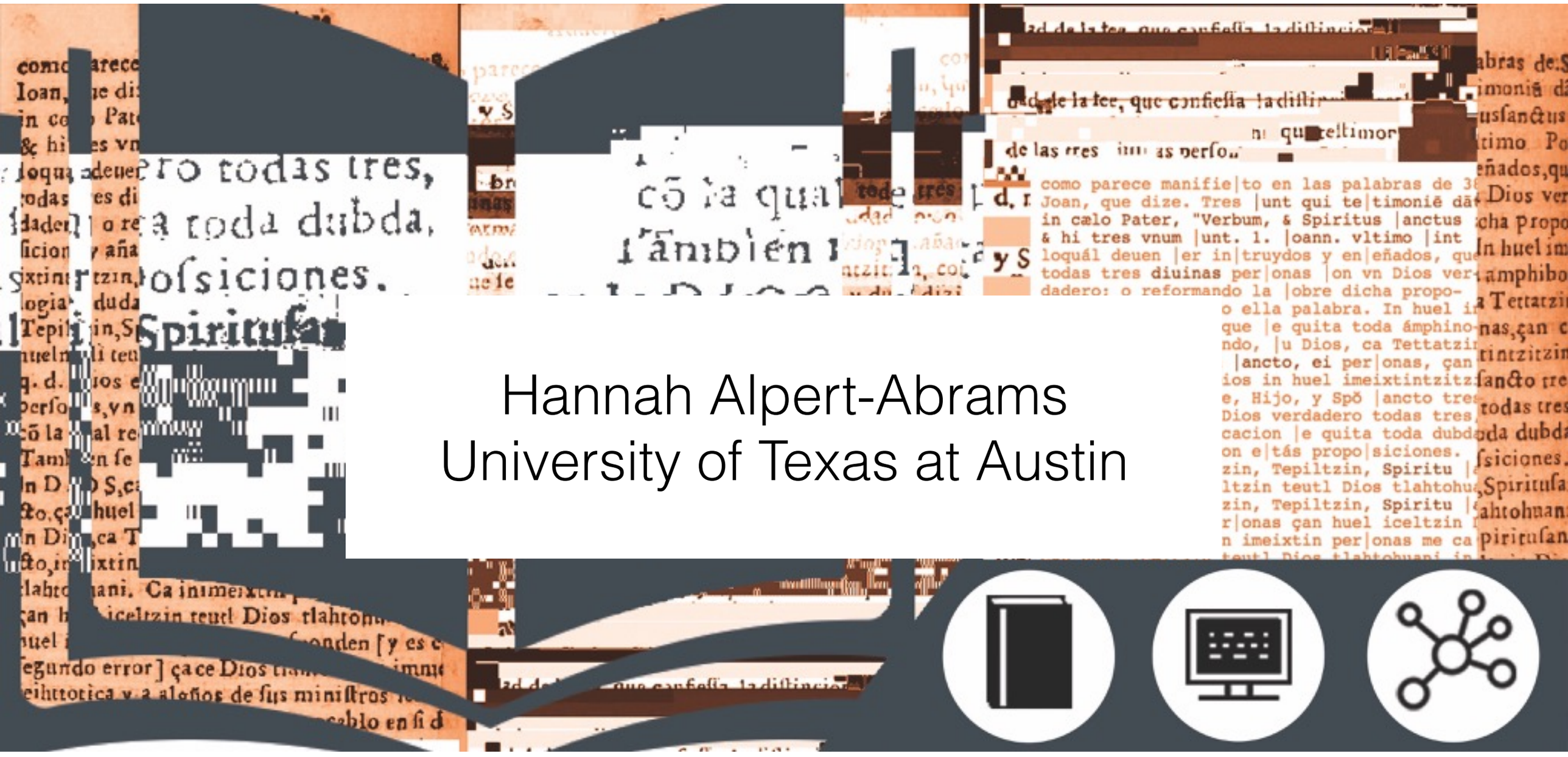
Transcribing Historical Texts



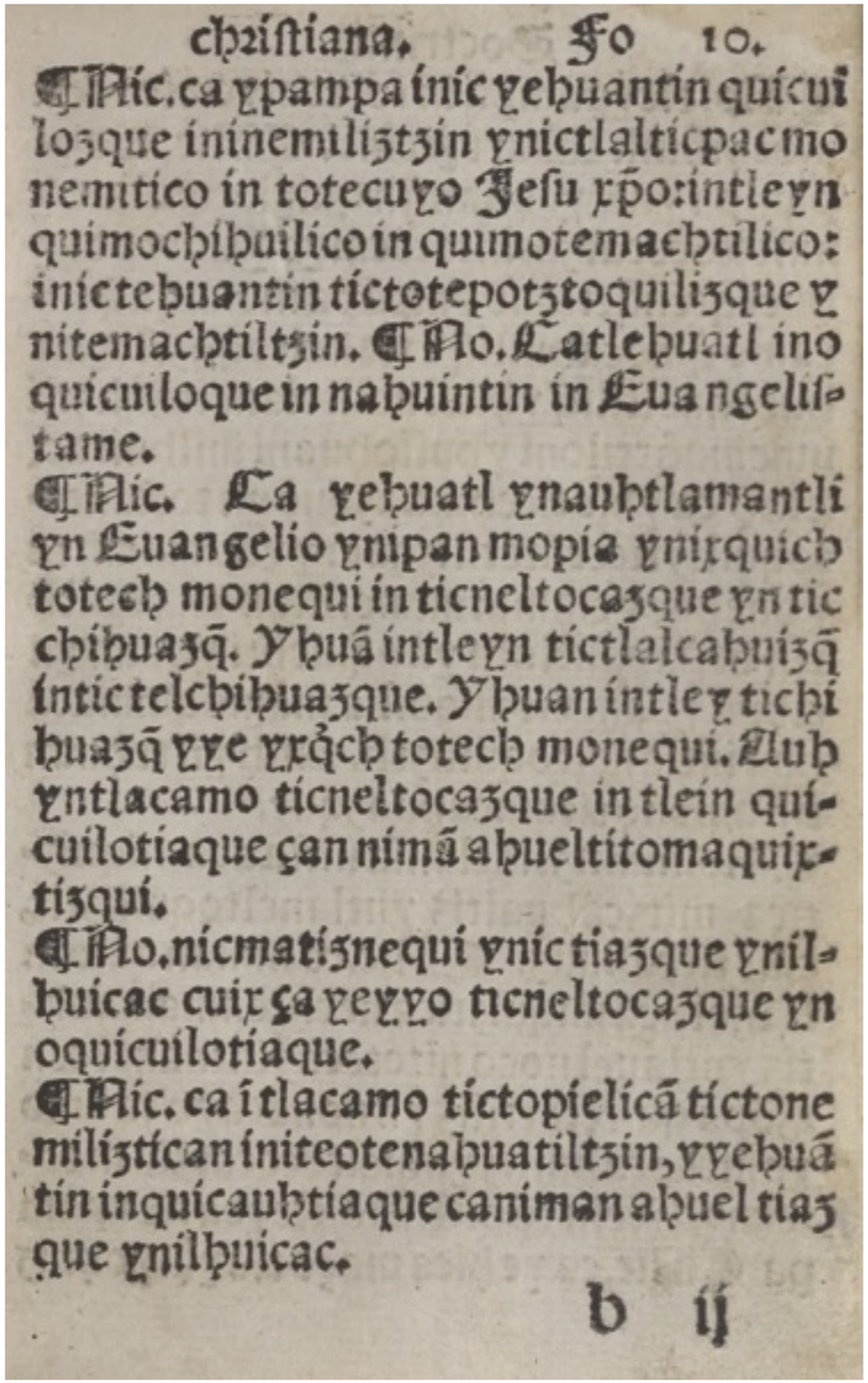
Hannah Alpert-Abrams
University of Texas at Austin

¶ Mic.ca ypampa inic yehuantin quicui lozque ininemiliztzin ynictlalticpac mo nemitico in totecuyo Jesu xpo:intleyn quimochihuilico in quimotemachtilico: inictehuantin tictotepotztoquilizque y nitemachtiltzin. ¶ No. Catlehuatl ino quicuiloque in nahuintin in Euangelis-tame.

¶ Mic. Ca yehuatl ynauhtlamantli yn Euangelio ynipan mopia ynixquich totech monequi in ticneltocazque yn tic chihuazq̃. Yhuã intleyn tictlalcahuizq̃ intictelchihuazque. Yhuan intley tichi huazq̃ yye yxc̃h totech monequi. Auh yntlacamo ticneltocazque in tlein qui-cuilotiaque çan nimã ahueltitomaquix-tizqui.

¶ No. nicmatiznequi ynic tiazque ynil-huicac cuix ça yeyyo ticneltocazque yn oquicuilotiaque.

¶ Mic. ca itlacamo tictopielicã tictone miliztican initeotenahuatiltzin, yyehuã tin inquicauhtiaque canimã ahuel tiaz que ynilhuicac.



christiana. Fo 10.

¶Nic. ca ypampa inic yehuantin quicui lozque ininemiliztzin ynictlalticpacmo nemitico in totecuyo Jesu xp̄o: intleyn quimochihuilico in quimotemachtilico: inictehuantin tictotepotztoquilizque y nitemachtiltzin. ¶No. Catlehuatl ino quicuiloque in nahuintin in Euangelis-tame.

¶Nic. Ca yehuatl ynauhtlamantli yn Euangelio ynipan mopia ynixquich totech monequi in ticneltocazque yn tic chihuazq̄. Yhuā intleyn tictlalcahuizq̄ intictelchihuazque. Yhuan intley tichi huazq̄ yye yxq̇ch totech monequi. Auh yntlacamo ticneltocazque in tlein qui-cuilotiaque çan nimā ahueltitomaquix-tizqui.

¶No. nicmatiznequi ynic tiazque ynil-huicac cuix ça yeyyo ticneltocazque yn oquicuilotiaque.

¶Nic. ca ītlacamo tictopielicā tictone miliztican initeotenahuatiltzin, yyehuā tin inquicauhtiaque caniman ahuel tiaz que ynilhuicac.

b ij

chri|tiana. Fo 10.

¶Nic. ca ypampa inic yehuantin quicui lozque ininemilztzin yn ictlalticpacmo nemitico in totecuyo Je|u x\~po: intleyn quimochihuilico in quimotemachtilico: inictehuantin tictotepotztoquilizque y nitemachtiltzin. ¶No. Catlehuatl ino quicuiloque in nahuintin in Euangeli|-tame.

¶Nic. Ca yehuatl yn auhtlamantli yn Euangelio yn ipan mopia yn ixquich totech monequi in ticneltocazque yn tic chihuaz\~q. Yhu\~a intleyn tictlalcahuiz\~q íntictelchihuazque. Yhuan intley tichi huaz\~q yye yx\'qch totech monequi. Tluh yntlacamo ticneltocazque in tlein qui-cuilotiaque çan nim\~a ahueltitomaquix-tizqui.

¶No. nicmatiznequi yn ic tiazque yn il-huicac cuix ça yeyyo ticneltocazque yn oquicuilotiaque.

¶Nic. ca \~itlacamo tictopielic\~a tictone miliztican initeotenahuatiltzin, yyehu\~a tin inquicauhtiaque caniman ahuel tiaz que yn ilhuicac.

b ij

Accessibility
Discoverability
Preservation
Research

Labor
Cost
Skill
Accuracy
Privacy

Crowdsourcing

Automatic Transcription

Smithsonian

HOME PROJECTS ABOUT TIPS NEWS

SMITHSONIAN DIGITAL VOLUNTEERS: TRANSCRIPTION CENTER

SIGNUP LOGIN

feedback

JOIN US!

LEARN HOW TO TRANSCRIBE

Become a Smithsonian Digital Volunteer and help us make historical documents and biodiversity data more accessble.

Join 8,650 volunteers to add more to the total 287,436 pages of field notes, diaries, ledgers, logbooks, currency proof sheets, photo albums, manuscripts, biodiversity specimens labels that have

BROWSE PROJECTS

Select a category below to begin browsing projects.

Select a Category

DISCOVER THE FREEDMEN'S BUREAU

North Carolina Assistant Commissioner, Registers of Letters

LATEST UPDATES

mandc transcribed a page from Charles Francis Hall Journal with Navigational Notes 1861

sagehen transcribed a page from North Carolina Assistant Commissioner, Endorsements Sent, Vol. 3 (20), Mar. 4-Dec. 30, 1868

Katies marked for review a page from Charles Lang Freer's letters to Frank

https://transcription.si.edu/

Sample Text

Axcan ipan miercoles çempuali macuiltonal metztli enero x-

Created

21-04-2016

Files

Transcribe This Item

1. AGCA A1 Leg6074 Exp 54891-1.JPG
2. AGCA A1 Leg6074 Exp 54891-2.JPG

http://nahuatl-nawat.org/

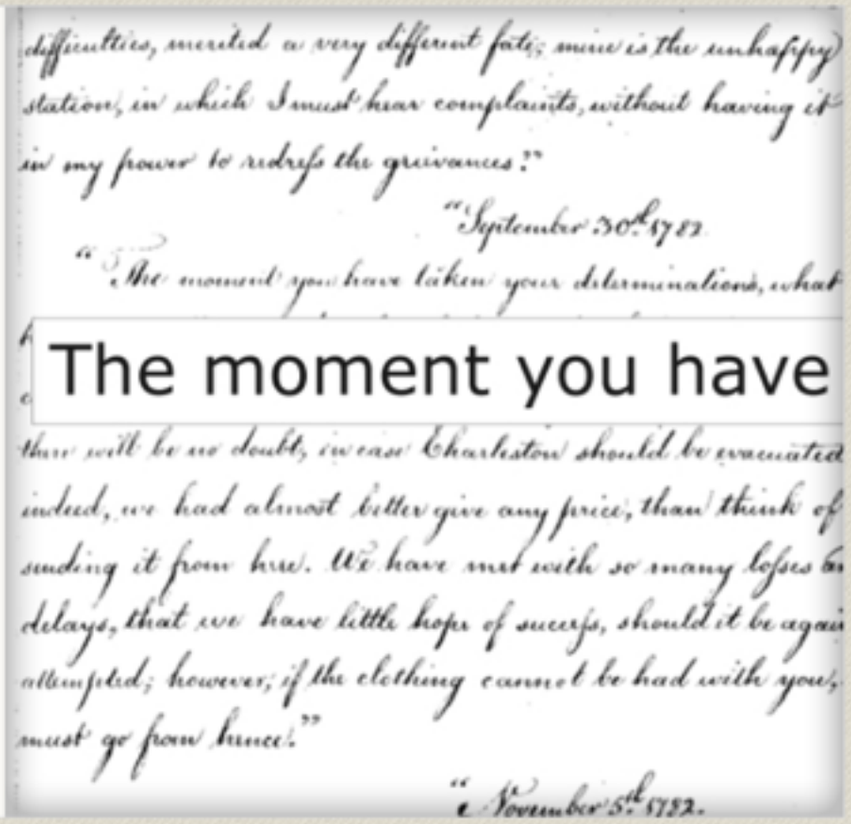
scripto a community transcription tool

Home | Showcase | User's Guide | Download | News | About

A free, open source tool enabling community transcriptions of document and multimedia files

Scripto brings the power of MediaWiki to your collections. Designed to allow members of the public to transcribe a range of different kinds of files, Scripto will increase your content's findability while building your user community through active engagement.

Download | Learn More



http://scripto.org/

Author: *Packe, Susanna, fl. 1674.*

Title: *Cookbook of Susanna Packe [manuscript].*

Catalogue

Original

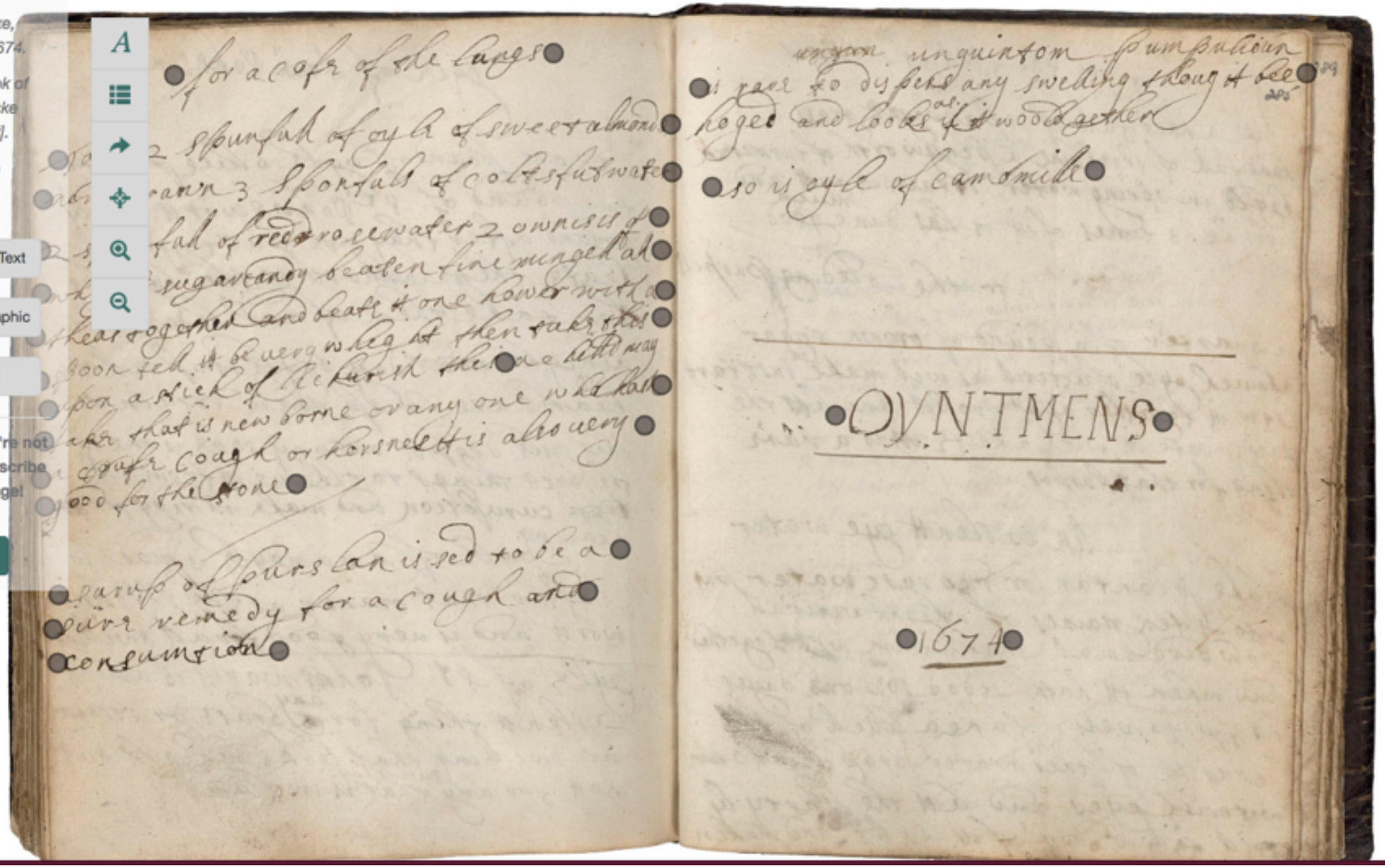
A Transcribe Text

Identify Graphic

Tutorial

Remember, you're not required to transcribe the whole page!

I'm done!



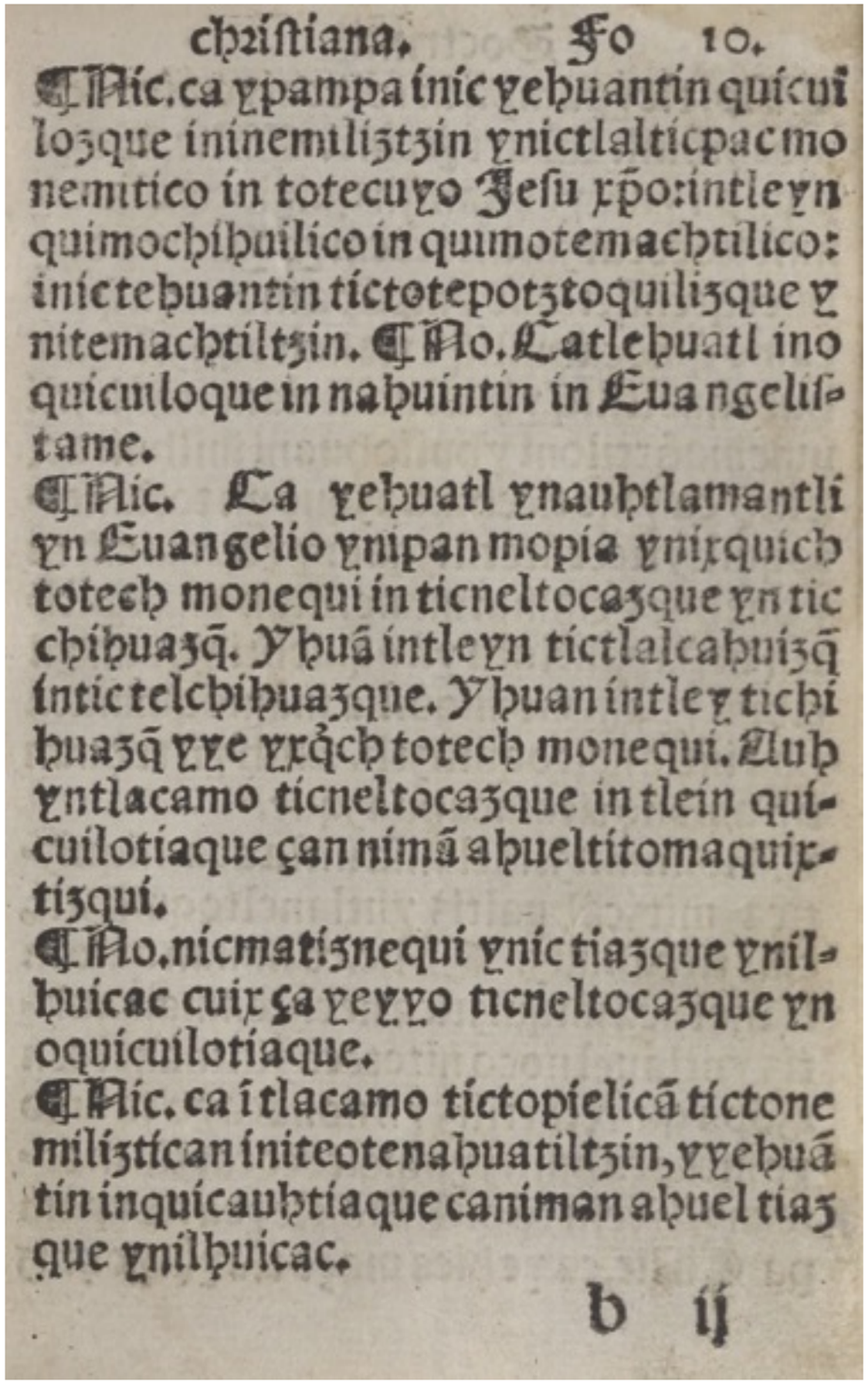
https://www.shakespearesworld.org/

Crowdsourcing

Accurate
Affordable
Needs Volunteers
Needs Infrastructure
Needs Maintenance

Crowdsourcing

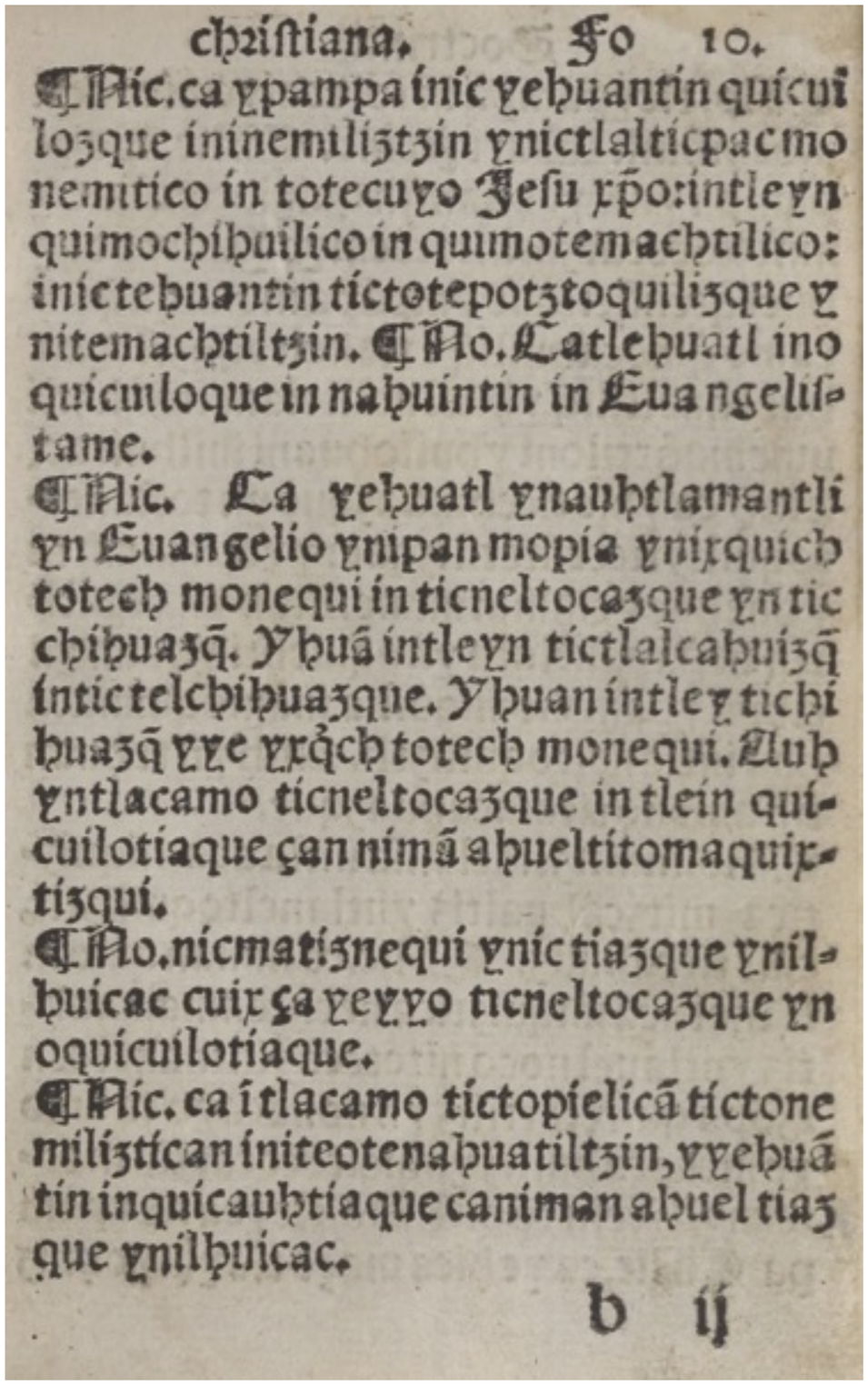
Automatic Transcription



christiana. Fo 10.
¶Nic. ca ypampa inic yehuantin quicui lozque ininemilztzin ynictlalticpacmo nemitico in totecuyo Jesu xpõ: intleyn quimochihuilico in quimotemachtilico: inictehuantin tictotepotztoquilizque y nitemachtiltzin. ¶No. Catlehuatl ino quicuiloque in nahuintin in Euangelis- tame.
¶Nic. Ca yehuatl ynauhtlamantli yn Euangelio ynipan mopia ynixquich totech monequi in ticneltocazque yn tic chihuazq̃. Yhuã intleyn tictlalcahuizq̃ íntictelchihuazque. Yhuan intley tichi huazq̃ yye yxc̃h totech monequi. Tluh yntlacamo ticneltocazque in tlein qui- cuilotiaque çan nimã ahueltitomaquix- tizqui.
¶No. nicmatiznequi ynic tiazque ynil- huicac cuix ça yeyyo ticneltocazque yn oquicuilotiaque.
¶Nic. ca itlacamo tictopielicã tictone miliztican initeotenahuatiltzin, yyehuã tin inquicauhtiaque caniman ahuel tiaz que ynilhuicac.
b ij

chri|tiana. Fo 10.
¶Nic. ca ypampa inic yehuantin quicui lozque ininemilztzin yn ictlalticpacmo nemitico in totecuyo Je|u x\~po: intleyn quimochihuilico in quimotemachtilico: inictehuantin tictotepotztoquilizque y nitemachtiltzin. ¶No. Catlehuatl ino quicuiloque in nahuintin in Euangeli|- tame.
¶Nic. Ca yehuatl yn auhtlamantli yn Euangelio yn ipan mopia yn ixquich totech monequi in ticneltocazque yn tic chihuaz\~q. Yhu\~a intleyn tictlalcahuiz\~q íntictelchihuazque. Yhuan intley tichi huaz\~q yye yx\'qch totech monequi. Tluh yntlacamo ticneltocazque in tlein qui- cuilotiaque çan nim\~a ahueltitomaquix- tizqui.
¶No. nicmatiznequi yn ic tiazque yn il- huicac cuix ça yeyyo ticneltocazque yn oquicuilotiaque.
¶Nic. ca \~itlacamo tictopielic\~a tictone miliztican initeotenahuatiltzin, yyehu\~a tin inquicauhtiaque caniman ahuel tiaz que yn ilhuicac.
b ij

Optical Character Recognition | OCR



christiana. Fo 10.

¶ Nic.ca ypampa inic yehuantin quicui lozque ininemilizt zin ynictlalticpac mo nemitico in totecuyo Jesu xpo:intleyn quimochihuilico in quimotemachtilico: inictehuantin tictotepotztoquilizque y nitemachtiltzin. ¶ No. Catlehuatl ino quicuiloque in nahuintin in Euangelis tame.

¶ Nic. Ca yehuatl ynauhtlamantli yn Euangelio ynipan mopia ynixquich totech monequi in ticneltocazque yn tic chihuazq̃. Yhuã intleyn tictlalcahuizq̃ intictelchihuazque. Yhuan intley tichi huazq̃ yye yxc̃h totech monequi. Auh yntlacamo ticneltocazque in tlein qui cuilotiaque çan nimã ahueltitomaquix tizqui.

¶ No. nicmatiznequi ynic tiazque ynil huicac cuix ça yeyyo ticneltocazque yn oquicuilotiaque.

¶ Nic. ca itlacamo tictopielicã tictone milizticanin iteotenahuatiltzin, yyehuã tin inquicauhtiaque canimanahuel tiaz que ynilhuicac.

b ij

o — M r^ -^ v^SO |>»00 Cv O — «
r^ -«t- «^\o r-^oo Oso — r^
vo SO vO \0 vO *0 vO vO nO so \0
sOvO vososOnOvOvOvosOvOvO

00 (7s O — «*< '◆N ^ w^sO r-»00

(7s o «- n r^ 't- «^VO f>.00 Cv o
w^ w^vO sovovosovovovo>ovo r**
r^r^j^j^c^r>.c^r>. r**oo

VO r**00 Cv o — *^ «^ -^ «^SO
t>i30 J\ o — r«

o — M r^ -^ v^SO |>»00 Cv O — «
r^ -«t- «^\o r-^oo Oso — r^
vo SO vO \0 vO *0 vO vO nO so \0
sOvO vososOnOvOvOvosOvOvO

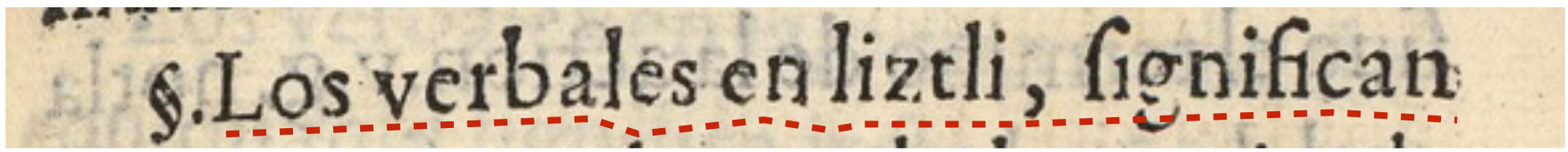
00 (7s O — «*< '◆N ^ w^sO r-»00

(7s o «- n r^ 't-

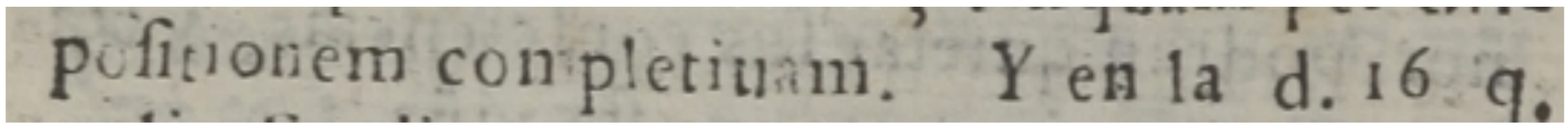
Optical Character Recognition | OCR

How does automatic transcription work?

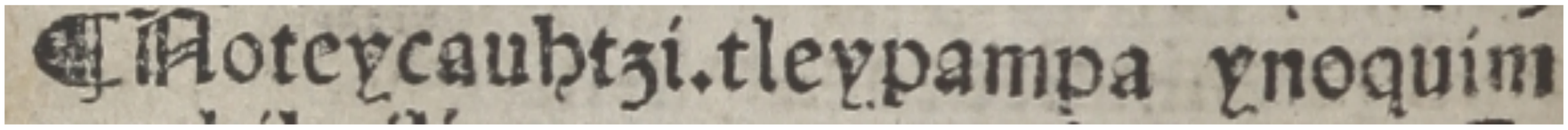
Wandering baseline

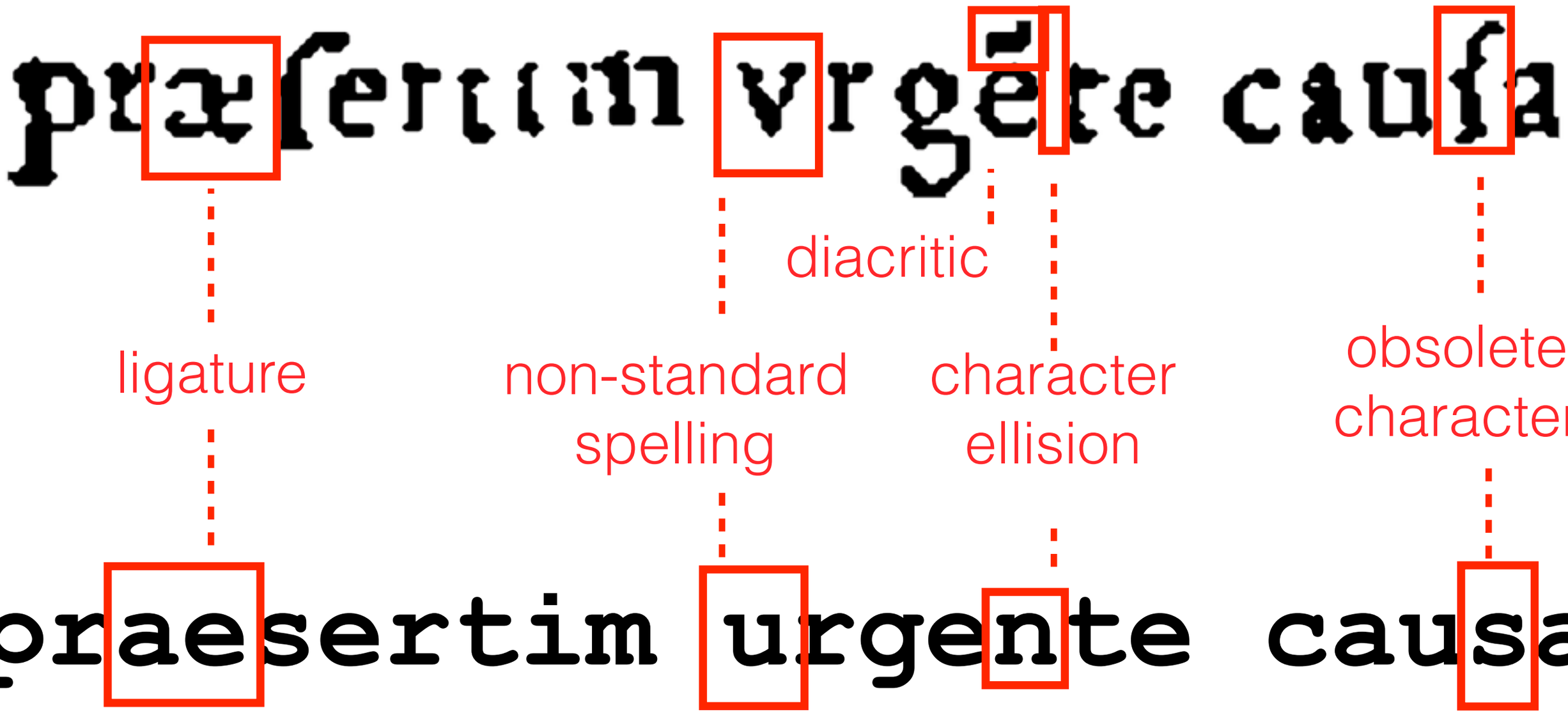


Uneven inking



Unfamiliar Typefaces




præſertim vrgẽte cauſa
ligature
diacritic
non-standard spelling
character ellision
obsolete character
praesertim urgente causa

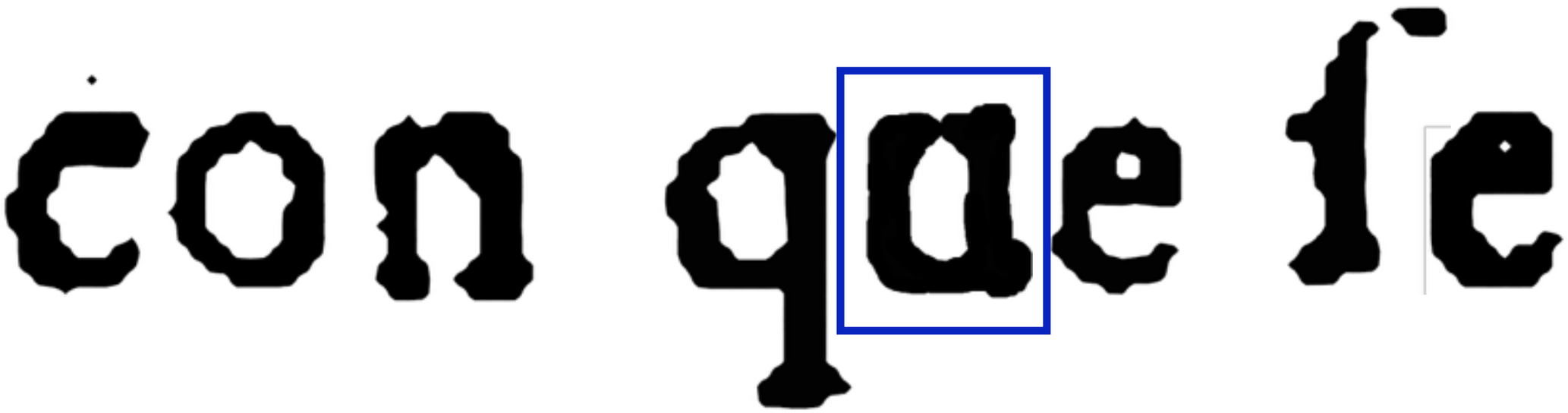
Font Model

+

Language Model

Font Model

Font Model



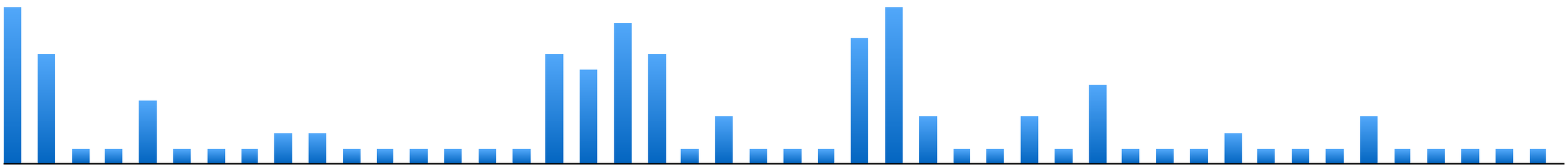
Font Model

Font Model

Font Model

Font Model

Font Model



a á b c d e é f g h i í j k l m n ñ o ó p q r s t u ú v w x y z 0 1 2 3 4 5 6 7 8 9 . , : ;

Language Model

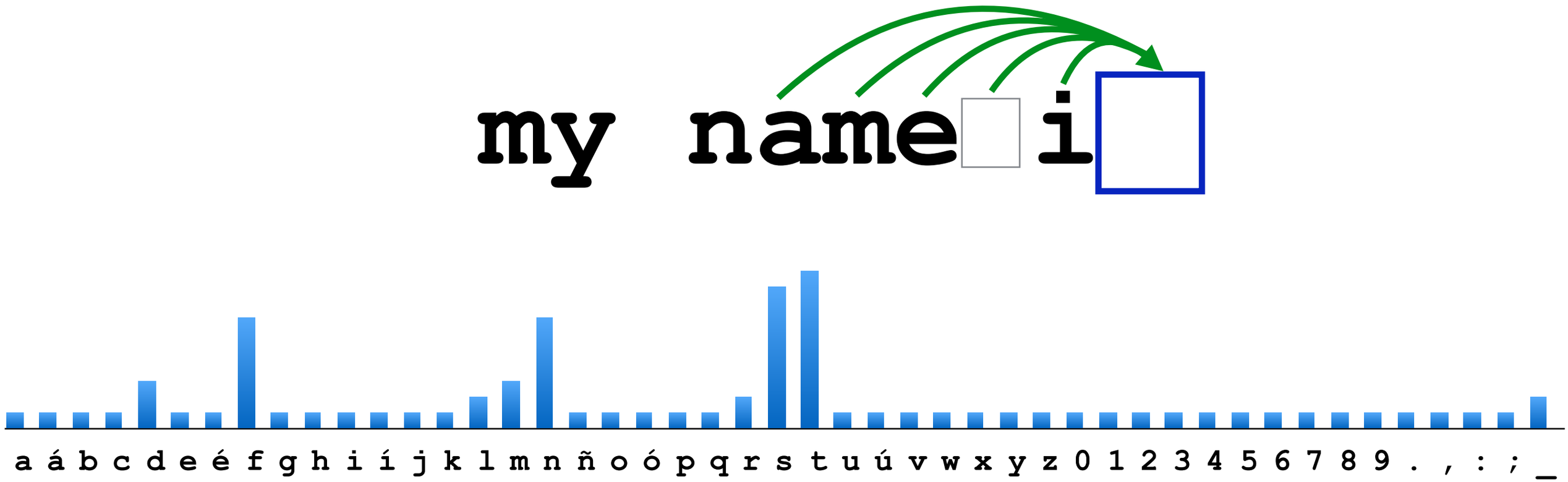
```
my name i
```

Language Model

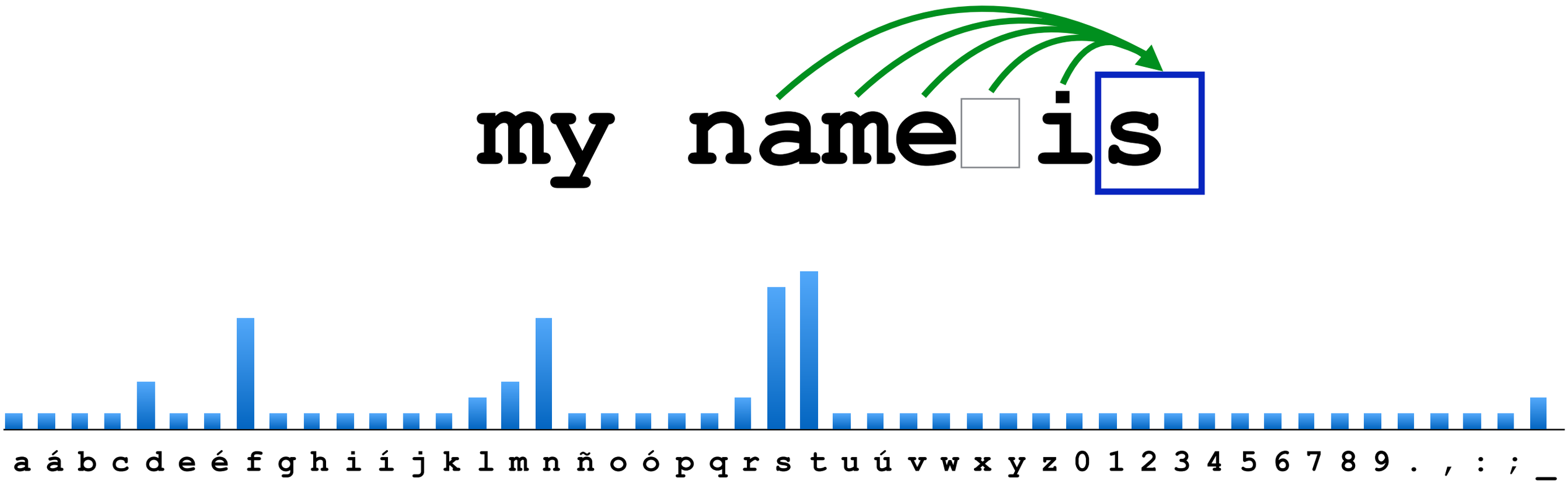
my name i

Language Model

Language Model

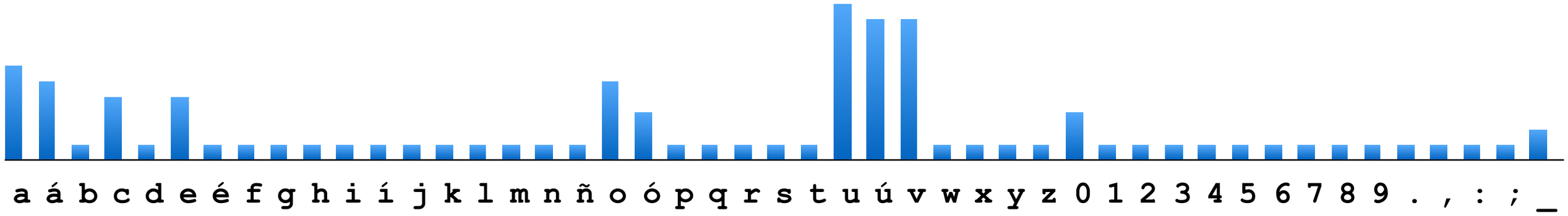


Language Model



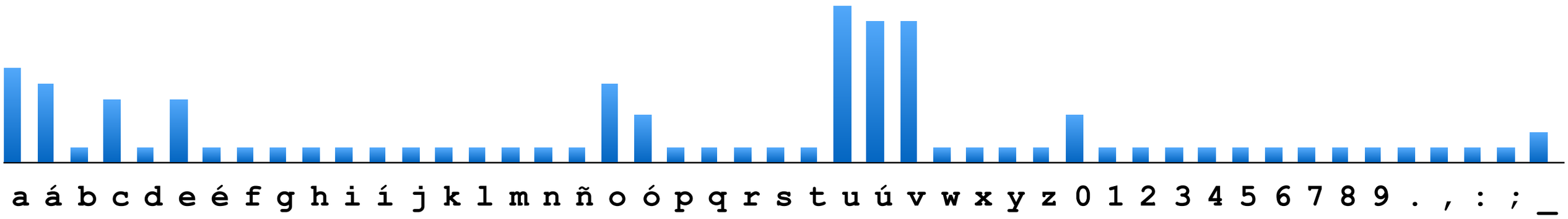
Font Model

Font Model

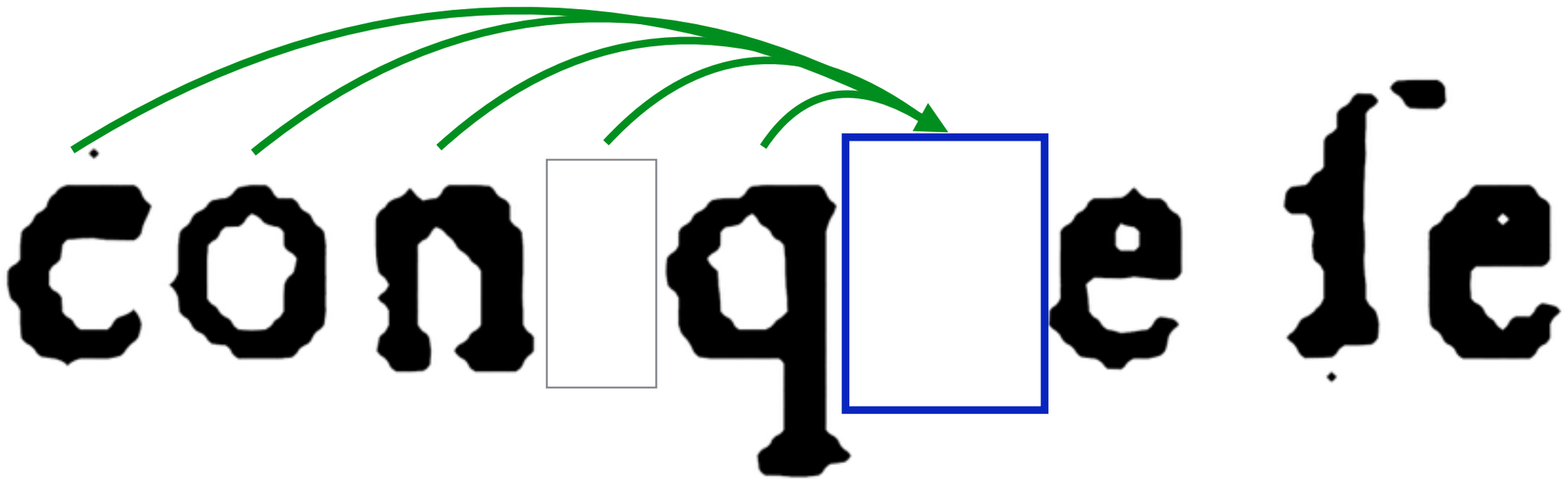


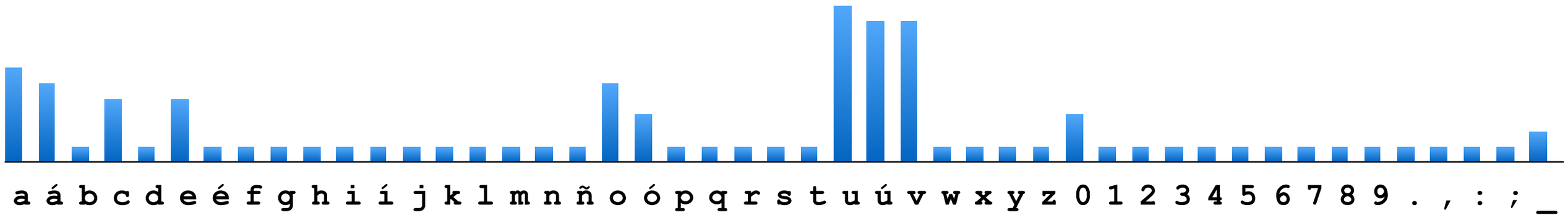
Language Model

con que le



Language Model





Language Model

con que fe

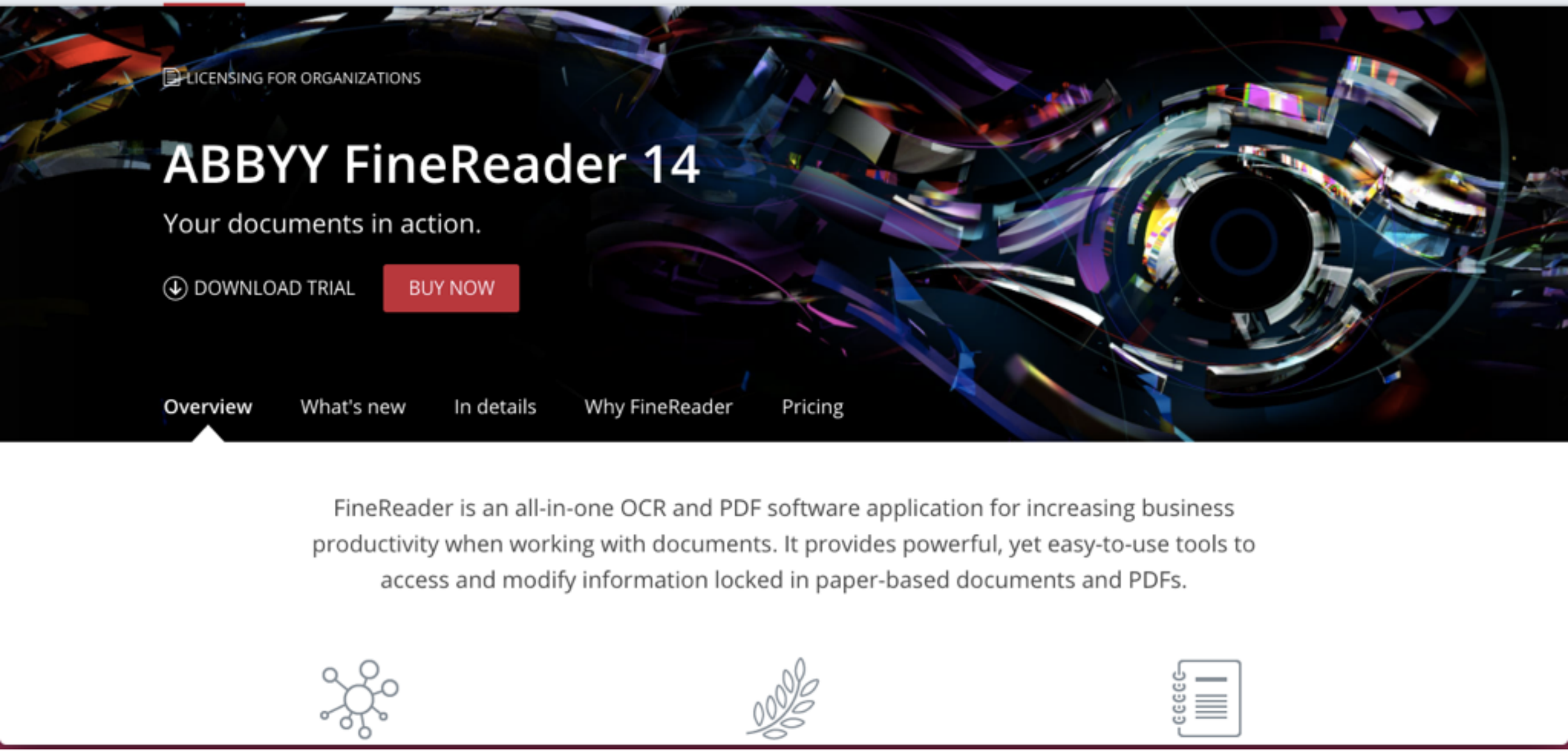
How does automatic transcription work?

Options for Automatic Transcription

ABBYY

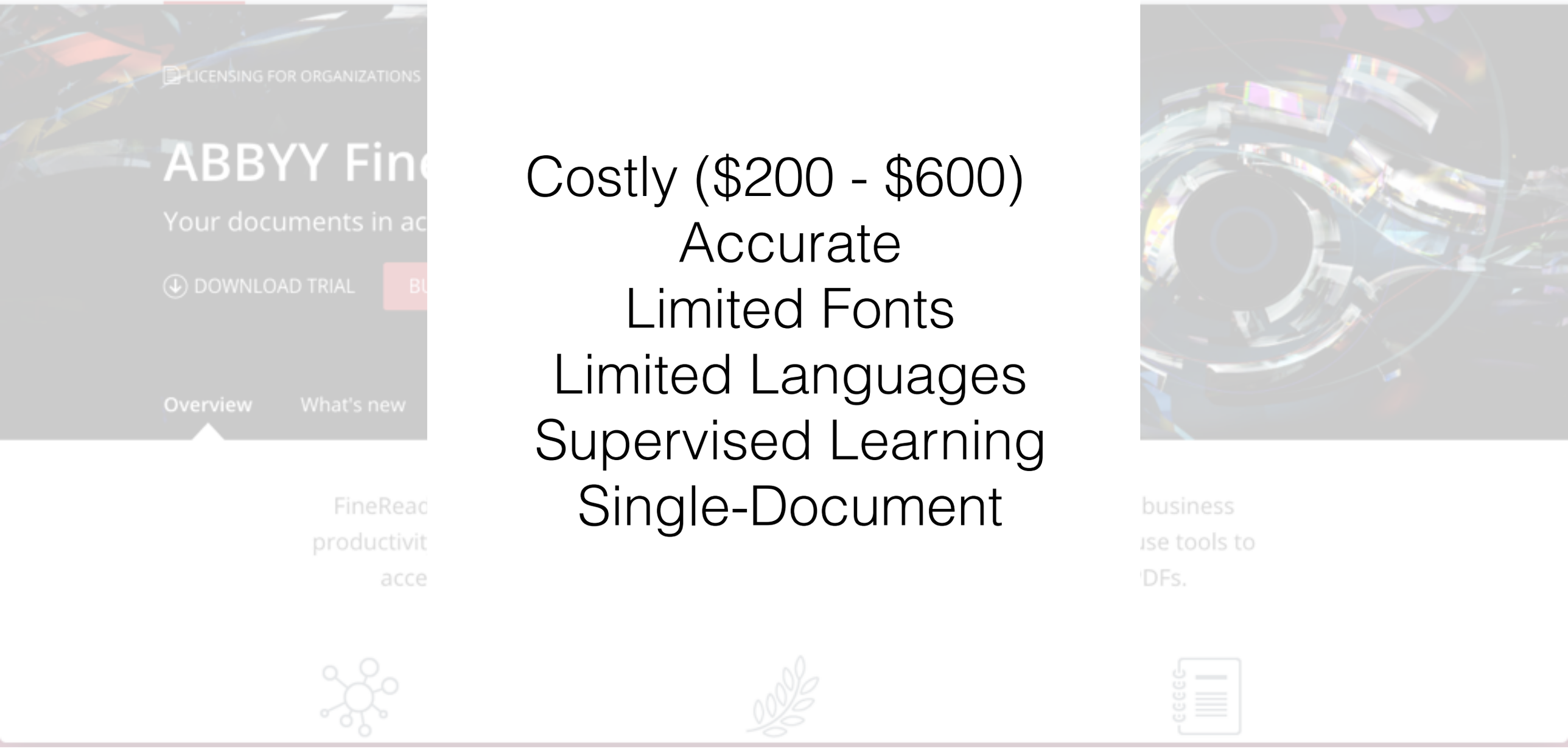
CONTACTS STORE SELECT REGION

Products Solutions Developer Tools Partners Support Company



FineReader is an all-in-one OCR and PDF software application for increasing business productivity when working with documents. It provides powerful, yet easy-to-use tools to access and modify information locked in paper-based documents and PDFs.

ABBYY Finereader



Costly ($200 - $600)
Accurate
Limited Fonts
Limited Languages
Supervised Learning
Single-Document

ABBYY Finereader

This organization | Search | Pull requests | Issues | Gist

tesseract-ocr

Repositories | People 0

Search repositories... | Type: All | Language: All

tesseract

Tesseract Open Source OCR Engine (main repository)

machine-learning | ocr | tesseract | lstm | tesseract-ocr

C++ ★ 9,893 2,277 Updated 3 hours ago

langdata

Source training data for Tesseract for lots of languages

★ 243 373 Updated on Feb 21

tessdata

★ 524 316 Updated on Dec 28, 2016

docs

Top languages

C++ HTML

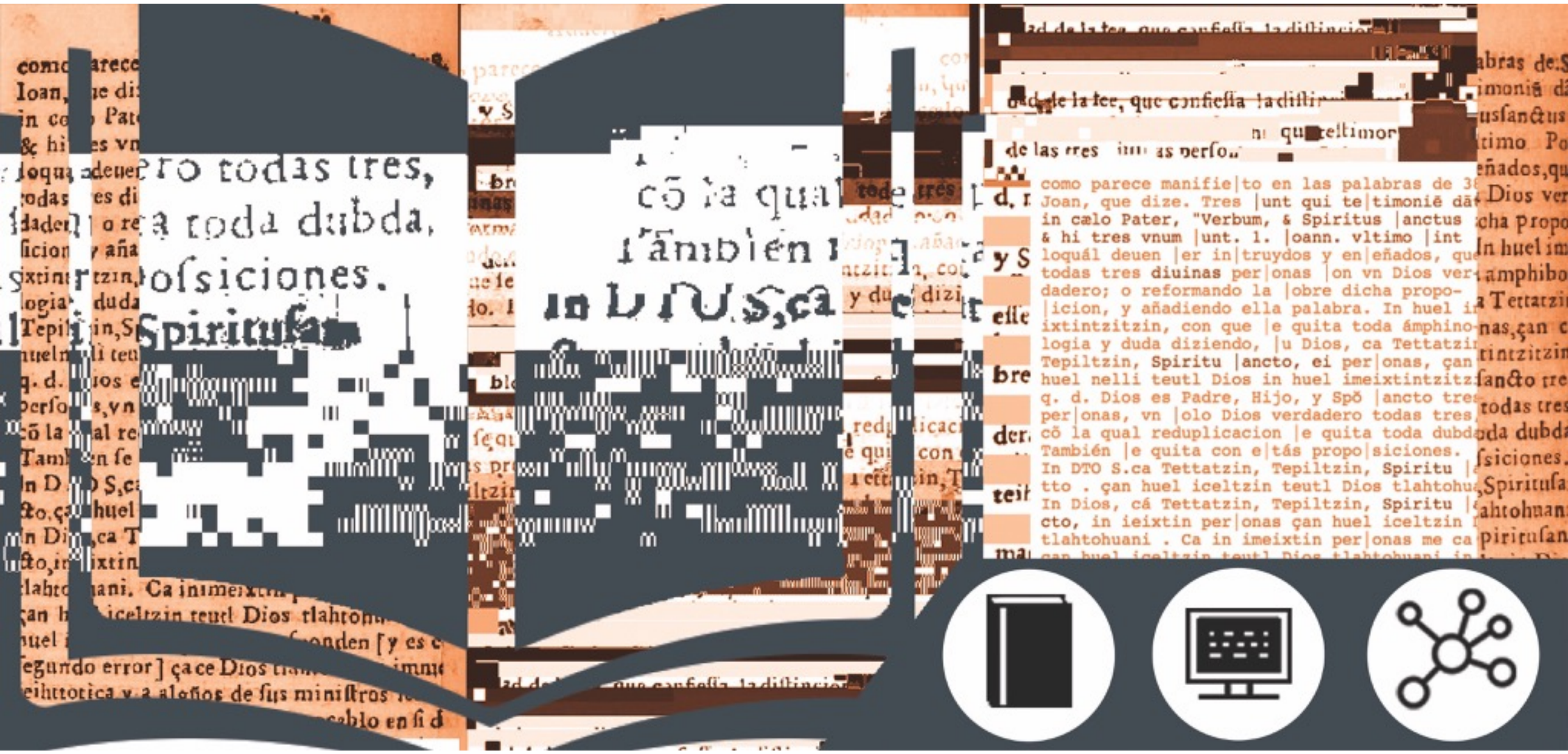
People 0 >

This organization has no public members. You must be a member to see who's a part of this organization.

Tesseract

Open-Access
Limited Fonts
Limited Languages
Modern Orientation
Command-Line Operations

Tesseract


como parece manifie|to en las palabras de 3
Joan, que dize. Tres |unt qui te|timonié dã
in cælo Pater, "Verbum, & Spiritus |anctus
& hi tres vnum |unt. 1. |oann. vltimo |int
loquál deuen |er in|truydos y en|eñados, qu
todas tres diuinas per|onas |on vn Dios ver
dadero; o reformando la |obre dicha propo-
|icion, y añadiendo ella palabra. In huel i
ixtintzitzin, con que |e quita toda ámphino-
logia y duda diziendo, |u Dios, ca Tettatzi
Tepiltzin, Spiritu |ancto, ei per|onas, çan
huel nelli teutl Dios in huel imeixtintzitz
q. d. Dios es Padre, Hijo, y Spŏ |ancto tres
per|onas, vn |olo Dios verdadero todas tres
cõ la qual reduplicacion |e quita toda dubda
También |e quita con e|tás propo|siciones.
In DTO S.ca Tettatzin, Tepiltzin, Spiritu |
tto . çan huel iceltzin teutl Dios tlahtohu
In Dios, cá Tettatzin, Tepiltzin, Spiritu |
cto, in ieixtin per|onas çan huel iceltzin
tlahtohuani . Ca in imeixtin per|onas me ca

Ocular

Designed for historical documents
Handles multiple languages
Handles nonstandard orthographies
Automated normalization
Unsupervised learning

Ocular

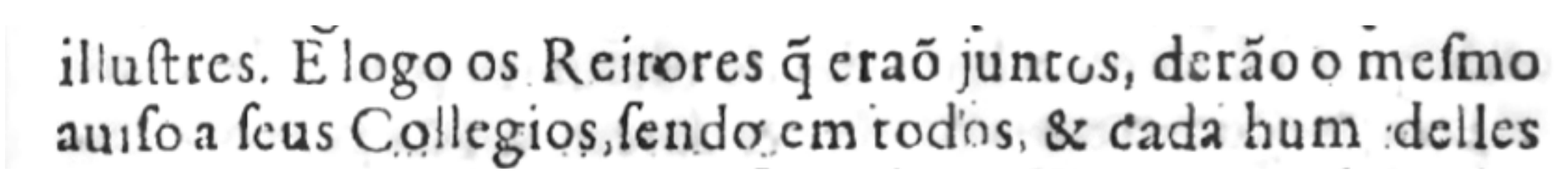
"¶ Q¿nid dicendum de Ocnamacaque. f. de
las pu|queras, o taberneros, qui vendunt vi-
num indorum, indis, quod dicitur Octli .&.
etiam de iis qui eis venduntvr nõ verum Hi|

¶ Quid dicendum de Ocnamacaque. ſ. de
las pulqueras, o taberneros, qui vendunt vi- 27
num Indorum, Indis, quod dicitur Octli :&
etiam de ijs qui eis vendunt vinũ verum Hiſ

Ocular

Normalized text: illustres. E logo os Reitores que taõ juntos, derão o mesmo
Historical text: illustres. E logo os Reitores \~qe taõ juntos, derão o meſmo

Normalized text: aviso a seus Collegios, sendo em todos, & cada hum delles
Historical text: aviſo a ſeus Collegios, ſendo em todos, & cada hum delles


illuſtres. Ẽ logo os Reitores q̃ eraõ juntos, derão o meſmo
auiſo a ſeus Collegios, ſendo em todos, & cada hum delles

Ocular

Technologically complex
Command-line operations
Single-document transcription

Ocular

[video using Ocular]

https://youtu.be/Xgys4dBbDfI

---

https://github.com/tberg12/ocular