Journal of Open Humanities Data

# Text Recognition Model for Yiddish in *Vaybertaytsh* Typeface, Based on Community Regulations

RESEARCH PAPER

RONNY RESHEF
MIRJAM GUTSCHOW
*Author affiliations can be found in the back matter of this article

ubiquity press

## ABSTRACT

We present a public text recognition PyLaia model accompanied by a baseline model for the layout of community regulations in Yiddish and a dataset for Yiddish texts printed in *Vaybertaytsh* typeface. The model was built using legal documents, namely regulations written by the Ashkenazi Jewish community in Amsterdam during the 18th century. The necessity of such a model for *Vaybertaytsh* typeface stems from the substantial differences between it and other Yiddish or Hebrew typefaces. Existing text recognition models for Yiddish are dedicated to handwritten texts or substantially other typefaces, followed by a short description of the dataset, its unique characteristics, and how it can be used further. The process of training the text recognition model is explained, and challenges encountered are specified, as well as strategies for coping with them. The model is publicly accessible via Transkribus, and the complete dataset used to train the model is available via Figshare. The models and dataset offer valuable contributions to the digital humanities, specifically for research on linguistics, Jewish History and related fields.

**CORRESPONDING AUTHOR:**
**Ronny Reshef**
Department of Business-Society Management, Rotterdam School of Management, Erasmus University, Rotterdam, The Netherlands
reshef@rsm.nl

**KEYWORDS:**
Yiddish printing; Vaybertaytsh; Transkribus; printing history; Western Yiddish; Amsterdam; Netherlands

**TO CITE THIS ARTICLE:**
Reshef, R., & Gutschow, M. (2024). Text Recognition Model for Yiddish in *Vaybertaytsh* Typeface, based on community regulations. *Journal of Open Humanities Data*, 10: 35, pp. 1–10. DOI: https://doi.org/10.5334/johd.194

# (1) OVERVIEW

## CONTEXT

This paper presents a typeface model, Vaybertaytsh.YidTakNL, that can facilitate reading and researching Yiddish texts printed in the Vaybertaytsh typeface. Yiddish is the historical, traditional language of Ashkenazi Jews. Ashkenazi Jewry originates from the Holy Roman Empire. In the 11th century, Ashkenazim gradually started spreading throughout Europe. Yiddish gradually developed into quite a uniform literary language and is one of the most popular languages in Jewish culture and history, succeeding classical Hebrew and (Jewish) Aramaic. Yiddish was the vernacular used by Ashkenazi Jews on a daily basis. Since its mediaeval origins, it has been used at home, in religious institutions, and in literature. Later, it also became prominent in theatres, politics, schools, and journalism. (Weinreich 2007: 332, 336).

The syntax of Yiddish was, in its inception, a combination of Germanic with Hebrew and Aramaic, yet it has always also been influenced by local languages. Yiddish is written in the Hebrew Alphabet, which is primarily vowel-free.[1] With time, the Semitic Hebrew script was adjusted into an alphabet containing consonants and vowels, for example, by using the letter Ayin (ע) for e, Aleph (א) for a and o, and the combination Vav-Yud (וי) and two Yuds (יי) as diphthongs. The Yiddish writing system continuously evolved, and conventions of word separation, the representation of unstressed vowels and diphthongs substantially changed during the 13th–19th centuries. Moreover, dialects differ in terms of grammar, as well as phonologically, in terms of the evolution of vowels. Scholars often use the distinct categories Old Yiddish, Modern Eastern Yiddish, and Western Yiddish (Schäfer 2023: 8–9; Jacobs 2005: 2–4, 10; Birnbaum 2016: xlviii note 2; Weinreich 2008: 4; Katz 2011; Weinreich 2007: 332–335; Beider 2015: 63–64).

In early modern times, Amsterdam had a flourishing Ashkenazi Jewish community that spoke Western Yiddish on a daily basis. Their dialect most probably originates mainly in German-speaking areas. During this period, Amsterdam thrived as one of the foremost centres of Yiddish book production, renowned for both the exceptional quality and impressive quantity of its literary output. Western Yiddish has a distinct grammar, phonology, pronunciation, and vocabulary compared to German. From the 17th century, Western Yiddish in the Netherlands gradually became influenced by Dutch. During the 19th century, Western Yiddish declined and was replaced by the local languages (Michman et al. 2007: 108; Fleischer 2018: 239, 250–251; Birnbaum 2016: xlviii note 2; Aptroot 1988: 8–11; Berger 2008: 203).

Yiddish texts were printed in the *Vaybertaytsh* typeface throughout Europe during the 16th–19th centuries. One of the oldest and most well-known examples is the Tz'enah Ur'enah prose work, also known as the Women's Bible, containing segments from the Torah and Haftarahs used in Jewish prayer services. The earliest edition, which survived, is dated 1622 and was written in Hanau. In Amsterdam, it was used to print regulations throughout the 18th century. These regulations have a distinct vocabulary and layout and give fascinating insight into the relationship between the Ashkenazi Jewish leadership and the Dutch government.

*Vaybertaytsh* is a semi-cursive Ashkenazi typeface, also called *vayberksav*, *taytsh*, *ivre-taytsh*, *Tsene-(u)rene-ksav*, *Tkhine-ksav*, *kleyn-taytsh* and *mashket/mesheyt* (Spinner 2019: 152; Zafren 1982: 138; Jacobs 2005: 47; Fishman 1991: 44; Fleischer 2018: 266). In a semi-cursive script, elements of both cursive and print (block) writing are combined. For the untrained eye, this typeface can be difficult to read. Hence, there was a need for a text recognition model.

## DATASET DESCRIPTION

### Object Name

Vaybertaytsh.YidTakNL

### Access

The model is publicly accessible via Transkribus: https://readcoop.eu/model/vaybertaytsh-typeface-18th-19th-century/; The complete dataset, including transcriptions and images of the texts used for training this model, can be accessed at: doi.org/10.6084/m9.figshare.25422844.

1 For example, the word "Parnassim" (community leaders) is written as "פרנסים", literally "prnsim"; "kahal" (crowd) is written as "קהל", literally "khl".

See also https://doi.org/10.5281/zenodo.10017358 for an overview of a corpus of regulations and announcements written by Amsterdam's Ashkenazi Jewish community between 1708 and 1846.

### Dataset creators

Ronny Reshef (creator, annotator), Mirjam Gutschow (annotator)

### Language

English, Yiddish, Hebrew

### Licence



### Reuse Potential

The text recognition model and accompanying baseline model can be used to advance research on Jewish history and Yiddish language, literature and culture. The model is very robust, and using it will enable the preparation and editing of digital transcriptions of Yiddish texts written in Vaybertaytsh efficiently, quickly and precisely. Transcriptions made using the model can especially contribute to research on the linguistic development of (Western) Yiddish, for example, by producing word lists for analysing linguistic structures and patterns. Moreover, transcriptions made using the model can also be processed for lexicographical and scientific purposes.

Historians can use the model to transcribe texts easily and conduct content analysis for efficient and thorough study of data, occurrences, and stories. Both models can be further improved by enriching the dataset in terms of vocabulary, text types (literature, religious texts), and layout. Such improvements will benefit all the users of the models and, subsequently, advance research on Jewish history and Yiddish language and literature. Finally, as Hodel et al. (2021: 8) suggest, collaboration is the key to advancing text recognition models and Vaybertaytsh.YidTakNL and its accompanying baseline model are no exception to this. Libraries and archives can provide substantial quantities of images for GT sets, academics can contribute transcriptions, and experts of digital humanities can assemble and unify the inputs in a standardised manner. To preserve, make accessible, and study our cultural heritage, we need to combine our wide range of expertise and unite our efforts.

## TEXT RECOGNITION OF VAYBERTAYTSH

The Vaybertaytsh.YidTakNL model was created via Transkribus using PyLaia, a Handwritten Text Recognition (HTR) engine. Transkribus is a platform that enables (automatic) text recognition, structure recognition, textual analysis, tagging, and image analysis of handwritten and printed historical documents while applying machine learning to improve its technology continuously. It is a user-friendly, accessible platform that makes historical texts with different formats readable, editable, and searchable by dividing pages into text regions, lines, and words and recognising the text. The machine learning principle uses HTR models of neural networks based on so-called ground truths (GTs) submitted to train the model. A GT is a manually transcribed version of a document which is accurate and has been manually verified insofar as possible. When thusly well-trained on a large number of GTs, Transkribus is able to provide accurate results.[2] (Muehlberger et al. 2019: 955, 957, 961; Ströbel, Clematide & Volk 2020: 3552, 3556, 3558).

Modern HTR tools like Transkribus can process handwritten and historical texts much more efficiently than traditional Optical Character Recognition (OCR). OCR was developed in the 1990s to identify single characters in modern printed texts accurately. HTR engines can be trained to recognise a practically limitless quantity of shapes representing a specific character and also strings of characters, meaning (sub) words belonging to a particular corpus (Couture, Verret, Gohier & Deslandres, 2023: 9; Nockels, Gooding, Ames & Terras 2022: 368–369).

HTR engines on Transkribus, such as the deep learning toolkit of PyLaia, can also be successfully used for printed documents. A GT of 10,000 words leads to results requiring relatively few corrections. The trained model is able to recognise printed texts which are comparable to the

2 See also https://readcoop.eu/what-is-ground-truth/ (Last accessed: 11 November 2023).

GT. Currently, Transkribus represents the state-of-the-art in text recognition systems available. At the time of this project, PyLaia was certainly one of the best HTR models available within Transkribus (Romein et al. 2020: 294; Muehlberger et al. 2019: 955, 957, 961; Ströbel, Clematide & Volk 2020: 3552, 3556, 3558; Maarand et al. 2022: 7, 10–11, 13; Rabus 2022: 182–183). A new feature is the "Transformer Based Models", or "Super Models", which could be even superior to PyLaia and may even better facilitate future projects.[3]

It is possible to train the model with a base model: an existing suitable HTR model and GT. A base model is not obligatory, yet it can be applied to transfer information from an already existing model to boost the creation of a new model. This allows the training algorithm to apply knowledge from an existing model based on a substantial dataset to perform pre-calibration and classification of the characters. This accelerates the process of training a new model and improves its performance (Ströbel et al. 2023: 7; Couture, Verret, Gohier & Deslandres 2023: 10).

## (2) METHOD

### STEPS

Transcriptions for Yiddish texts in the Vaybertaytsh typeface were prepared using Transkribus in 2023. Since Transkribus already had two public PyLaia HTR models for Yiddish, the *Dybbuk* and *DiJeSt*,[6] we chose to base our language model on one of the two existing models. However, *Vaybertaytsh* is substantially different from other common Hebrew and modern Yiddish typefaces; therefore, much work needed to be done before the transcriptions were at a GT level (See Table 1 for a comparison of scripts).[7]

**Table 1** The classic Hebrew typeface Meruba and Vaybertaytsh.

| MERUBA[4] | VAYBERTAYTSH[5] | UNICODE ID | LETTER NAME |
|---|---|---|---|
| א | א | U+05D0 (1488) | Alef |
| ב | בּ ב | U+05D1 (1489) | Bet |
| ג | ג | U+05D2 (1490) | Gimel |
| ד | ד | U+05D3 (1491) | Dalet |
| ה | ה | U+05D4 (1492) | He |
| ו | ו | U+05D5 (1493) | Vav |
| ז | ז | U+05D6 (1494) | Zayin |
| ח | ח | U+05D7 (1495) | Het |
| ט | ט | U+05D8 (1496) | Tet |

(Contd.)

3 See https://readcoop.eu/introducing-transkribus-super-models-get-access-to-the-text-titan-i/, (Last accessed: 06 March 2024).

4 We used the Drugulin font (sometimes mistakenly also Drogolin), one of the most typical Meruba fonts, to illustrate the difference between the typical typeface used in the Dybbuk model and the Vaybertaytsh typeface. The image of the letters originates in the following link: https://he.wikipedia.org/wiki/%D7%93%D7%A8%D7%95%D7%92%D7%95%D7%9C%D7%99%D7%9F_(%D7%92%D7%95%D7%A4%D7%9F)#/media/%D7%A7%D7%95%D7%91%D7%A5:Drogolin_hebrew_font2.jpg. (Last accessed: 10 November 2023).

5 The Vaybertaytsh letters are taken from: Amsterdam Ashkenazic Community. ([5]497 [i.e. 1737]). תקנות הקהלה דקהל קדש אשכנזים אשר באמשטרדם. Abraham ben Raphael Athias. https://www.hebrewbooks.org/37002. A transcription of the Takkanot (regulations) from 1737 can be found here: https://doi.org/10.5281/zenodo.8328503 (Last accessed: 25 November 2023).

6 https://www.dybbuk.co/; http://dijest.net/gtmodel; (Last accessed: 10 November 2023).
The project is headed by Dr Ruthie Abeliovich from Tel Aviv University.
The Dybbuk model is based on handwritten texts, which are typically more complex and have lower consistency. Handwritten materials have a larger variation in terms of size, spacing, slant and style. They therefore require a bigger and more diverse training set compared with printed texts, which are more standardised.

7 We compare the Vaybertaytsh typeface to the much-used Meruba (square) Hebrew font, since the Meruba font has clear, distinct letter shapes that are easily readable compared to Vaybertaytsh. The comparison is meant to show why we encountered challenges while developing the model, as Vaybertaytsh is very different from this classic Hebrew typeface. See Eldar 2023: 22-23; Blum 2017: 5; Kerler 1993: 14; We thank Dr. Rusinek for advising us on this.

| MERUBA[4] | VAYBERTAYTSH[5] | UNICODE ID | LETTER NAME |
|---|---|---|---|
| י | י | U+05D9 (1497) | Yud |
| ך | ך | U+05DA (1498) | Kaf Sofit |
| כ | כ כּ | U+05DB (1499) | Kaf |
| ל | ל | U+05DC (1500) | Lamed |
| ם | ם | U+05DD (1501) | Mem sofit |
| מ | מ | U+05DE (1502) | Mem |
| ן | ן | U+05DF (1503) | Nun Sofit |
| נ | נ | U+05E0 (1504) | Nun |
| ס | ס | U+05E1 (1505) | Samech |
| ע | ע | U+05E2 (1506) | Aiyn |
| ף | ף | U+05E3 (1507) | Pe Sofit |
| פ | פ פּ | U+05E4 (1508) | Pe |
| ץ | ץ | U+05E5 (1509) | Tsadi Sofit |
| צ | צ | U+05E6 (1510) | Tsadi |
| ק | ק | U+05E7 (1511) | Qof |
| ר | ר | U+05E8 (1512) | Resh |
| ש | ש | U+05E9 (1513) | Shin |
| ת | ת תּ | U+05EA (1514) | Tav |

The first transcriptions were made using the *Dybbuk* model since it seemed most suitable for our purposes. They were then corrected and double-checked. The transcriptions corrected manually finally had a high degree of precision so that the output could be used for scientific purposes. Since the initial results from the first models we created based on GT and the pre-existing Yiddish HTR models were unsatisfactory, we added more GT. We deployed a few private models to keep correcting and improving the model before making it public. In Table 2, we present a comparison of the word error rates for the three publicly available Yiddish handwritten text recognition (HTR) models within Transkribus. This analysis was conducted using a selection of ground truth texts written in Vaybertaytsh typeface. The texts used for the comparison are distinct from those employed for the development of the Vaybertaytsh.YidTakNL model.

**Table 2** Word error rate for texts written in Vaybertaytsh typeface.

| MODEL | # OF WORDS | # OF PAGES | WORD ERROR RATE | WORD ACCURACY |
|---|---|---|---|---|
| **The Dybbuk for Yiddish Handwriting (Nov 13, 2022)** | 6307 | 22 | 92,546 | 4,856 |
| **DiJeSt 2.0 (Nov 10, 2022)** | 6307 | 22 | 42,122 | 43,036 |
| **Vaybertaytsh.YidTakNL (Nov 29, 2023)** | 6307 | 22 | 9,26 | 85,468 |

The CER measures the character level error rate of incorrect transcriptions. A CER of 1% and lower can be achieved with a large enough corpus and sufficient training. However, a CER of less than 10% is generally considered sufficient for automatic transcriptions, depending on the desired application and goals (Li & Hill 2023: 2; Muehlberger et al. 2019: 962; Ströbel, Clematide & Volk 2020: 3552). The first private Vaybertaytsh.YidTakNL model was run in March 2023, using the *Dybbuk* as a base model and a training set size of 7163 words. The CER of the first model, Yid.Dutch.ver3, was 0.20%, as can be seen in Table 3. This model was then used to transcribe the next texts, which were also corrected by both authors.

| MODEL | YID.DUTCH.VER3 | YIDNL.7 | VAYBERTAYTSH.YIDTAKNL |
|---|---|---|---|
| **ID** | 51030 | 52365 | 57147 |
| **Date** | 2023-03-30 | 2023-05-22 | 2023-11-29 |
| **No of words** | 7 163 | 21 661 | 66 497 |
| **No of lines** | 752 | 2 434 | 8 062 |
| **Max epochs** | 250 | 100 | 100 |
| **Early stopping** | 20 | 20 | - |
| **Epochs trained** | 250 | 100 | 100 |
| **Learning rate** | 0.0003 | 0.0003 | 0.0003 |
| **Base model** | 46159 /The Dybbuk for Yiddish Handwriting | 46159 /The Dybbuk for Yiddish Handwriting | |
| **CER on Training Set** | 0.60% | 0.80% | 0.60% |
| **CER on Validation Set** | 0.20% | 0.20% | 0.91% |
| **No of pages in Training Validation Set** | 21 | 71 | 23 |

**Table 3** Overview of the 3 models with additional information.

The second private model, YidNL.7, was deployed in May 2023, again using the *Dybbuk* as a base model and a training set size of 21661 words. Although it was based on more GTs, the CER of the second model was also 0.20%, as presented in Table 3. A possible explanation could be that the vocabulary of the second model was still limited, and additional GT was required to decrease the errors it made. However, considering the satisfactory score, it could also well be that a training set size of about 7,000 words is sufficient for this sort of typeface.

We transcribed additional texts in the *Vaybertaytsh* typeface to improve its performance. The public Vaybertaytsh.YidTakNL text recognition model has a training set size of 66497 words and a CER of 0.90% (see Table 3). Despite having a slightly less good CER (0.90% compared with 0.20%), Vaybertaytsh.YidTakNL performs considerably better than our private text recognition models. A possible explanation could be related to the increase in vocabulary and source diversity.

We also deployed a baseline model for layout to work more efficiently with the printed texts. The Vaybertaytsh.YidTakNL accompanying layout model was necessary since the default layout model was not able to recognise the unique layout of the regulations. The default model ineffectively divided the page into a few random regions when one region was sufficient. It also over-segmented the lines, needlessly splitting multiple lines into 2 or 3 parts.

In order to build the layout model, we manually separated running titles and leave numbers, signatures and catchwords and placed the numbering and titles[8] of the chapters (or, as in this genre, mainly paragraphs) on separate lines. After manually correcting the layout of around 50 pages according to the set rules we developed, a baseline model was trained and constantly improved. Using the new baseline model, most pages were rendered, to a large extent, the way we wanted them to appear. The new baseline model, trained on a set of 60,036 words, renders most pages largely the way we want them to appear.

## QUALITY CONTROL

The main challenges with training a model for *Vaybertaytsh* were recurrent confusions in the recognition of some characters, which are very similar to each other.[9] Some examples are

8 These "titles" mostly consist of the first word of the paragraph and are typeset in a bold and large font size. When preparing an edition at a later stage, these words and their positions need to be addressed somehow – either by imitating the original style, or by adding a new marker to indicate the beginning of a new paragraph. By placing the words on a separate line, their identification is more straightforward.

This is related to one persisting problem, namely the fact that Transkribus is generally based on the assumption that the reading order of the text is from left to right, whereas the Yiddish texts, and their layout, are drawn up from right to left. The consequence is that in Transkribus the first line of a paragraph is placed ahead of its heading, and the catchword (bottom left) is placed ahead of the signature (bottom centre).

9 Maarand et al. 2022: 12 studied handwritten Norwegian texts. They discovered that the HTR model (PyLaia) is more spread across alternatives when compared with printed OCR. Moreover, the common OCR substitutions of single characters are present in the HTR model, but their relative weight is lower. Similar results were found for usual OCR substitutions of special characters. Furthermore, Maarand et al. noticed a few examples of confusion which is not typical for OCR and are quite important.
Their findings could explain the phenomena of confusion we encountered to some extent.

DOI: 10.5334/johd.194

confusions between Yud (י), Vav (ו) and Geresh (׳), Resh (ר) and Dalet (ד), and Tsadi Sofit (ץ) and Fe Sofit (ף). Other challenges were omitting superscripts (diacritical signs), errors with word separation, irregular spacing, disregarding hyphenations, punctuation marks, and some hypercorrection errors,[10] where letters were added for no clear reason. Rabus (2022: 186–187) had similar issues training a Transkribus model in Croatian Glagolitic. Table 4 summarises the most common letter confusions we encountered. These confusions were repeatedly manually corrected to teach the algorithm the difference between these letters. With every new model, there were fewer letter confusions and hypercorrections, and the overall result became more satisfactory. We attribute the ongoing enhancement of the model to the systematic incorporation of additional texts at a GT level, coupled with meticulous rectification of recurring issues. The public text recognition model functions much better than the previous models developed, and there is substantially less confusion.

**Table 4** Recurrent issues with training the *Vaybertaytsh* model.[11]

| LETTER | PICTURE | CONFUSED WITH (AND VICE VERSA) | PICTURE |
|---|---|---|---|
| Mem מ |  | Aiyn ע |  |
| Vav + Nun ו + נ |  | Mem מ |  |
| Vav ו |  | Yud י |  |
| Tet ט |  | Shin ש |  |
| Dalet ד |  | Resh ר |  |
| Mem מ |  | Alef א |  |
| ) |  | Lamed ל |  |
| He ה |  | Het ח |  |
| Geresh ' |  | Yud י |  |
| Geresh ' |  | Vav ו |  |
| Vav ו |  | Nun Sofit ן |  |
| Vav ו |  | Zayin ז |  |
| Samech ס |  | Mem sofit ם |  |
| Tsadi Sofit ץ |  | Fe Sofit ף |  |
| Aiyn ע |  | Tet ט |  |
| Samech ס |  | Tet ט |  |
| Nun נ |  | Gimel ג |  |
| Nun נ |  | Kaf כ |  |
| Bet ב |  | Kaf כ |  |
| Shin ש |  | Samech ס |  |
| Tav ת |  | Het ח |  |
| Lamed ל |  | Tsadi צ |  |

10 Throughout the years, hypercorrections are becoming increasingly rare, yet they occur in models dedicated to one language, including Vaybertaytsh.YidTakNL. Furthermore, models that integrate diverse sources over an extended period and across multiple languages rarely experience hypercorrections.

The hypercorrections in Vaybertaytsh.YidTakNL could stem from overgeneralization of an orthographic or linguistic pattern by the model, leading to a wrong transcription (as observed by Rabus 2019, 12).

Since Vaybertaytsh.YidTakNL is only based on Yiddish, and since RtL languages still do not have TrOCR models (pre-trained image and text Transformer models which are based on more than a million tokens) on Transkribus, the chance of hypercorrections remains. This could lead to misinterpretations and pose a challenge for scholarly editions, where precision is vital, and each letter needs to be the same as in the original.

11 The letters originate from the following 2 sources: Takkanot 1737 op cit; Amsterdam, Ashkenazic Community. ([5]532 [1772]). Translat, Ampliatsien ʻal takanot kehilatenu (10 Detsember).
(10 דעצעמבר) טראנסלאט אמפליאצינען על תקנות קהלתינו. Kosman ben Joseph Baruch. https://www.hebrewbooks.org/46840 & https://www.nli.org.il/en/books/NNL_ALEPH990012586850205171 (Last accessed: 26 January 2024).

## CONCLUSION

The less-than-satisfactory initial results in early 2023 underscore the importance of tenacity and perseverance when developing an HTR model. Proofreading and correcting additional texts to GT status and periodically updating the model significantly improved the quality, rendering the next batch of pages with its subsequent version. Basing the first models on existing Yiddish HTR models was crucial since they provided a solid foundation for Vaybertaytsh.YidTakNL. The model can automatically recognise 18th century Yiddish text from Amsterdam, enabling information searches within the text without additional human editing. While achieving a flawless text suitable for a scholarly edition requires additional human work, this is common when preparing scholarly editions.

Given the model's derivation from specific text types with characteristic layout and a moderate vocabulary (See Reshef & Gutschow 2023), pages of several other Yiddish texts printed in Amsterdam during the 18th century will be added to it in the near future. These texts feature alternative layout structures and distinct vocabularies, enriching and strengthening the Vaybertaytsh.YidTakNL model. Furthermore, as certain community regulations include cursive Ashkenazi, a distinctive typeface mimicking handwriting, the next model version will incorporate several fully set book pages using this style. Afterwards, we aspire to continue improving and elaborating the model by adding a variety of texts from other Ashkenazi communities in Europe. With time, we anticipate that the Vaybertaytsh.YidTakNL model will become more diverse and, as a result, improve significantly, thanks to community efforts.

## ACKNOWLEDGEMENTS

Dr. Sinai Rusinek, University of Haifa

Dr. C. Annemieke Romein, Huygens Instituut, Amsterdam

Assaf Urieli, Joliciel Informatique

Jenneken Schouten, Subject Librarian for Hebrew, New Greek and Religious Studies at the Library of the University of Amsterdam

Prof. Dr. Marion Aptroot, Heinrich Heine University Düsseldorf

Help Desk team, READ-COOP

## COMPETING INTERESTS

The authors have no competing interests to declare.

## AUTHOR CONTRIBUTIONS

Ronny Reshef: writing – original draft, writing – review and editing, project administration

Mirjam Gutschow: writing – review and editing, data curation, visualisation

## AUTHOR AFFILIATIONS

**Ronny Reshef** orcid.org/0009-0006-3583-7005
Department of Business-Society Management, Rotterdam School of Management, Erasmus University, Rotterdam, The Netherlands

**Mirjam Gutschow** orcid.org/0000-0002-0356-5475
Department of Jewish Studies, Yiddish Culture, Language and Literature, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

## REFERENCES

**Aptroot, M.** (1988). Dutch Impact on Amsterdam Yiddish Prints. In D. Katz (Ed.), *Dialects of the Yiddish Language. Papers from the Second Annual Oxford Winter Symposium in Yiddish Language and Literature, 14–16 December 1986* (pp. 7–11). Oxford: Pergamon. DOI: https://doi.org/10.1016/B978-0-08-036564-0.50005-3

**Beider, A.** (2015). *Origins of Yiddish dialects*. Oxford: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780198739319.001.0001

**Berger, S.** (2008). "Yiddish book production in Amsterdam Between 1650–1800: Local and international aspects". In *The Dutch Intersection*. Leiden, The Netherlands: Brill. DOI: https://doi.org/10.1163/ej.9789004149960.i-450.52

**Birnbaum, S. A.** (2016). *Yiddish: a survey and a grammar*. Toronto: University of Toronto Press. DOI: https://doi.org/10.3138/9781442665330

**Blum, S. T.** (2017). Typography and the Evolution of Hebrew Alphabetic Script: Writing Method of the Sofer. *Faculty and Staff Publications*, *49*. https://digitalcommons.xula.edu/fac_pub/49

**Couture, B., Verret, F., Gohier, M.,** & **Deslandres, D.** (2023). The Challenges of HTR Model Training: Feedback from the Project Donner le gout de l'archive a l'ere numerique. *Journal of Data Mining & Digital Humanities*. DOI: https://doi.org/10.46298/jdmdh.10542

**Eldar, G.** (2023). The Visual Type as an Image of a People: Hebrew Typography throughout History and Its Representation of Jewish and Israeli Identity. In *Imagined Israel(s): Representations of the Jewish State in the Arts* (pp. 19–39). Leiden: Brill. DOI: https://doi.org/10.1163/9789004530720_003

**Fishman, J. A.** (1991). *Yiddish: Turning to life*. John Benjamins Publishing. DOI: https://doi.org/10.1075/z.49

**Fleischer, J.** (2018). Western Yiddish and Judeo-German. In *Languages in Jewish communities, past and present (Vol. 112)*. B. Hary & S. B. Benor (Eds.), pp. 239–275. *Berlin; Boston: De Gruyter Mouton*. DOI: https://doi.org/10.1515/9781501504631-009

**Hodel, T., Schoch, D., Schneider, C.,** & **Purcell, J.** (2021). General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example. *Journal of Open Humanities Data*, *7*, 13.DOI: https://doi.org/10.5334/johd.46

**Jacobs, N. G.** (2005). *Yiddish: A linguistic introduction*. Cambridge: Cambridge University Press.

**Katz, D.** (2011, October 31). Language: Yiddish. *YIVO Encyclopedia of Jews in Eastern Europe*. Retrieved November 10, 2023, from https://yivoencyclopedia.org/article.aspx/Language/Yiddish

**Kerler, D. B.** (1993). *The Origins of Modern Literary Yiddish. Oxford Modern Languages and Literature Monographs*. Oxford: Oxford University Press. https://oxford.universitypressscholarship.com/10.1093/acprof:oso/9780198151661.001.0001/acprof-9780198151661

**Li, S.,** & **Hill, N.** (2023). Printed Text Recognition for Lexical Lists in Chinese-International Phonetic Alphabet (IPA) Glossing. *Journal of Open Humanities Data*, *9*, 15, pp. 1–8. DOI: https://doi.org/10.5334/johd.119

**Maarand, M., Beyer, Y., Kåsen, A., Fosseide, K. T.,** & **Kermorvant, C.** (2022, May). A comprehensive comparison of open-source libraries for handwritten text recognition in Norwegian. In *International Workshop on Document Analysis Systems* (pp. 399–413). Cham: Springer International Publishing. DOI: https://doi.org/10.1007/978-3-031-06555-2_27

**Michman, J., Wallet, B., Michman, D., van Bekkum, W., Brasz, C.,** & **Adler, I.** (2007). Amsterdam. In M. Berenbaum & F. Skolnik (Eds.), *Encyclopaedia Judaica* (2nd ed., Vol. 2, pp. 106–120). Macmillan Reference USA. https://link.gale.com/apps/doc/CX2587501034/GVRL?u=duessel&sid=bookmark-GVRL&xid=6e0a5866

**Muehlberger, G., Seaward, L., Terras, M., Oliveira, S. A., Bosch, V., Bryan, M., ...** & **Zagoris, K.** (2019). Transforming Scholarship in the Archives Through Handwritten Text Recognition: Transkribus as a Case Study. *Journal of documentation*, *75*(5), 954–976. DOI: https://doi.org/10.1108/JD-07-2018-0114

**Nockels, J., Gooding, P., Ames, S.,** & **Terras, M.** (2022). Understanding the Application of Handwritten Text Recognition Technology in Heritage Contexts: a Systematic Review of Transkribus in Published Research. *Archival Science*, *22*(3), 367–392. DOI: https://doi.org/10.1007/s10502-022-09397-0

**Rabus, A.** (2019). Training Generic Models for Handwritten Text Recognition using Transkribus: Opportunities and Pitfalls. In *Proceedings of the Dark Archives Conference*. https://www.academia.edu/49356690/Training_generic_models_for_Handwritten_Text_Recognition_using_Transkribus_Opportunities_and_pitfalls

**Rabus, A.** (2022). Handwritten Text Recognition for Croatian Glagolitic. *Slovo: časopis Staroslavenskoga instituta u Zagrebu*, *72*(1), 181–192. DOI: https://doi.org/10.31745/s.72.5

**Reshef, R.,** & **Gutschow, M.** (2023). YidTakNL Corpus: 18th–19th Centuries Regulations of the High German Jewish Community in Holland. *Journal of Open Humanities Data*, *9*(1), 29. DOI: https://doi.org/10.5334/johd.161

**Romein, C. A., Kemman, M., Birkholz, J. M., Baker, J., De Gruijter, M., Meroño-Peñuela, A., Ries, T., Ros, R.** & **Scagliola, S.** (2020). State of the field: digital history. *History*, *105*(365), 291–312. DOI: https://doi.org/10.1111/1468-229X.12969

**Schäfer, L.** (2023). Yiddish. In *Oxford Research Encyclopedia of Linguistics*. DOI: https://doi.org/10.1093/acrefore/9780199384655.013.946

**Spinner, S. J.** (2019). Reading Jewish. *PMLA Modern Language Association of America*, *134*(1), 150–156. DOI: https://doi.org/10.1632/pmla.2019.134.1.150

**Ströbel, P. B., Clematide, S.,** & **Volk, M.** (2020, May). How Much Data Do You Need? About the Creation of a Ground Truth for Black Letter and the Effectiveness of Neural OCR. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 3551–3559). DOI: https://doi.org/10.5167/uzh-197209

**Ströbel, P. B., Hodel, T., Boente, W.,** & **Volk, M.** (2023, August). The Adaptability of a Transformer-Based OCR Model for Historical Documents. In *International Conference on Document Analysis and Recognition* (pp. 34–48). Cham: Springer Nature, Switzerland. https://www.zora.uzh.ch/id/eprint/235624/1/adapda_workshop_2023.pdf

**Weinreich, M.** (2008). *History of the Yiddish Language* (Vol. 1). New Haven, Connecticut: Yale University Press.

**Weinreich, U.** (2007). Yiddish Language. In M. Berenbaum & F. Skolnik (Eds.), *Encyclopaedia Judaica* (2nd ed., Vol. 21, pp. 332–338). Macmillan Reference USA. https://link.gale.com/apps/doc/CX2587521264/GVRL?u=duessel&sid=bookmark-GVRL&xid=bb601baa

**Zafren, H. C.** (1982). Variety in the Typography of Yiddish: 1535–1635. *Hebrew Union College Annual*, 137–163. https://www.jstor.org/stable/23507628

**TO CITE THIS ARTICLE:**
Reshef, R., & Gutschow, M. (2024). Text Recognition Model for Yiddish in *Vaybertaytsh* Typeface, based on community regulations. *Journal of Open Humanities Data*, 10: 35, pp. 1–10. DOI: https://doi.org/10.5334/johd.194

**Submitted:** 23 December 2023
**Accepted:** 08 April 2024
**Published:** 06 May 2024



*Journal of Open Humanities Data* is a peer-reviewed open access journal published by Ubiquity Press.