático e independiente del lenguaje que genere una ruta que permitan segmentar líneas de texto manuscrito que además sea un método no supervisado y no dependa de los materiales del documento, la época del documento y la caligrafía del autor? Se planteó mejorar la exactitud en la BRSLT calculando una ruta no lineal con una función que trate de cruzar por la menor cantidad de letras minimizada globalmente desde el punto inicial hasta el

punto final del documento. La búsqueda global de la ruta para segmentar líneas de texto ha permitido superar la exactitud de los métodos del estado del arte en la tarea de la SLT.

Los resultados presentados en el capítulo anterior indican que con los métodos propuestos y las hipótesis planteadas en esta tesis se ha superado la exactitud de los métodos actuales para la LLT y BRSLT.

Los objetivos completados durante el desarrollo de esta tesis son los siguientes:

- Desarrollar un método independiente del lenguaje.
- Mejorar los resultados del estado del arte.
- Desarrollar un método no supervisado.
- Desarrollar todo este trabajo mediante la metodología de desarrollo de software orientado a componentes.

Como parte del trabajo futuro hace falta validar que el proceso de reconocimiento de escritura manuscrita mejore después de implementar los métodos que se han propueso en esta tesis.

El análisis de imágenes de documentos manuscritos es una etapa importante para facilitar el acceso a la información que contienen. Es importante recalcar que el tesoro más valioso de la humanidad es el conocimiento que se genera por diversas culturas y que se encuentra digitalizado. En este trabajo se presentaron dos métodos MEPP-Alpha y el método BRSLT-GA para la segmentación multilenguaje de líneas de texto, que es una etapa del reconocimiento de escritura manuscrita.

El método MEPP-Alpha que se propuso para la etapa de la LLT puede ajustarse automáticamente para localizar las líneas de texto en diferentes lenguajes, por lo tanto, el método que aquí se propone es un método no supervisado. Usando la colección estándar propuesta en (Saabni et al., 2014) para cuatro lenguajes e incluso para lenguajes mezclados nuestro método MEPP-Alpha supera a otros sistemas en todas las sub-colecciones. Por lo tanto, se concluye que para la tarea de la LLT es mejor buscar los mínimos locales en un histograma de proyección horizontal de un mapa de energía basado en la mezcla alfa.

Para el problema de la búsqueda de una ruta que mejor divida líneas de texto vecinas, se propuso el método BRSL-GA que permite buscar una ruta no lineal que minimice los cortes entre letras y la distancia entre los puntos iniciales y finales. De acuerdo con la experimentación, nuestro método supera al método presentado en (Saabni et al., 2014). Como puede verse, en la Figura 5,1 y 5.2 nuestro método tiene una similaridad mayor a los trazos que genera el humano.

Es necesario enfatizar que para los métodos propuestos no fue necesario ajustar empíricamente los parámetros para cada subcolección, esto es un avance en comparación

a los métodos propuestos hasta el momento. Los métodos que aquí se proponen superan a otros sistemas en documentos con lenguaje español, árabe, Ingles, chino y documentos que contienen dos lenguajes (árabe y español). Además, los métodos propuestos en este trabajo son métodos no supervisados para la segmentación de líneas de texto en documentos manuscritos.

A diferencia de los métodos presentados en (Arvanitopoulos & Süsstrunk, 2014; Du et al., 2009; Ptak et al., 2017) en este trabajo se demostró que para el método que aquí se propone es posible definir sus parámetros de manera automática.

6.2. Aportaciones

En los algoritmos desarrollados:

Una medida que permite comparar los resultados de los métodos para la localización de líneas de texto manuscrito.

Un nuevo algoritmo para extraer el mapa de energía de documentos manuscritos que permite incrementar las diferencias entre los máximos y mínimos locales en un histograma de proyección horizontal.

Un alogitmo para calcular automáticamente los mejores parámetros para la extracción del mapa de energía propuesto. En todos los métodos analizados es necesario definir empíricamente los parámetros para los métodos.

Un algoritmo para la localización de líneas de texto tomando como referencia los mínimos locales en el histograma de proección horizontal.

Un nuevo método que realiza una búsqueda global de la ruta que permita segmentar líneas de texto manuscrito con un algoritmo genético.

En datos

En esta investigación se generó un corpus para la segmentación de líneas de texto manuscrito antiguo en español antiguo. El corpus generado durante el desarrollo de este trabajo se encuentra disponible para evaluar los métodos para la segmentación de líneas de texto manuscrito.

Se calculó la varianza que existe entre diferentes personas para segmentar las mismas líneas de texto manuscrito, esto permitió desarrollar el algoritmo de evaluación propuesto en este trabajo. Esta varianza fue utilizada como parámetro para el método de evaluación propuesto en este trabajo. La varianza calculada siempre es diferentes dependiendo del corpus.

La varianza que existe en el humano para segmentar líneas de texto manuscrito fue de 8 píxeles. La varianza en el humano para segmentar las líneas de texto en el corpus de (Arvanitopoulos & Süsstrunk, 2014) fue de 10 píxeles.

Se usó la técnica de histograma de proyección vertical para establecer un baseline para el corpus propuesto en este trabajo y para el corpus propuesto en (Arvanitopoulos & Süsstrunk, 2014).

Trabajo futuro

En el futuro es interesante hacer pruebas en documentos con una complejidad mayor, donde incluso para el humano es más difícil trazar una línea de corte.

Como trabajo futuro se plantea agregar una etpa de agrupamiento al método propuesto para permitir que sea aplicado a documentos como los mostrados en las figuras 3.1 y 3.6. Después de identificar los grupos de texto es corregir el ángulo de inclinación para cada grupo y posteriormente se aplicar los métodos que se proponen en esta tesis

Al revisar los trabajos del estado del arte se ha observado que no existe una comparación de todos los métodos actuales usando un mismo corpus por lo cual se pretende trabajar en una comparación de los métodos del estado del arte actuales. Para el proceso de segmentación de líneas de texto manuscrito.

Es necesario compara los métodos actuales con corpus que tengan diferente complejidad para segmentar líneas de texto.

El método propuesto actual define un conjunto de puntos que se mueven sobre el eje y, como trabajo futuro se propone mejorar el método para que la cantidad de puntos sea variable y que el ajuste de puntos se realice también en el eje x.

En esta tesis se ha mejorado la etapa de SLT, por lo cual es necesario pasar a la etapa de reconocimiento para medir la influencia que tiene la SLT en el reconocimiento de escritura manuscrita.

Referencias

A. Prachanucroa, & S. Phongsuphap. (2013). Marginal noise removal for scanned document images by projection profile based method. In *The 2013 10th International Joint Conference on Computer Science and Software Engineering (JCSSE)* (pp. 17–20). https://doi.org/10.1109/JCSSE.2013.6567312

Archivo General de Indias. (n.d.). [Portada del Archivo General de Indias]. Retrieved from http://www.mecd.gob.es/cultura/areas/archivos/mc/archivos/agi/portada.html

Arica, N., & Yarman-Vural, F. T. (2001). An overview of character recognition focused on off-line handwriting. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *31*(2), 216–233. https://doi.org/10.1109/5326.941845

Arvanitopoulos, N., & Süsstrunk, S. (2014). Seam Carving for Text Line Extraction on Color and Grayscale Historical Manuscripts. In *2014 14th International*

Conference on Frontiers in Handwriting Recognition (pp. 726–731). https://doi.org/10.1109/ICFHR.2014.127

Avidan, S., & Shamir, A. (2007). Seam Carving for Content-aware Image Resizing. *ACM Trans. Graph.*, *26*(3). https://doi.org/10.1145/1276377.1276390

B. Gatos, K. Ntirogiannis, & I. Pratikakis. (2009). ICDAR 2009 Document Image Binarization Contest (DIBCO 2009). In *2009 10th International Conference on Document Analysis and Recognition* (pp. 1375–1382). https://doi.org/10.1109/ICDAR.2009.246

Bagdanov, A., & Kanai, J. (1998). Projection profile based skew estimation algorithm for JBIG compressed images. *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, *1*(1), 43–51. https://doi.org/10.1109/icdar.1997.619878

Bagley, R. W. (2004). Anyang Writing and the Origin of the Chinese Writing System. In S. D. Houston (Ed.), *The first writing: script invention as history and process* (pp. 190–249). Cambridge: Cambridge University Press. Retrieved from https://contentstore.cla.co.uk//secure/link?id=5e2dff9d-5c36-e711-80c9-005056af4099

Baines, J., Bennet, J., & Houston, S. D. (2008). *The disappearance of writing systems : perspectives on literacy and communication*. London; Oakville, CT: Equinox.

Biswas, S., & Das, A. K. (2012). Writer Identification of Bangla Handwritings by Radon Transform Projection Profile. In *2012 10th IAPR International Workshop on Document Analysis Systems* (pp. 215–219). https://doi.org/10.1109/DAS.2012.98

Burger, W., & Burge, M. J. (2008). *Digital Image Processing: An Algorithmic Introduction Using Java*. Springer-Verlag London.

Causer, T., & Wallace, V. (2012). Building a Volunteer Community: Results and Findings from Transcribe Bentham. *Digital Humanities Quarterly*, *6*(2). Retrieved from http://www.digitalhumanities.org/dhq/vol/6/2/000125/000125.html

Du, X., Pan, W., & Bui, T. D. (2009). Text line segmentation in handwritten documents using Mumford-Shah model. *Pattern Recognition*, *42*(12), 3136–3145. https://doi.org/10.1016/j.patcog.2008.12.021

García Castro, R. (2013). *Suma de visitas de pueblos de la Nueva España*. Universidad Autónoma del Estado de México. Retrieved from http://ri.uaemex.mx/handle/123456789/33111

García-Hernández, R. A., & Ledeneva, Y. (2013). Single Extractive Text Summarization Based on a Genetic Algorithm. In J. A. Carrasco-Ochoa, J. F. Martínez-Trinidad, J. S. Rodríguez, & G. S. di Baja (Eds.), *Pattern Recognition: 5th Mexican Conference, MCPR 2013, Querétaro, Mexico, June 26-29, 2013. Proceedings* (pp. 374–383). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-38989-4_38

Gomez-Allende, D. M., & Gómez-Allende, D. M. (1993). *Reconocimiento de formas y visión artificial*. RA-MA. Retrieved from https://books.google.com.mx/books?id=iGY4AAAACAAJ

Gonzalez, R. C., Woods, R. E., Davue Rodríguez, F., & Rosso, L. (1996). *Tratamiento digital de imágenes*. Wilmington (Delaware, U.S.A.); [Madrid]: Addison-Wesley ; Díaz de Santos.

Gray, N. M. B. (1948). *The Paleography of Latin Inscriptions in the Eighth, Ninth and Tenth Centuries in Italy*. Macmillan. Retrieved from https://books.google.com.mx/books?id=xlGhOQAACAAJ

Gy\Hory, H. (2008). Medicine in Ancient Egypt. In H. Selin (Ed.), *Encyclopaedia of the History of Science, Technology, and Medicine in Non-Western Cultures* (pp. 1508–1518). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-1-4020-4425-0_9748

Ha, J., Haralick, R. M., & Phillips, I. T. (1995). Document page decomposition by the bounding-box project. In *Proceedings of 3rd International Conference on Document Analysis and Recognition* (Vol. 2, pp. 1119–1122 vol.2). https://doi.org/10.1109/ICDAR.1995.602115

Helgason, S. (1999). *The Radon Transform*. Birkhäuser Boston. Retrieved from https://books.google.com.mx/books?id=tq3eStnBwIUC

I. Pratikakis, K. Zagoris, G. Barlas, & B. Gatos. (2016). ICFHR2016 Handwritten Document Image Binarization Contest (H-DIBCO 2016). In *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)* (pp. 619–623). https://doi.org/10.1109/ICFHR.2016.0118

J. A. Sánchez, A. H. Toselli, V. Romero, & E. Vidal. (2015). ICDAR 2015 competition HTRtS: Handwritten Text Recognition on the tranScriptorium dataset. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 1166–1170). https://doi.org/10.1109/ICDAR.2015.7333944

Kesiman, M. W. A., Burie, J.-C., & Ogier, J.-M. (2016a). A New Scheme for Text Line and Character Segmentation from Gray Scale Images of Palm Leaf Manuscript. *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 325–330. https://doi.org/10.1109/icfhr.2016.0068

Kesiman, M. W. A., Burie, J.-C., & Ogier, J.-M. (2016b). A New Scheme for Text Line and Character Segmentation from Gray Scale Images of Palm Leaf Manuscript. *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 325–330. https://doi.org/10.1109/icfhr.2016.0068

Khandelwal, A., Choudhury, P., Sarkar, R., Basu, S., Nasipuri, M., & Das, N. (2009). Text Line Segmentation for Unconstrained Handwritten Document Images Using Neighborhood Connected Component Analysis. In *Proceedings of the 3rd International Conference on Pattern Recognition and Machine Intelligence* (pp. 369–374). New Delhi, India: Springer-Verlag.

Koppula, V. K., & Negi, A. (2014). Segmentation of closely set and touching lines in handwritten document images using fringe maps. *International Conference for Convergence for Technology-2014*, 1–6. https://doi.org/10.1109/i2ct.2014.7092176

Likforman-Sulem, L., Zahour, A., & Taconet, B. (2006). Text line segmentation of historical documents: a survey. *International Journal of Document Analysis and Recognition (IJDAR)*, 9(2–4), 123–138. https://doi.org/10.1007/s10032-006-0023-z

Liwicki, M., Indermuhle, E., & Bunke, H. (2007). On-Line Handwritten Text Line Detection Using Dynamic Programming. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, 447–451. https://doi.org/10.1109/icdar.2007.4378749

Martinsanz, G. P., PAJARES, G., & de la Cruz García, J. M. (2007). *Visión por computador. Imágenes Digitales y Aplicaciones. 2a Edición*. RA-MA S.A. Editorial y Publicaciones. Retrieved from https://books.google.com.mx/books?id=EQqsPgAACAAJ

Mauricio, V., Alejandro, T., Joan-Andreu, S., & Enrique, V. (n.d.). Overview of the ImageCLEF 2016 Handwritten Scanned Document Retrieval Task. Presented at the ImageCLEF.

Medina Morán, S. (2011). ¿Un error en la piedra de rosetta?, *30*(1), 21–27.

Muñoz y Rivero, J. (1880). *Manual de paleografia diplomatica española de los siglos XII al XVII : Método teórico-prático para aprender á leer los documentos españoles de los siglos XII al XVII*. Madrid : Moreno y Rojas.

Nicolaou, A., & Gatos, B. (2009). Handwritten Text Line Segmentation by Shredding Text into its Lines. In *Proceedings of the 2009 10th International Conference on Document Analysis and Recognition* (pp. 626–630). IEEE Computer Society.

O'Gorman, L., Sammon, M. J., & Seul, M. (2008). *Practical Algorithms for Image Analysis with CD-ROM*. Cambridge University Press. Retrieved from https://books.google.com.mx/books?id=8dXkUPv2DGYC

Peng, G., Yu, P., Li, H., & He, L. (2016). Text line segmentation using Viterbi algorithm for the palm leaf manuscripts of Dai. *2016 International Conference on Audio, Language and Image Processing (ICALIP)*, 336–340. https://doi.org/10.1109/icalip.2016.7846561

Prachanucroa, A., & Phongsuphap, S. (2013). Marginal noise removal for scanned document images by projection profile based method. In *The 2013 10th International Joint Conference on Computer Science and Software Engineering (JCSSE)* (pp. 17–20). https://doi.org/10.1109/JCSSE.2013.6567312

Ptak, R., Zygadlo, B., & Unold, O. (2017). Projection Based Text Line Segmentation with a Variable Threshold. *International Journal of Applied Mathematics and Computer Science*, *27*(1), 195–206. https://doi.org/10.1515/amcs-2017-0014

Q. N. Vo, & G. Lee. (2016). Dense prediction for text line segmentation in handwritten document images. In *2016 IEEE International Conference on Image Processing (ICIP)* (pp. 3264–3268). https://doi.org/10.1109/ICIP.2016.7532963

Rendón Rojas, M. Á. (n.d.). Relación entre los conceptos: información, conocimiento y valor. Semejanzas y diferencias. *Ci. Inf*, *34*(2), 52–61.

Saabni, R., Asi, A., & El-Sana, J. (2014). Text line extraction for historical document images. *Pattern Recognition Letters*, *35*, 23–33. https://doi.org/10.1016/j.patrec.2013.07.007

Steinherz, T., Rivlin, E., & Intrator, N. (1999). Offline cursive script word recognition – a survey. *International Journal on Document Analysis and Recognition*, *2*(2), 90–110. https://doi.org/10.1007/s100320050040

Sumano, M. E. B. (2002). *Texto de paleografía y diplomática*. Universidad Autónoma del Estado de México. Retrieved from https://books.google.com.mx/books?id=-7mSPnIcnp4C

Tseng, Y.-H., & Lee, H.-J. (1999). Recognition-based Handwritten Chinese Character Segmentation Using a Probabilistic Viterbi Algorithm. *Pattern Recogn. Lett.*, *20*(8), 791–806. https://doi.org/10.1016/S0167-8655(99)00043-4

U. V. Marti, & H. Bunke. (2001). On the influence of vocabulary size and language models in unconstrained handwritten text recognition. In *Proceedings of Sixth International Conference on Document Analysis and Recognition* (pp. 260–265). https://doi.org/10.1109/ICDAR.2001.953795

V. Romero, J. A. Sánchez, V. Bosch, K. Depuydt, & J. de Does. (2015). Influence of text line segmentation in Handwritten Text Recognition. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 536–540). https://doi.org/10.1109/ICDAR.2015.7333819

Valy, D., Verleysen, M., & Sok, K. (2016). Line Segmentation Approach for Ancient Palm Leaf Manuscripts Using Competitive Learning Algorithm. *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 108–113. https://doi.org/10.1109/icfhr.2016.0032

Voynich Manuscript, Beinecke MS 408, General Collection. (1912). Yale University.

Yanikoglu, B. A., & Vincent, L. (1998). Pink Panther: A Complete Environment For Ground-Truthing And Benchmarking Document Page Segmentation. *Pattern Recognition*, *31*(9), 1191–1204. https://doi.org/10.1016/S0031-3203(97)00137-4

Yarushkina, N. G. (2002). Genetic algorithms for engineering optimization: theory and practice. In *Proceedings 2002 IEEE International Conference on Artificial Intelligence Systems (ICAIS 2002)* (pp. 357–362). https://doi.org/10.1109/ICAIS.2002.1048127

Z. Shi, & Venu Govindaraju. (2004). Line separation for complex document images using fuzzy runlength. In *First International Workshop on Document*

Image Analysis for Libraries, 2004. Proceedings. (pp. 306–312). https://doi.org/10.1109/DIAL.2004.1263259

Anexo 1.

Muestras del corpus generado

En esta sección se presentan muestras del corpus generado en esta investigación. En la primera sub-sección se describe la necesidad actual de generar corpus para el reconocimiento de escritura en documentos manuscritos antiguos.

En los documentos de esta sección se puede distinguir que la ortografía de esa época difiere de la ortografía actual. En cada época se tienen diferentes peculiaridades de en caligrafía y ortografía. También se puede ver que se usan alguas palabras que en esta época no tienen signficado, en algunos casos usaban abreviaciones para optimizar la cantidad de recursos necesarios durante la escritura. En los siguientes documentos se muestran ejemplos de ornamentación (Ver figura 7.17). Aunque los documentos del corpus generado son de la misma época, país e idioma en todas las figuras de esta sección se presentan documentos con patrones diferentes de caligrafía, abreviaciones, interlineado, espacio entre caracteres, espacio entre palabras.

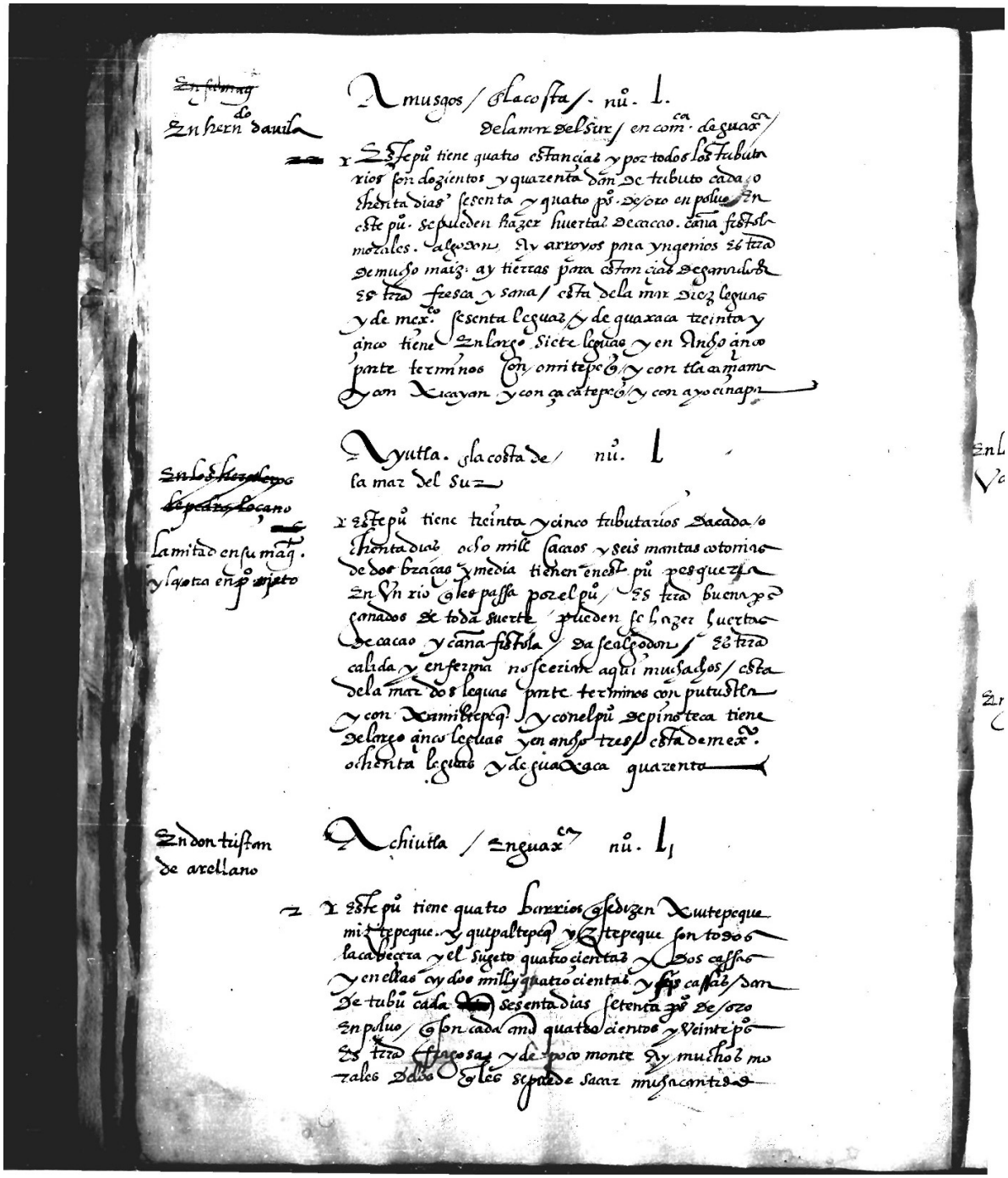


Figura 7.1. Muestra de documento del corpus generado. Página 49 en español mexicano de un period entre 1538-1540 (García Castro, 2013).

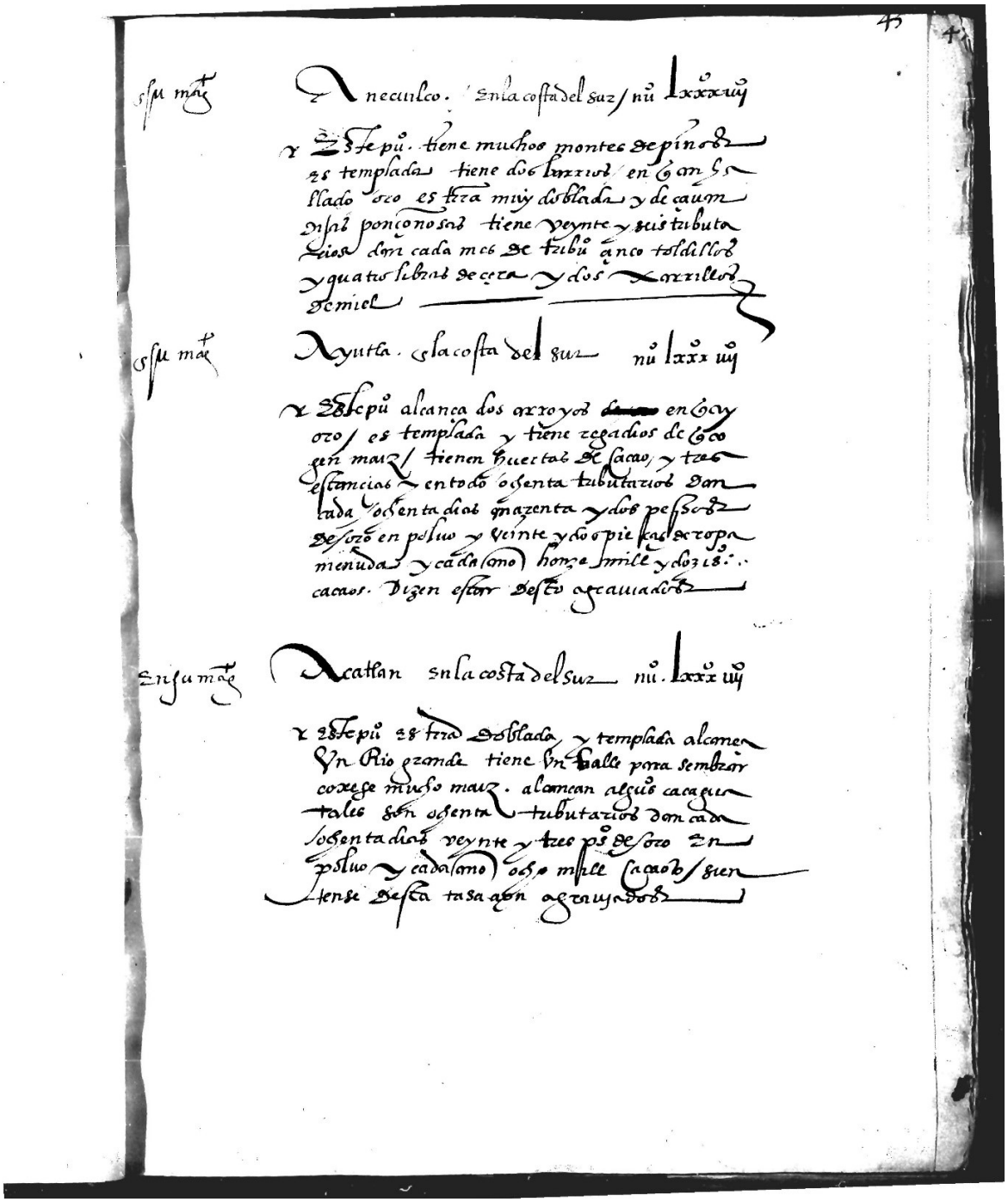


Figura 7.2. Muestra de documento del corpus generado. Página 74 en español mexicano de un period entre 1538-1540 (García Castro, 2013).

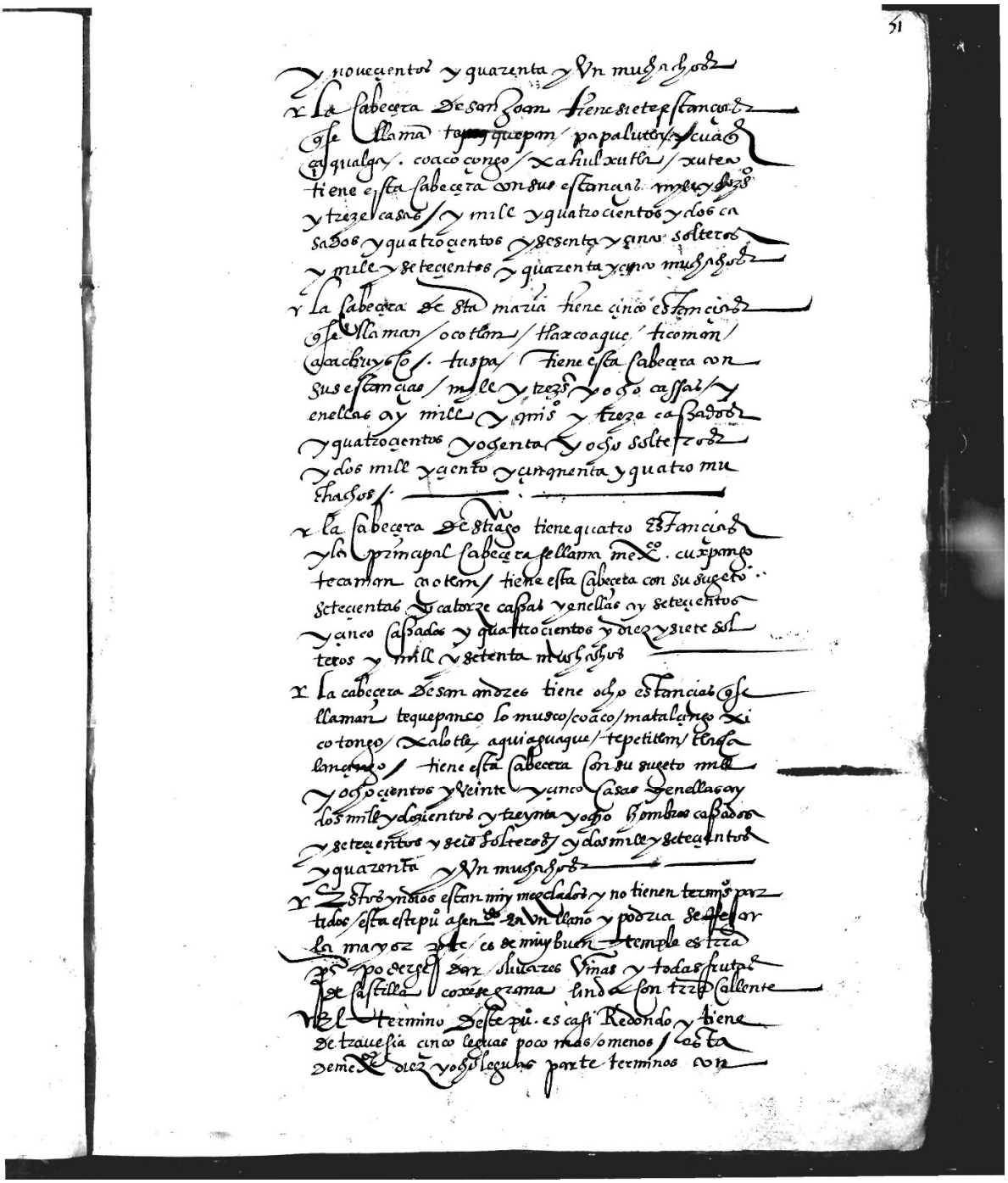


Figura 7.3. Muestra de documento del corpus generado. Página 85 en español mexicano de un period entre 1538-1540 (García Castro, 2013).

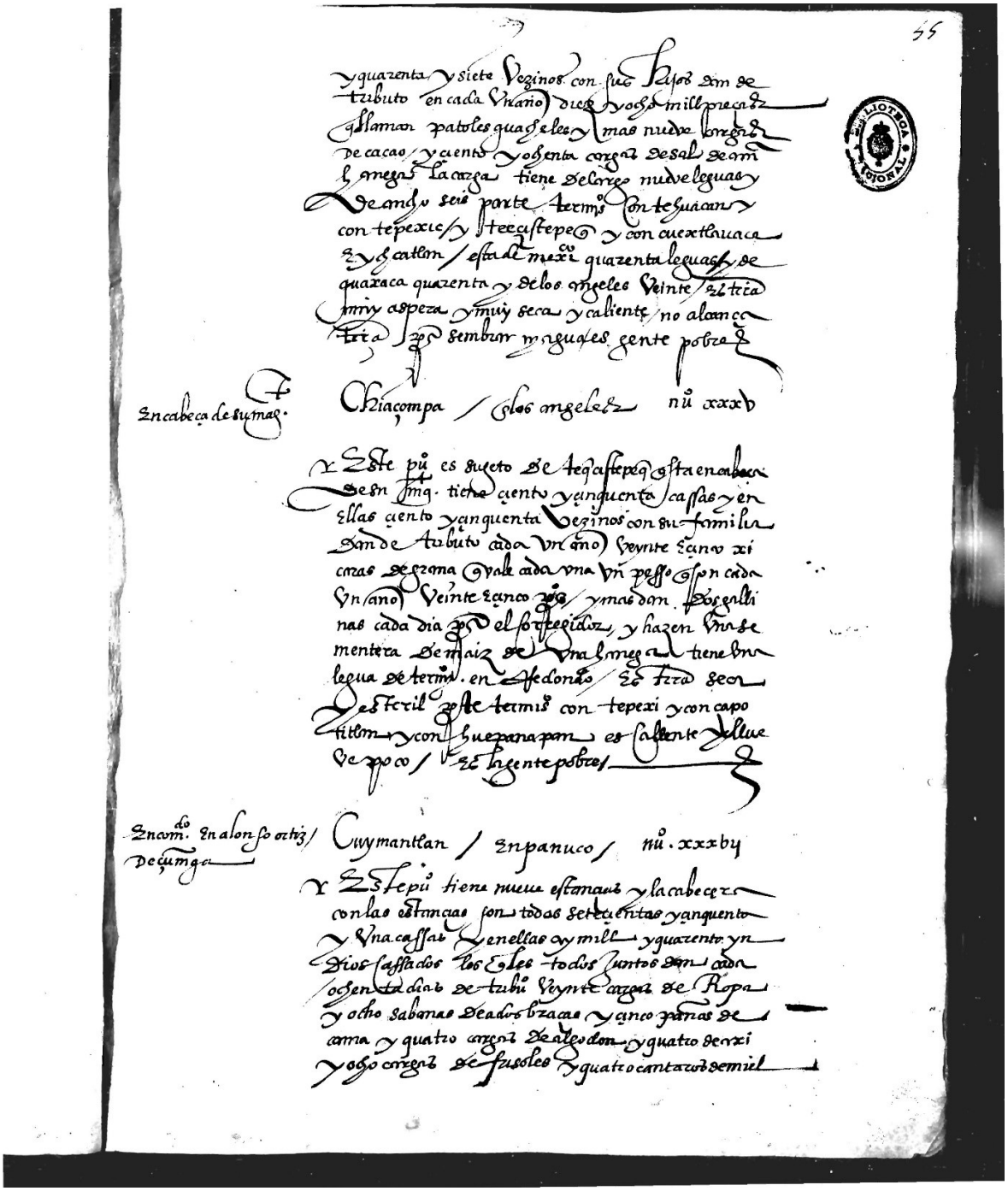


Figura 7.4. Muestra de documento del corpus generado. Página 85 en español mexicano de un period entre 1538-1540 (García Castro, 2013).

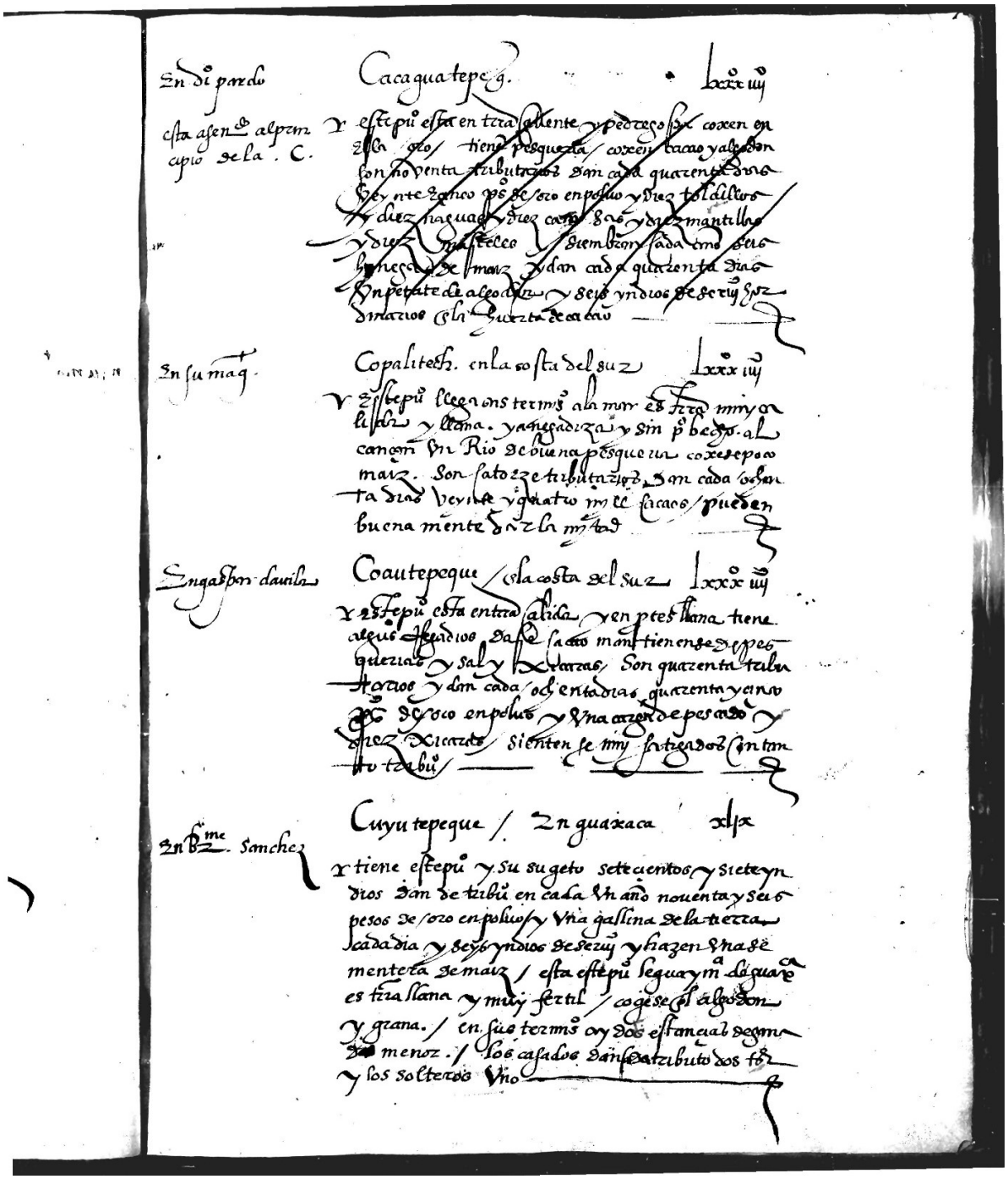


Figura 7.5. Muestra de documento del corpus generado. Página 135 en español mexicano de un periodo entre 1538-1540 (García Castro, 2013).

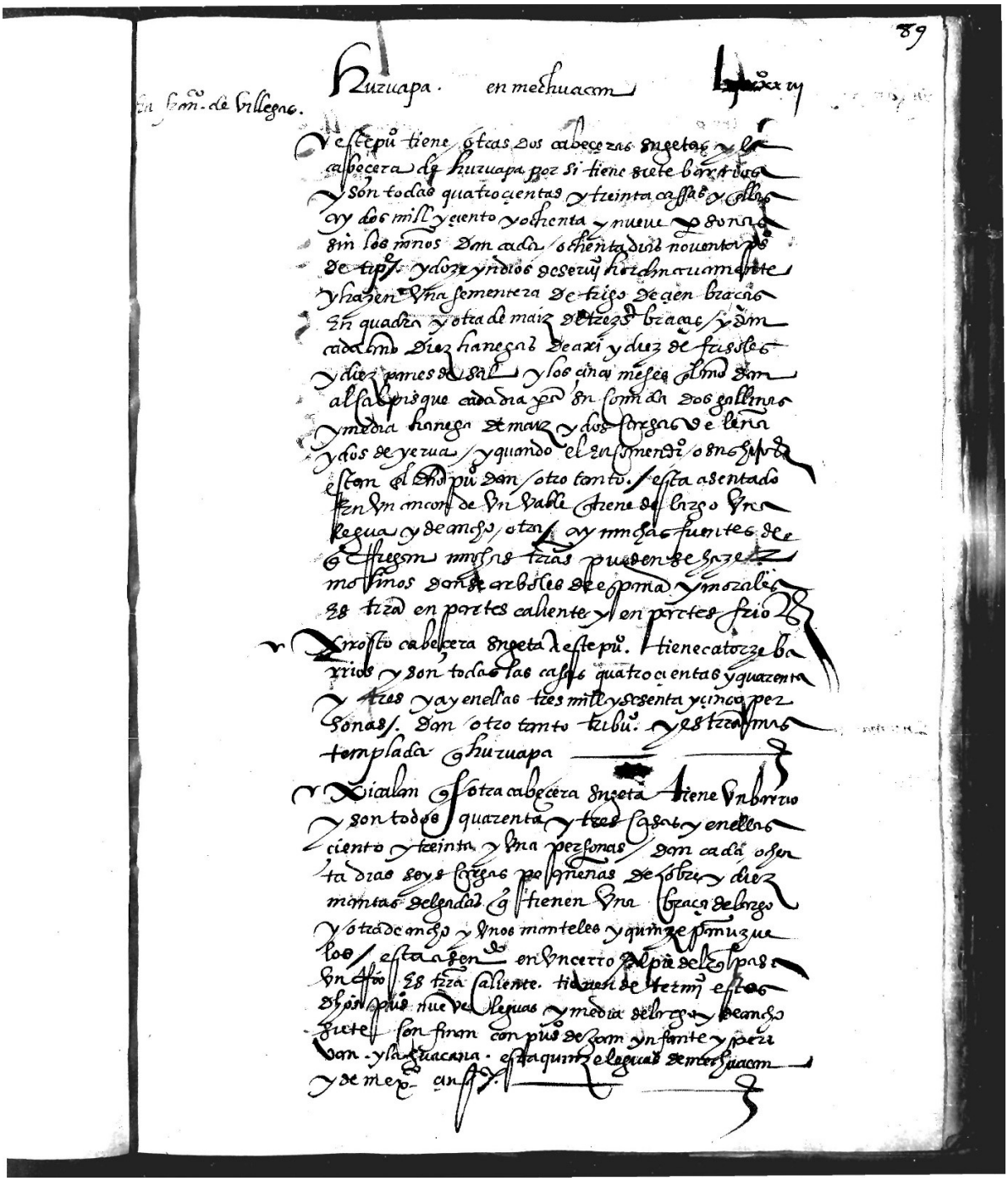


Figura 7.6. Muestra de documento del corpus generado. Página 135 en español mexicano de un periodo entre 1538-1540 (García Castro, 2013).

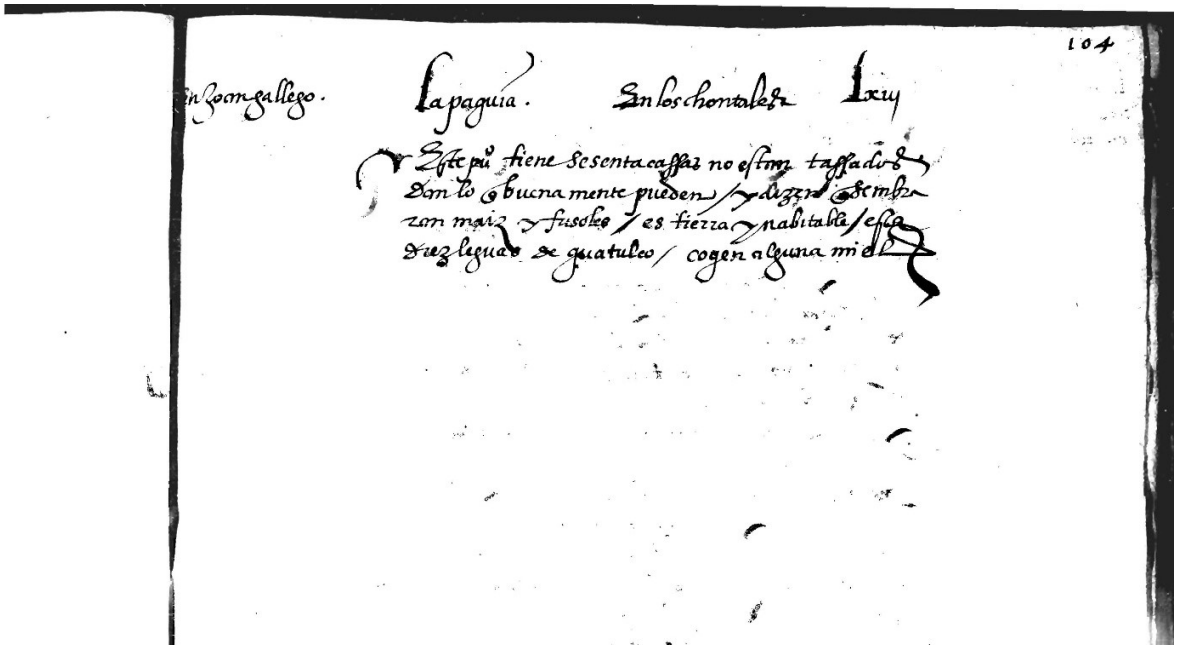


Figura 7.7. Muestra de documento del corpus generado. Página 135 en español mexicano de un period entre 1538-1540 (García Castro, 2013).

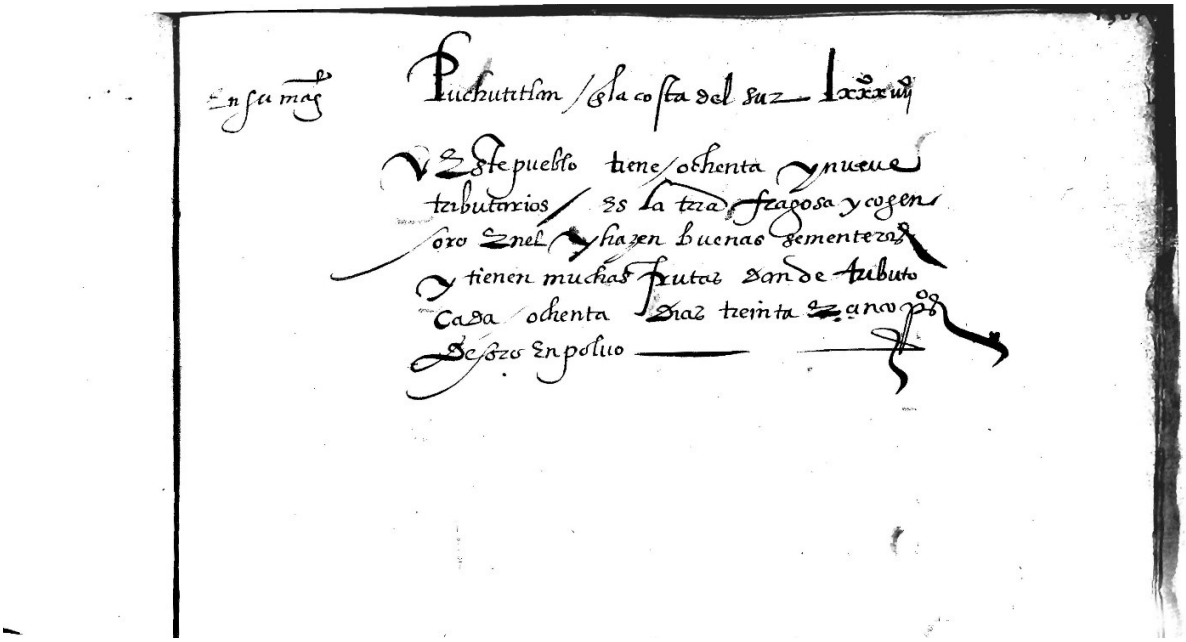


Figura 7.8. Muestra de documento del corpus generado. Página 247 en español mexicano de un period entre 1538-1540 (García Castro, 2013).

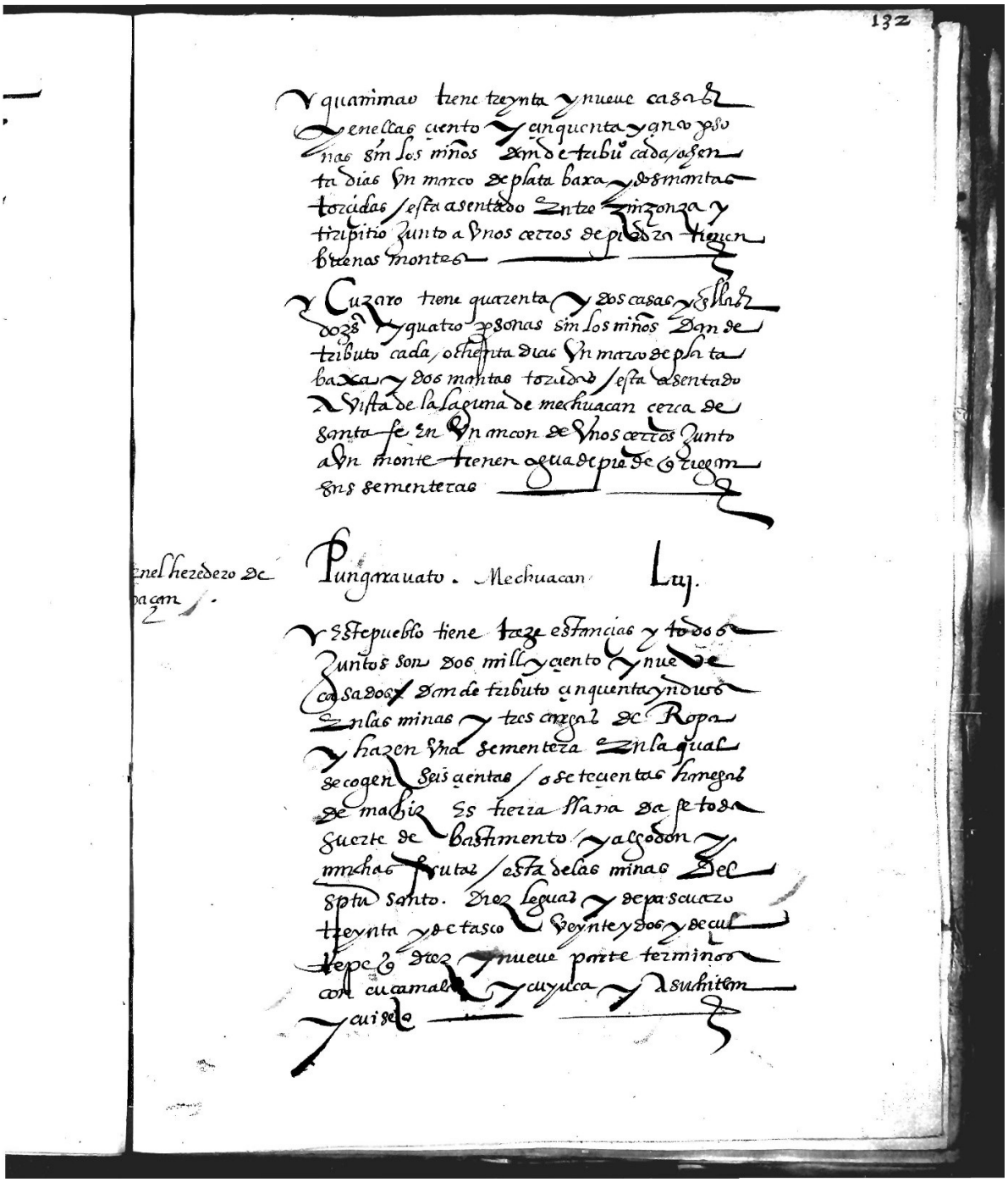


Figura 7.9. Muestra de documento del corpus generado. Página 239 en español mexicano de un period entre 1538-1540 (García Castro, 2013).

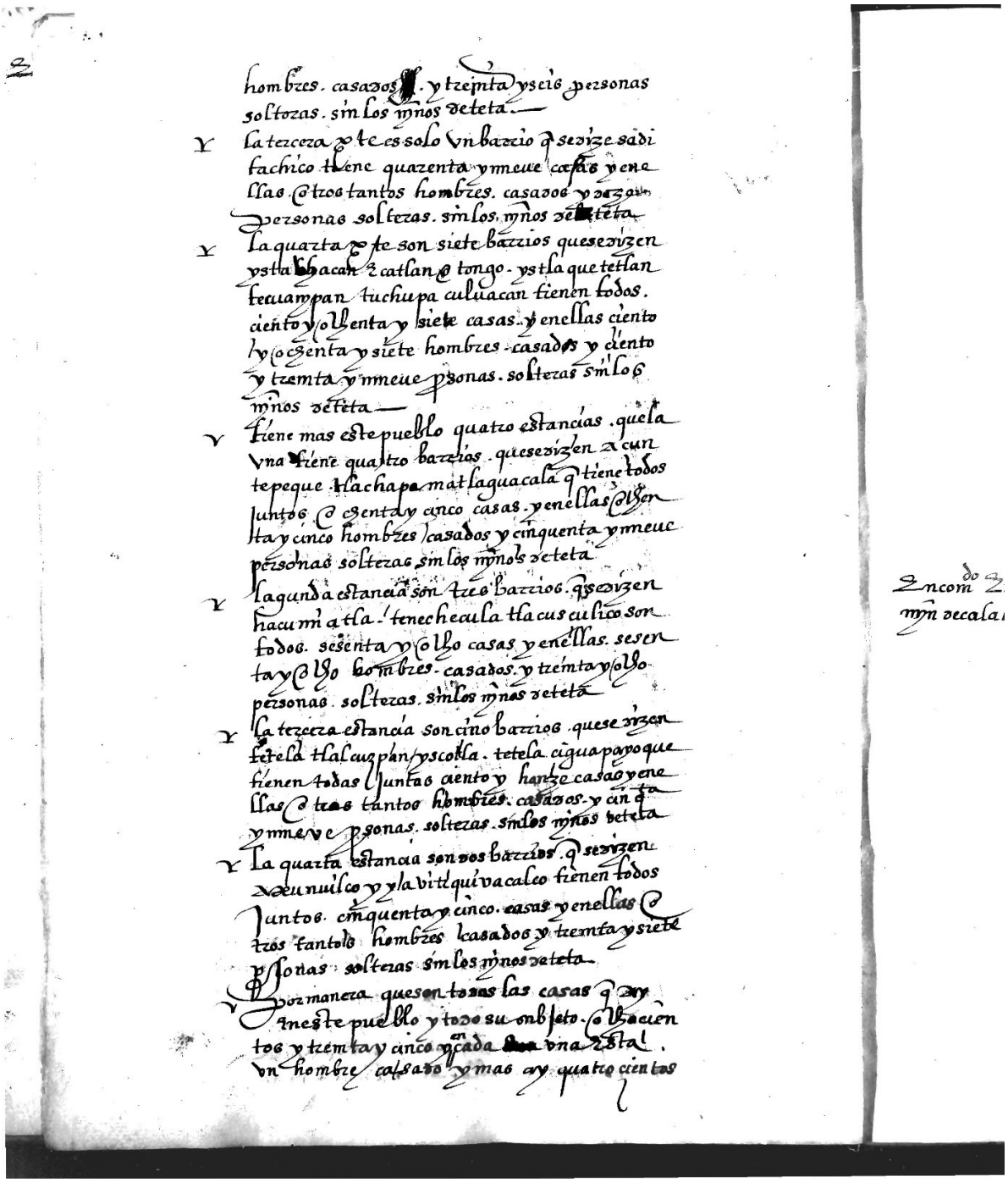


Figura 7.10. Muestra de documento del corpus generado. Página 282 en español mexicano de un period entre 1538-1540 (García Castro, 2013).

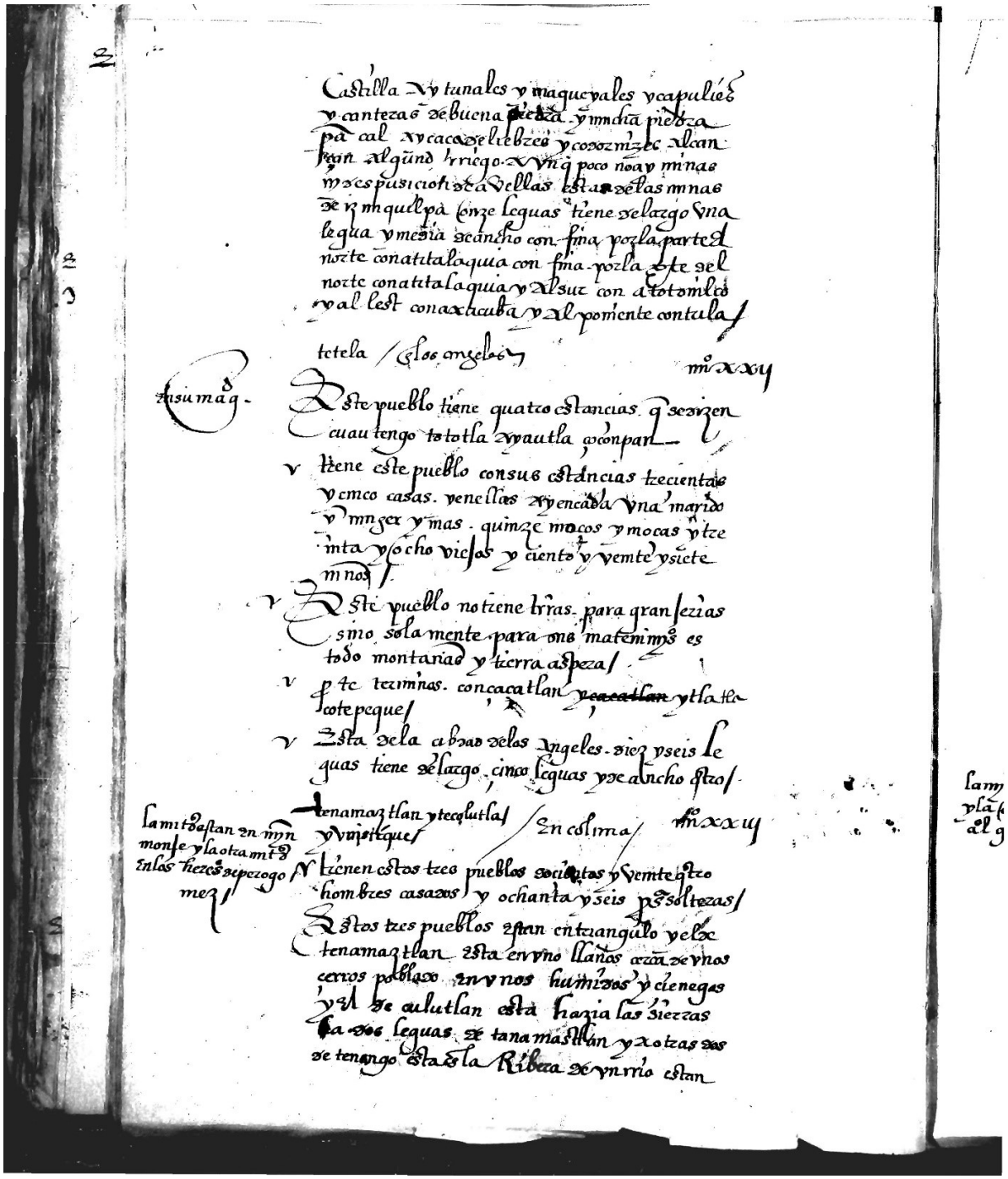


Figura 7.11. Muestra de documento del corpus generado. Página 300 en español mexicano de un period entre 1538-1540 (García Castro, 2013).

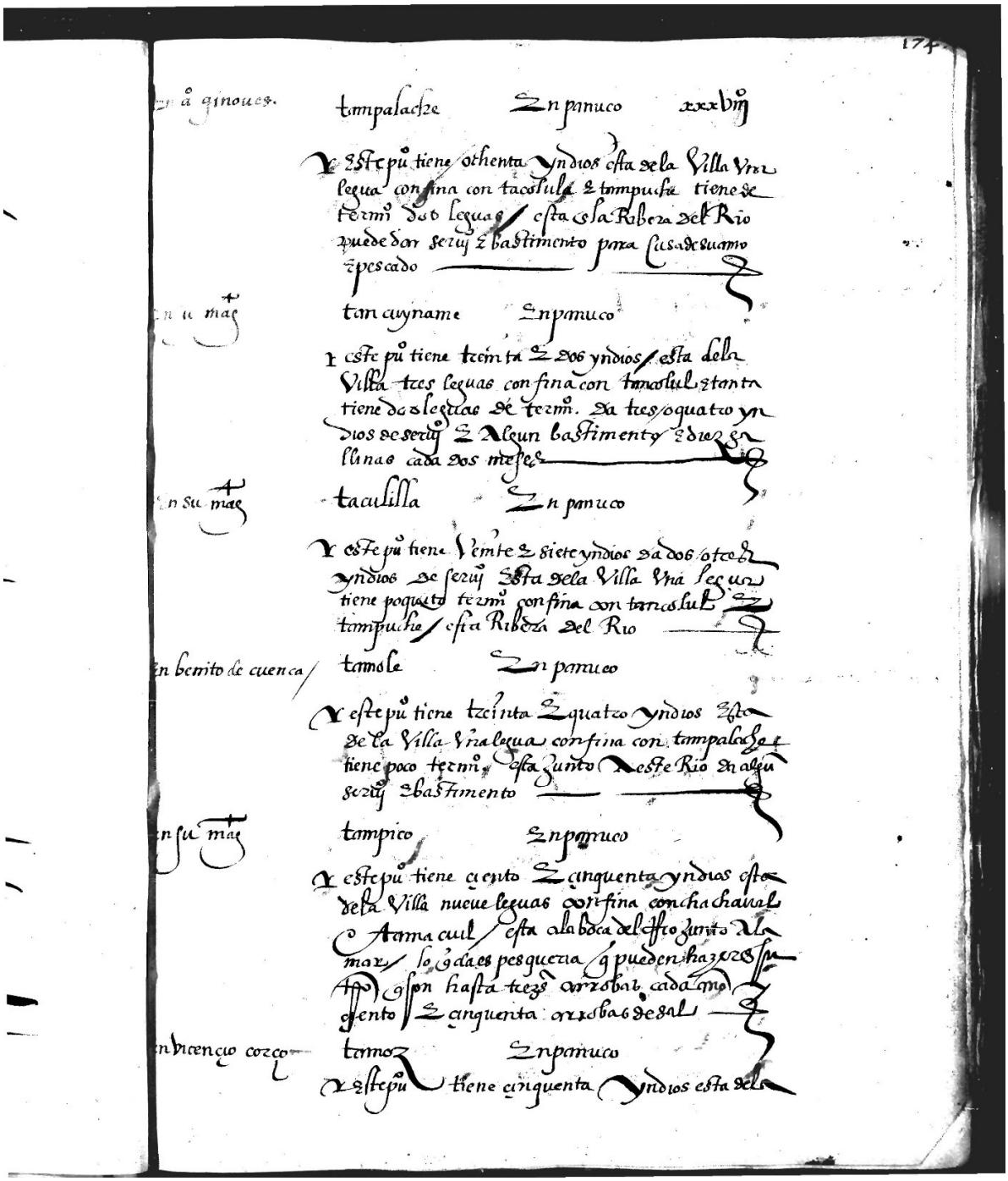


Figura 7.12. Muestra de documento del corpus generado. Página 325 en español mexicano de un period entre 1538-1540 (García Castro, 2013).

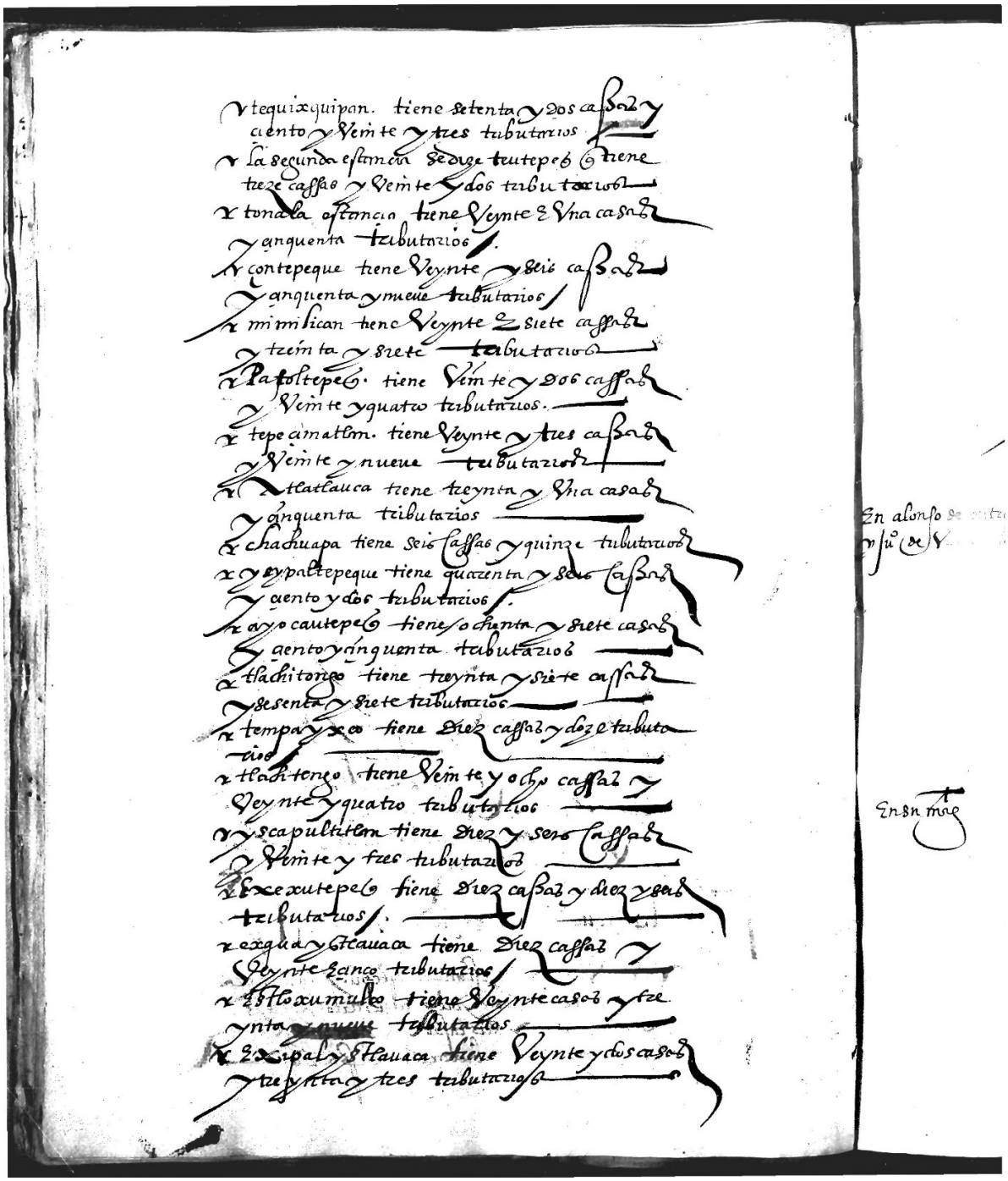


Figura 7.13. Muestra de documento del corpus generado. Página 381 en español mexicano de un periodo entre 1538-1540 (García Castro, 2013).

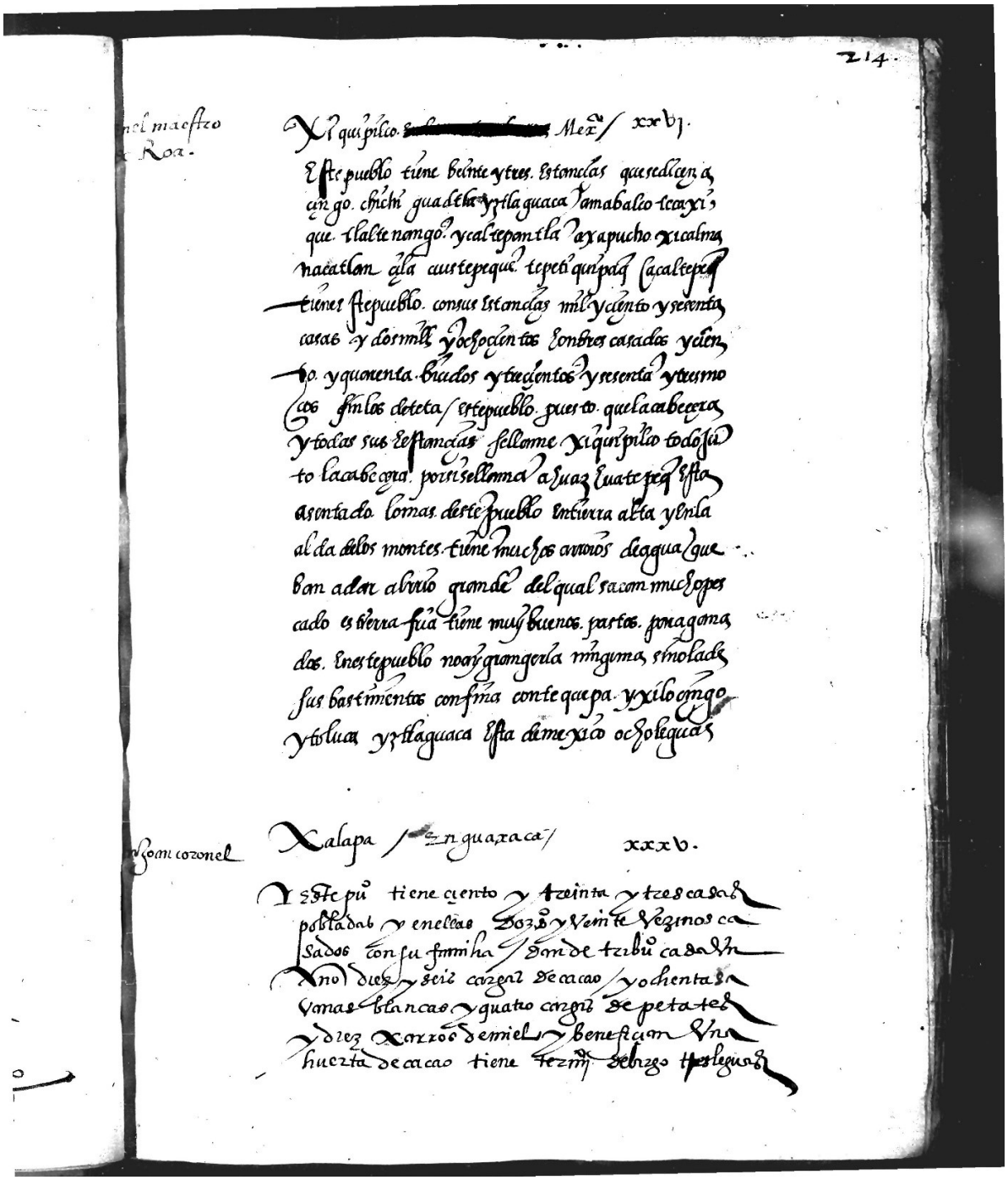


Figura 7.14. Muestra de documento del corpus generado. Página 404 en español mexicano de un period entre 1538-1540 (García Castro, 2013).

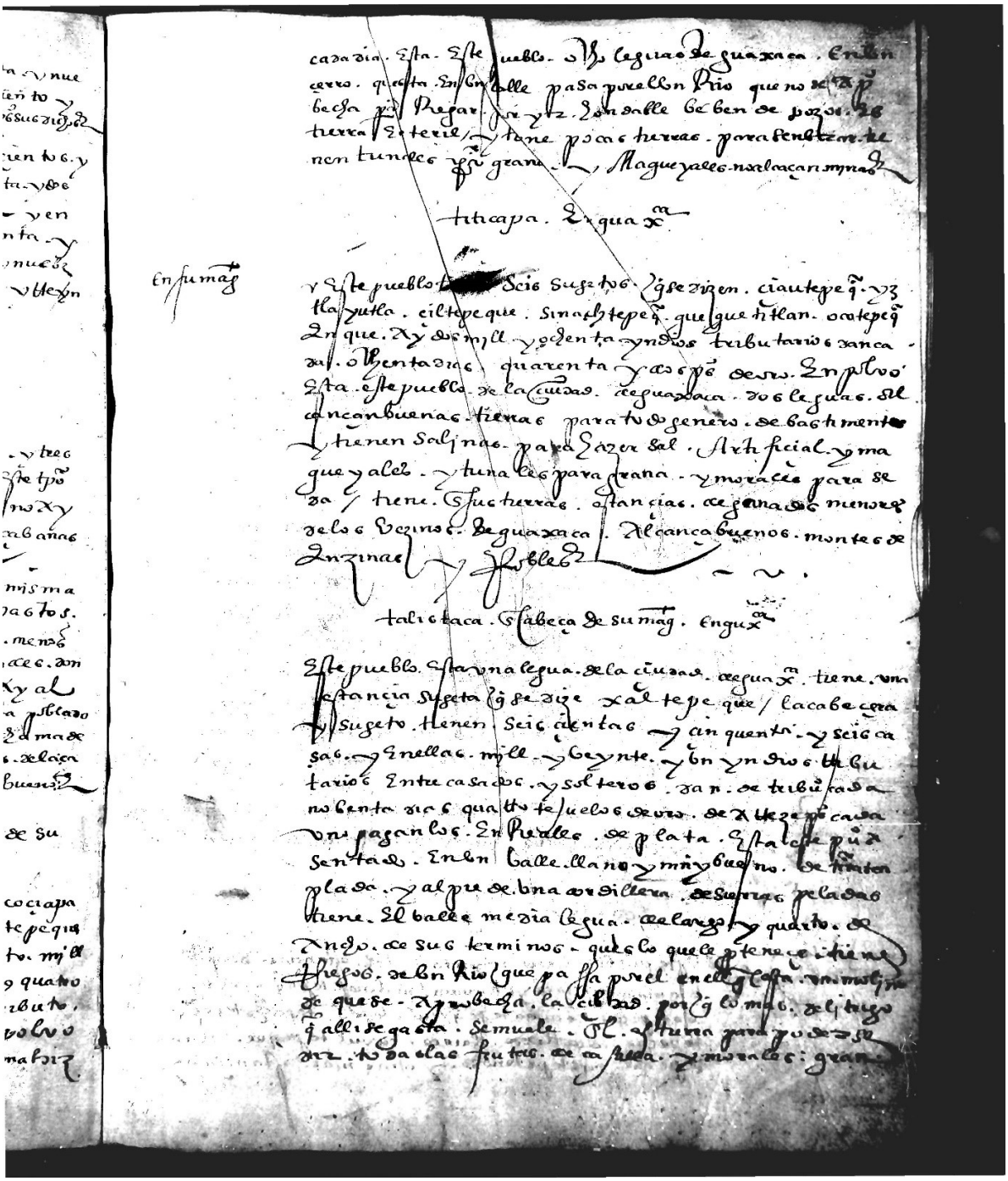


Figura 7.15. Muestra de documento del corpus generado. Página 427 en español mexicano de un periodo entre 1538-1540 (García Castro, 2013).

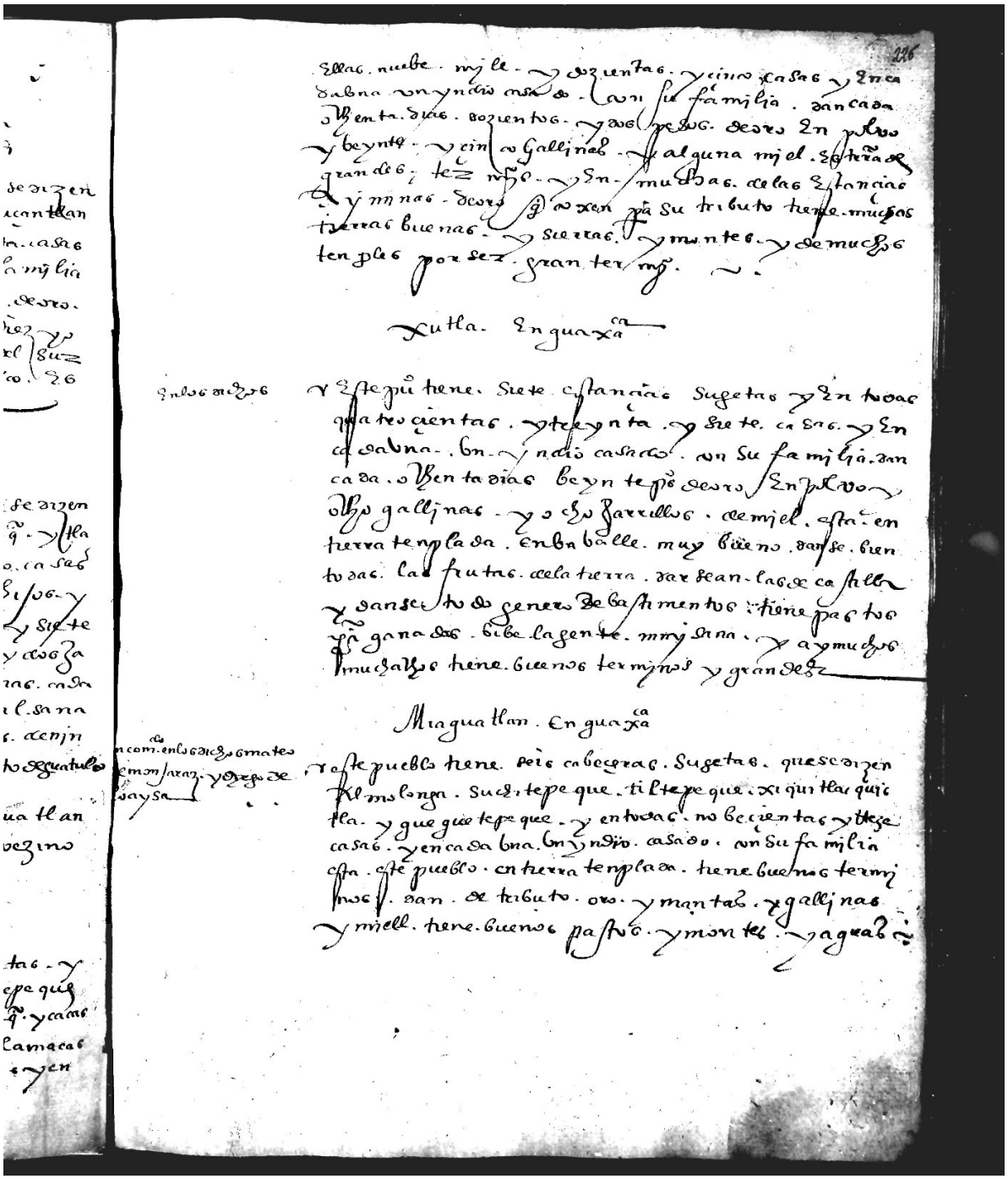


Figura 7.16. Muestra de documento del corpus generado. Página 429 en español mexicano de un period entre 1538-1540 (García Castro, 2013).

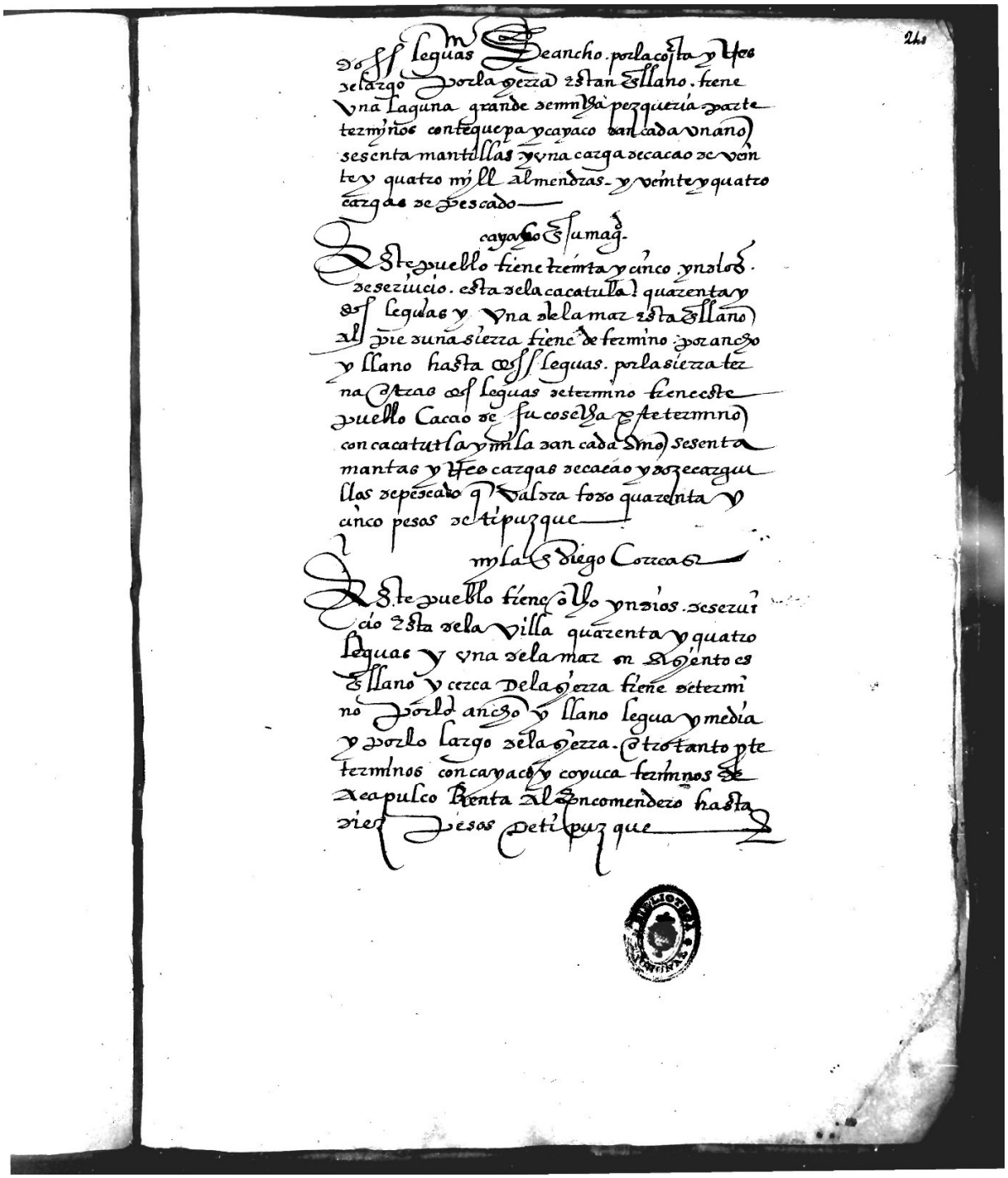


Figura 7.17. Muestra de documento del corpus generado. Página 444 en español mexicano de un period entre 1538-1540 (García Castro, 2013).