Translating, Transcribing and Summarizing Documents Using AI

Thomas MacEntee, of Genealogy Bargains
https://genealogybargains.com hidefgen@gmail.com

The power of artificial intelligence can be used to assist genealogists in translating, transcribing, and summarizing a variety of genealogy records. A handwritten baptismal record from the 1800s? No problem. A faded newspaper obituary? No problem. Most AI platforms can help generate useful content for genealogical research.

What is Artificial Intelligence and How Does it Work?

AI stands for Artificial Intelligence and represents computer-based systems that can “mimic” human intelligence. The goal is to have these systems perform human tasks.

The most discussed features of artificial intelligence are “deep learning” and “generative AI.” Deep learning mimics the human brain in that it looks for patterns using vast amounts of information to interpret photos, audio, and text. Generative AI actually “generates” new photos, audio, and text, based on information provided by the user, and again, uses its own database of “training data” to understand patterns and generate output that matches the user’s query.

AI Platforms for Translating, Transcribing and Summarizing Documents

While genealogy vendors are incorporating AI into features provided to users, there are some popular AI platforms open to the public that you might want to consider using.

- **ChatGPT:** Meaning “Chat Generative Pre-trained Transformer,” ChatGPT (https://chat.openai.com/) is the most popular publicly-accessible artificial intelligence platform.

- **Claude:** Claude (https://claude.ai/) is an artificial intelligence chatbot created by the company Anthropic that is designed to generate text content and engage in conversations with users using human-like responses.

- **Copilot:** Copilot (https://copilot.microsoft.com) is an AI-powered intelligent assistant that helps you get answers and inspirations from across the web, supports creativity and collaboration, and helps you focus on the task at hand.

- **Gemini:** Developed by Google, Gemini (https://gemini.google.com/) formerly known as Bard, describes itself as “a family of AI models developed by Google's AI research labs DeepMind and Google Research. Gemini is Google's largest and most flexible AI model, able to run on data centers and mobile devices.”

Summarization

Ever find yourself staring down a massive text—maybe a thick local history volume, an old genealogy journal, a lengthy family narrative, or even a centuries-old account of your ancestors' homeland—wondering how you'll uncover the key details hidden inside? Let's face it: reading through hundreds of pages can drain your time and energy. This is where artificial intelligence steps in with a game-changing advantage. AI-powered summarization tools can quickly zero in on the names, themes, and must-have tidbits, letting you skip straight to the parts that matter most to your research.

How To

- Start a new chat on the AI platform.

- Upload the document to be summarized. Best formats are PDF or Microsoft Word. NOTE: some FREE versions of AI platforms may limit your ability to upload documents.

- Enter your summary prompt. Example: *"Create a summary of this book The Descendants of David Putman published in 1916 extracting family groups and tracing migration routes and patterns by generation."*

- Review generated text and refine summary using modifying language.

Providing Contextual Prompts

AI models perform better when given context. When requesting a summary, explain the nature of the document, its time period, and why it matters in the prompt. For example:

> *"Please summarize This local history book about rural New York in the late 19th century. I am looking for information related to the Crawford family, their property holdings, and any mention of their participation in the community church."*

Results will vary based on which AI platform you use and if you use the FREE or PAID version. Also consider adding "style modifiers" such as "use bullet point style".

Iterative Refinement

If the initial summary feels too vague or misses key details, refine your request. Add instructions like:

> *"Please revise the summary to highlight any mention of the Crawford family. Focus on names, dates, and property transactions between 1850 and 1900."*

This iterative process helps guide the AI toward more relevant and accurate outputs.

Transcription

Transcription is one of those essential skills that can make a world of difference in your genealogy research. It turns those tough-to-read documents—faded newspaper clippings, typed census sheets, scribbled diary pages, and old family letters—into text you can actually work with. Each record type has its own quirks: early newspapers might have ink that's barely there and columns that twist and turn, vital records sometimes mix tidy printed text with scrawled annotations, and those personal letters you inherited often feature challenging handwriting and outdated language that'll send you flipping through old dictionaries.

Thanks to recent advances in artificial intelligence, we're now seeing tools that can handle a lot of that tricky transcription work for us. Instead of squinting at spidery handwriting and guessing at every other word, you can rely on AI-driven OCR (optical character recognition) and handwriting recognition to do much of the heavy lifting. These cutting-edge models are trained on massive sets of historical texts and fonts, which means less time straining your eyes and more time putting that data to use in your family history projects.

Types of Documents Commonly Transcribed

- **Newspaper Articles and Obituaries:** Historical newspapers are treasure troves of genealogical data—local news, birth and death announcements, marriage notices, and social columns can reveal personal details about ancestors. After using OCR, you can transform these clippings into text that is easily searchable and can be summarized or analyzed using AI tools.
- **Vital Records (Birth, Marriage, Death Certificates):** Vital records often present a mix of printed headings and handwritten entries. OCR can handle the standardized printed parts, while specialized handwriting recognition tools (like Transkribus) interpret handwritten names, dates, and places. Once transcribed, these records can be incorporated into your genealogical database, making it simpler to connect individuals and trace lineages.
- **Census Records and Immigration Forms:** Census documents and passenger lists are often filled out by various clerks with varying handwriting styles. Handwriting recognition can help convert these complex tabular documents into structured text. With the resulting transcription, you can easily search for family members or analyze demographic trends.
- **Personal Letters, Diaries, and Family Correspondence:** These sources are highly valuable but often the most challenging to decipher. AI-powered transcription can handle much of the leg work, allowing you to focus on interpreting the meaning of the text, the relationships mentioned, and the social or emotional context revealed in private writings.

How To

- **Digitize Your Source:** Start by creating high-quality digital images or scans of the documents. Aim for a high-resolution format (e.g., 300 dpi or higher) and ensure the image is as clear, well-lit, and straight as possible. Correct skewed pages and remove backgrounds or stains if you can, as clear inputs yield better transcription results.
- **Choose the Appropriate Tool**
  - **For Printed Text:** ChatGPT, Claude, CoPilot, and Google Gemini handle modern or clearly printed documents well.
  - **For Historical Handwriting:** ChatGPT, Claude, CoPilot, and Google Gemini had all proven successful with transcribing handwriting. The PAID version of each platform often does a better job than the free version. NOTE: before paying to upgrade an AI platform, test the FREE version of all major AI platforms since results may vary from platform to platform.
- **Pre-Processing the Image:** Improve image quality by adjusting contrast, brightness, and orientation. Remove any extraneous borders or watermarks that might confuse the OCR engine. For handwriting, ensure that the text lines are as straight and isolated as possible.
- **Run the Transcription:** Upload your images to the chosen platform and run the transcription. For large volumes, consider batching the process. Many tools allow you to review transcribed text side-by-side with the original image, making it easier to identify errors.

  Use prompts that help place the document in context such as *"Transcribe this handwritten letter in English dated 1821 – an affidavit in support of an American Revolutionary War Pension Application for Jacob Hoff."*
- **Post-Processing and Correction:** Even the most advanced AI model will make mistakes. Review the resulting transcription for accuracy, focusing on surnames, place names, and old-fashioned terms. Correct errors directly in the tool if it allows, or in your preferred text editor afterward. This iterative correction process not only improves the current transcript but can, in some systems, train the model to perform better on future documents.

Translation

As you delve deeper into your family history, it's almost guaranteed you'll bump into documents that aren't in your native language. Our ancestors often crossed borders—both real and cultural—leaving behind a paper trail recorded in Latin from church registers, German in Austro-Hungarian civil documents, Italian parish books, Cyrillic-script manifests for Eastern European immigrants, or the kind of old-time English that reads more like a riddle than a record. Without a proper translation, these rich sources remain locked away, their hidden stories untold.

This is where artificial intelligence can lend a much-needed helping hand. Today's translation tools are more advanced than ever, and when you guide them with the right context and historical insight, they can break down those language barriers and give you the keys to your ancestors' lives. But keep in mind—old documents don't always play by the rules. Obsolete terms, unusual spellings, unique handwriting styles, and local dialects can throw a wrench into the works of a generic translation engine.

The Importance of Contextualizing Historical Documents

- **Time Period and Region:** Language evolves over time. Terms that were common in the 17th century may no longer be in use, or their meanings may have shifted. Specifying the era and region of the document (e.g., "a German-language marriage record from 1870 in rural Austria") helps AI tools focus on historically appropriate vocabulary and spelling variants.

- **Document Type:** Different types of documents use distinct language registers and terminologies. A birth certificate from early 20th-century Italy may have standardized legal terminology, while a personal letter from the 18th century may feature colloquialisms or family-specific nicknames. Identifying the nature of the document—church record, civil register, probate record, personal letter—guides the AI to expect certain patterns and vocabulary sets.

- **Cultural and Religious Context:** Religious documents, such as Latin baptismal or marriage entries, might include abbreviations and phrases derived from liturgical texts. Understanding the religious or cultural context encourages the AI to interpret terms according to their appropriate cultural and historical meanings.

- **Known Family Data:** If you know that your ancestors lived in a particular village, had certain surnames, or were associated with certain professions, share these details upfront. For example: "I am translating a Polish marriage record from the late 19th century. The family's surname is Kowalski, and they lived in the Mazovia region." This information helps the AI correctly identify key personal and place names.

How To

- **OCR/HTR Process:** Use ChatGPT to transcribe a handwritten German church register entry.
- **Contextual Prompting:** Instruct the AI translator: *“This is an 1870 marriage record from Lower Austria, written in German. Please translate it into English. The groom’s surname is ‘Müller’ and the bride’s surname is ‘Schmidt.’ Focus on accuracy of names, occupations, and any mentions of parental details.”*
- **Refine and Verify:** Check the translated text. If certain words seem off, ask the AI to clarify: “What does the term ‘Leineweber’ mean in this historical context?” The AI can explain that it means “linen weaver,” a common occupation at the time. Also, consider copying the transcribed text and ask the AI platform to translate it back to the original language. Compare the two versions.

Translating Various Document Types

Legal Documents and Affidavits

Documents like wills, property deeds, or affidavits may use formal, legalistic language and standardized phrases. Providing this context helps the AI predict the tone and vocabulary. For example:

> *“This is a 19th-century Spanish will, written by a landowner in rural Andalusia. Please translate it into English and note any mentions of property boundaries, heirs, or family relationships.”*

Correspondence and Diaries

Personal letters and diaries often include colloquial expressions, emotional tone, and unique nicknames. Make sure the AI knows it’s a personal communication:

> *“This is a personal letter from 1890 in French, likely between two family members discussing a recent wedding. Please translate and, where necessary, explain idiomatic phrases that no longer have a direct modern equivalent.”*

Affidavits and Immigration Papers

For immigration documents, explaining the nature of the text is crucial.

> *“This is an early 20th-century Russian immigration manifest listing passengers traveling to the United States. Translate the details about each passenger’s name, place of origin, and occupation. The text uses older Russian spellings, so pay attention to archaic forms.”*

Improving Accuracy

Accuracy is dependent upon a variety of factors:

- **Document/Image Quality:** If the image or document is faded or damaged, consider using a photo editor to improve the quality.
- **Document/Image Type:** Not every platform will be able to handle PDF documents. Try using JPG, PNG, or other formats.
- **Document Length/Number of Images:** For large documents or a large number of images, try loading only a few at a time, in batches.
- **Print vs. Handwriting:** Printed text will always be converted faster and with more accuracy. Depending on the handwriting quality, you will see less accuracy. Also forms completed with handwriting (US World War I Draft Registration Cards) will be difficult due to the formatting of the form, especially the lines.
- **Platform:** Some of the AI platforms listed above will do a better job with better accuracy. Try different platforms for best results.

AI and Source Citations

Those new to genealogy and family history soon learn the importance of source citations in proving relationships as well as facts about an ancestor. Usually source citations document how we find and use records such as census population schedules, death certificates, and even letters or diaries.

Citing sources need not be intimidating or time consuming. Stick to the basics: the information found, how it was found, information about where it was found, and locator data so another researcher can find the information.

For artificial intelligence content, here’s the formula you might consider using as proposed by the Modern Language Association of America (MLA):

> “[QUERY]” prompt. [NAME OF AI PLATFORM], [DATE OR VERSION OF PLATFORM], [NAME OF AI COMPANY], [DATE OF QUERY], [PLATFORM URL]

So, if I asked ChatGPT to translate a page from the book *Le troisième centenaire de l'Édit de Nantes en Amérique et en France*, here is the source citation I would use:

> “Translate to English” prompt using digital image of *Le troisième centenaire de l'Édit de Nantes en Amérique et en France*, page 3, published 1989. ChatGPT, ChatGPT 3.5 version, OpenAI, 12 March 2024, https://chat.openai.com/.

Tips and Tricks

General

- **Provide clear and concise instructions**. Make your prompt clear as to the desired result, such as “transcribe exactly” instead of “transcribe.” For translation use “translate to English” rather than “translate.”

- **Remember to provide feedback.** Most AI platforms provide a way to give feedback on the results generated. Remember these are “learning” platforms and this feedback is essential to better results in the future.

- **Break down the task into separate prompts.** You may find more success if you first ask the AI platform to transcribe a foreign language document and then follow up with the “translate to English” prompt.

- **Beware of platforms recommended by search engines.** There are many “specialized” AI platforms stating that they do a better job of transcribing or translating than “the bigger, better-known” platforms.

- **Review and test customized ChatGPTs.** Most of the offerings in the Explore ChatGPTs section are merely leads to external websites trying to sell an AI tool. Look for the number of chat conversations executed; the higher the number, the more popular the tool.

Summarization

- **Specify format.** Add modifiers to your prompt such as “bullet points.”

Transcription

- **Break down into smaller segments**. Especially when uploading a document or image containing handwritten text, use single pages or smaller segments for better accuracy.

- **Transcribe audio and video content**. Consider using AI to also upload video or audio files to create a transcription.

Translation

- **Craft specific prompts.** When uploading a document or image, use “translate to English” or other language instead of just “translate.”

- **Carefully review translated text.** The translated text should be similar in length to the original text. Some AI platforms may generate lengthy translations that are not useful.

- **Check the context for translated text.** While most AI platforms do a good job of translating text from one language to another, many results are literal rather than contextual.

Resources

- ***AI and Genealogy: A Practical Guide to Summarizing, Transcribing, and Translating Historical Documents*** – by Thomas MacEntee
  https://amzn.to/42lVdzq
- **ChatGPT**
  https://chat.openai.com/
- **Claude**
  https://claude.ai/
- **Coalition for Responsible AI in Genealogy**
  https://craigen.org/
- **CoPilot**
  https://copilot.microsoft.com/
- **Find Results with Full Text** - FamilySearch Labs
  https://www.familysearch.org/search/full-text
- **Gemini**
  https://gemini.google.com/
- **Genealogy and Artificial Intelligence (AI)** - Facebook group
  https://www.facebook.com/groups/1255245945084761