Shorthand Secrets: Deciphering Astrid Lindgren's Stenographed Drafts with HTR Methods

Raphaela Heil[⋆][0000−0002−5010−9149], Malin Nauwerck[‡†], and Anders Hast[⋆‡][0000−0003−1054−2754]

[⋆] Department of Information Technology, Uppsala University, Uppsala, Sweden
{raphaela.heil, anders.hast}@it.uu.se
[†] Department of Literature and Rhetoric, Uppsala University, Uppsala, Sweden
[‡] The Swedish Institute for Children's Books, Odengatan 61, 113 22 Stockholm, Sweden
malin.nauwerck@barnboksinstitutet.se
https://www.barnboksinstitutet.se/en/om-sbi/

Abstract. Astrid Lindgren, Swedish author of children's books, is known for having both composed and edited her literary work in the Melin system of shorthand (a Swedish shorthand system based on Gabelsberger). Her original drafts and manuscripts are preserved in 670 stenographed notepads kept at the National Library of Sweden and The Swedish Institute of Children's Books. For long these notepads have been considered undecipherable and are until recently untouched by research.

This paper introduces handwritten text recognition (HTR) and document image analysis (DIA) approaches to address the challenges inherent in Lindgren's original drafts and manuscripts. It broadly covers aspects such as preprocessing and extraction of words, alignment of transcriptions and the fast transcription of large amounts of words.

This is the first work to apply HTR and DIA to Gabelsberger-based shorthand material. In particular, it presents early-stage results which demonstrate that these stenographed manuscripts can indeed be transcribed, both manually by experts and by employing computerised approaches.

Keywords: Stenography · Handwritten Text Recognition · Digital Transcription · Document Image Analysis.

1 Introduction

Swedish author Astrid Lindgren holds a unique position within world literature, yet her enigmatic creative process, progressing from idea to finished work, has for many years been hidden in her original drafts and manuscripts, written in the stenographic system of Melin shorthand. These supposedly indecipherable original drafts have to a large extent contributed to Lindgrens creative process being surrounded by myth.

Another factor in this myth-making relates to the role Lindgren played in the production process of her own books. The year after her breakthrough with *Pippi Longstocking* (1945), Lindgren was employed by her publishing house Rabn & Sjgren and became her own editor and publisher. Lindgren who used shorthand both for writing and editing typically typed up her own notepads into full prose manuscripts, which subsequently were delivered directly to the typesetter and printer. Her colleagues at Rabn & Sjgren reputedly first met her new books as they arrived in cardboard boxes from the printer [4]. This image is confirmed by the typed up manuscripts preserved in the Astrid Lindgren Archives, which have been described as flawless transcriptions which contain little to none information beyond the printed book [1,15]. Consequently, Lindgrens shorthand notepads is the only first hand source to her creative process and the development of her literary work.

Altogether, 670 shorthand notepads are known to have been preserved. The major part (660) is in the possession of the Astrid Lindgren Archives at the National Library of Sweden[1], whereas the remaining ten are held at the Swedish Institute for Childrens Books[2]. The notepads include most of Lindgrens production, from her first published novels *Confidences of Britt-Mari* (1944) and *Pippi Longstocking* (1945) to her last, *Ronia, the Robbers Daughter* (1981). So far, 52 out of 670 notepads primarily containing shorthand drafts and manuscripts to *The Brothers Lionheart* (1973) have been digitized for the purposes of the ongoing interdisciplinary project, the Astrid Lindgren Code (20202022)[3].

1.1 Deciphering the "undecipherable" notepads: The Astrid Lindgren Code project

In contrast to those of other world famous authors of shorthand such as Charles Dickens [5], Lindgren's original manuscripts in shorthand have never been subject to research. Although an inventory of their content has been compiled by parliamentary secretary Britt Almstrm, manual transcription and analysis have been dismissed as highly time consuming or even impossible [16].

[1] The Astrid Lindgren Archives in the National Library Manuscript Division in Stockholm is the largest ever bequeathed by a single Swedish individual. In 2005 the Astrid Lindgren Archives was inscribed in UNESCO's Memory of the World register.

[2] These 10 notepads were donated by Lindgren to the Swedish Institute for Children's Books in 1968 and contain original drafts to the last part of the *Karlsson on the roof* trilogy (19551968).

[3] Full title "The Astrid Lindgren Code: Accessing Astrid Lindgrens shorthand manuscripts through handwritten text recognition, media history, and genetic criticism". The project can be further explored at the official website: https://www.barnboksinstitutet.se/en/forskning/astrid-lindgren-koden/

The main purpose of the Astrid Lindgren Code project is the deciphering of Lindgrens stenography through HTR, applied and developed in interactive combination with collective transcription through volunteer expert sourcing.[4]

The Melin system which Lindgren used is the only form of shorthand widely practised in Sweden. Lindgren learned it in the 1920s as part of her professional secretarial training. The system is based partially on the German Gabelsberger system, working according to the frequency of particular sounds in the Swedish language, and is built on phonetic symbols which basic elements are vowels, consonants and consonant combinations, as well as a wide range of abbreviations and suffixes.

The materiality of the notepads from pen colour to paper quality the stenographic practice of for example flipping the notepad and writing backwards in order to change topic or save paper, and the specific elements of a particular stenographic system (i.e. phonetically composed word images, personal abbreviations, contextual interpretation) all pose challenges to existing methods for HTR analysis.

2 Challenges and Solutions

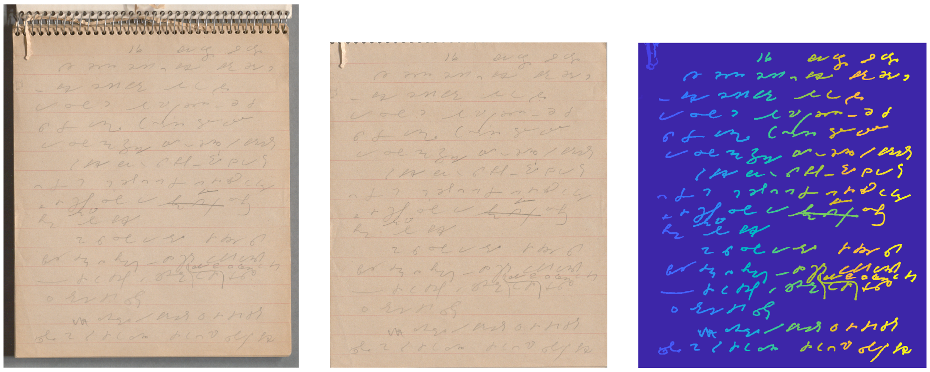
Generally, shorthand recognition has not been the focus of much research. Some works were performed in the 1980s to 90s, focusing on Pitman's shorthand, with one of the main resources being [11]. In addition, some research has been performed with regard to Gregg's shorthand system, most recently [19]. To the best of our knowledge, HTR has not yet been applied to manuscripts written in Gabelsberger or the Swedish Melin system.

For the present corpus, 52 notepads, containing approximately 7000 pages, have been digitised up until now. These page images pose a number of challenges that have to be addressed before automatic recognition methods can be applied to the material. A selection of these challenges and our proposed solutions are briefly outlined in the following sections.

2.1 Preprocessing and Segmentation

Firstly, the individual pages are handled by a preprocessing pipeline that prepares the material for the subsequent extraction and recognition stages. Due to the way the pages were written and digitised, the rotation of many images has to be adjusted in the first step, so that the text runs from the top left to the bottom right of the image. Generally, the first line of the writing follows after the ring binding, the latter of which can therefore be used to identify the actual

[4] An open digital platform for collective transcription and peer-editing of the 52 digitized notepads primarily containing drafts to *The Brothers Lionheart* is to be released in 2021. Although inspired by existing transcription projects based on crowdsourcing, a more appropriate term might be expert sourcing, as two very specific skills are required for volunteer participation: knowledge of shorthand and knowledge of the Swedish language.



(a) Original notepad (b) Preprocessed page (c) Text only

Fig. 1: Example of a) a notepad page containing draft to the last chapter (16) of *The Brothers Lionheart* b) the same page automatically cropped and with corrected illumination c) the cropped page with background and red lines removed in order to detect individual word images; coloured components indicate the different detected words.

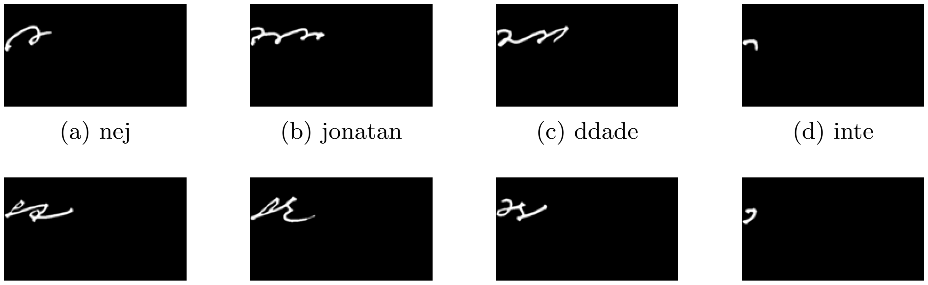
top of the page. In a next step, the extent of the textual content is detected and the images are cropped accordingly (cf. Figure 1 a) and b)). Additionally, empty pages are identified and discarded.

Subsequently, the illumination is corrected using [8] and the background is removed by applying [17], leaving only the greyscale pencil strokes. Finally, the boundaries of individual words are detected. In addition to this, the printed red lines, as shown e.g. in Figure 1 a), are identified and used to uniformly position the words during the extraction process. Examples for this positioning can be seen in Figure 2. The words are arranged so that the original printed line would run through the centre of each image. This step is especially crucial for the correct transcription of words, as the positioning of strokes, relative to the baseline of the text affects their transliteration. It should be noted that although this line may be of relevance for human transcribers, it is removed in the extracted words and represented implicitly by the relative positioning, in order to minimise the amount of distractions for the recognition algorithm.

Regarding the extraction of words, another challenge sometimes occurs in relation to the type of pencil Astrid Lindgren used. While most of the text was written with a common, grey lead pencil, for some portions, a red pen was used. As the lines are also printed in a shade of red, this overlap of colours may interfere with the regular processing as described above. Nevertheless, this problem can be solved by adapting the colour content before processing.

2.2 Transcription Alignment

At the time of writing, approximately 70 pages have been transcribed by stenographers. These transcriptions are currently only available as pure text, on a



(a) nej (b) jonatan (c) ddade (d) inte

(e) katla (f) karm (g) gjorde (h) det

Fig. 2: Words extracted and placed correctly in a box. In the Melin system of shorthand, interpunctuation is only implied. The literal transcription of the sentence is "a) nej b) jonatan c) ddade d) inte e) katla f) karm g) gjorde h) det". In English translation with added interpunctuation: "No, Jonathan didn't kill Katla, Karm did that."

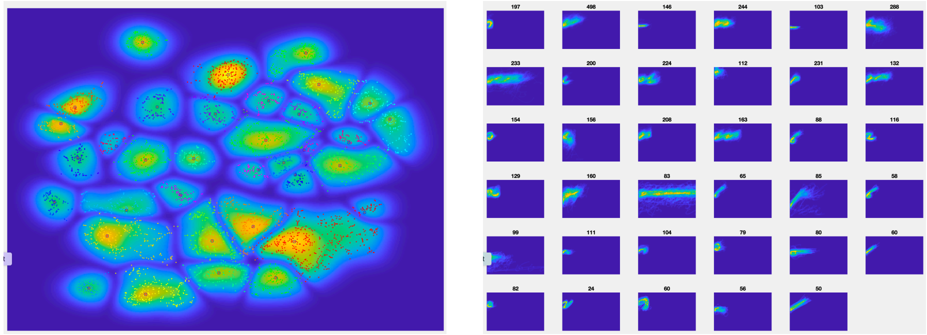
page-level, and not yet associated with the respective word images which were obtained by the previous processing steps. In order to use this data for training and evaluation of segmentation- and learning-based recognition approaches, a mapping between word image and transcription is required. A number of approaches have been proposed in the past to facilitate such an alignment. Several of these are based on Hidden Markov Models, such as [7], [14] and most recently [18]. We intend to explore a similar approach for the ground-truth creation, possibly also employing more advanced natural language processing techniques.

2.3 Semi-Automatic Transcription using Clustering

Given the limited ground-truth annotations that are currently available and the constraints on suitable transcribers, introduced by the stenographic material, we intend to apply a semi-automatic approach to speed up the transcription process. Once a larger amount of data has been transcribed, other methods, such as learning-based approaches, can be explored and applied to the remaining content.

Our semi-automatic transcription approach is based on [9] and is related to active learning, e.g. [2], and visual-interactive labelling, e.g. [3]. Taking the segmented words, obtained by the preprocessing step in section 2.1, hand-crafted features, such as histograms of oriented gradients (HOG) [6] are extracted for each word image. Other feature extraction methods, e.g. via neural networks, are conceivable as well and will be evaluated in the context of this approach.

The extracted features are mapped to a 2D representation, using dimensionality reduction techniques, such as t-SNE [12] and UMAP [13]. As has been proposed for handwritten digits (MNIST [10]), using HOG and t-SNE, in [9], the datapoints in the 2D representation are clustered together based on visual simi-



(a) Cluster Atlas

(b) Cluster Heatmaps

Fig. 3: Example of clustering and heatmaps of the contents of each cluster. The number above the heatmaps indicate the number of words in each cluster. The first level of clustering finds words of similar appearance, mainly based on width. Datapoints are coloured based on cluster membership and not by transcription label.

larities as shown in Fig. 3. This aspect of the visualisation will be employed in an interactive user interface (UI). The UI, which is currently under development, presents users (here: stenographers) with a rendering of the 2D datapoints. Users can select single or multiple points or whole clusters, inspect the associated word images and subsequently annotate them with a transcription.

One objective in this regard is to identify a combination of feature extractor and dimensionality reduction technique, such that the 2D representation results in homogeneous clusters that can be extracted and annotated, thereby transcribing large amounts of word images within a short amount of time and with limited user input.

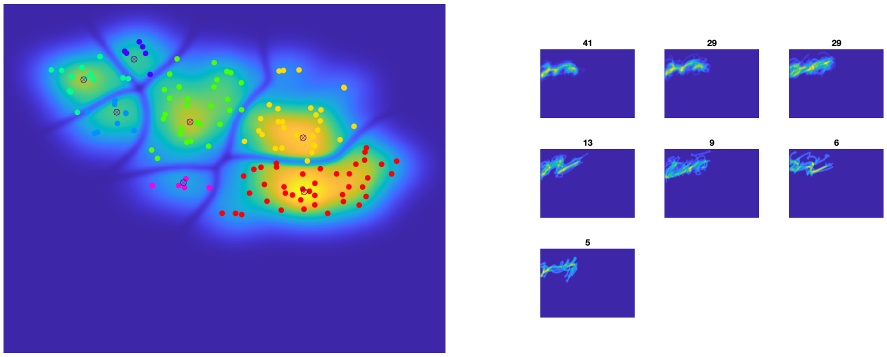
On the first cluster level words are clustered mainly based on their width as shown in Fig. 3. Sub-clustering is done by performing the t-SNE again on each cluster and an example is shown in Fig. 4. Depending on how many notepads are being processed, this process might need to be repeated several times. In the example, only one notepad has been processed and each cluster now contains rather few words and can therefore easily be examined visually. Fig. 5 shows the content of the two first heatmaps that mainly contain the word 'jonatan'. This shows that the technique described can be used to speed up transcription by facilitating annotation of batches of words, rather than transcribing word by word and line by line.

3 Conclusions

In this work we have introduced the Astrid Lindgren Code project, its inaccessible manuscripts as well as a selection of challenges that these pose to handwritten text recognition. We address these challenges by introducing a preprocessing pipeline and proposing solutions for the annotation and transcription of the material, both by employing already made transcriptions and by providing an approach for the fast, semi-automatic processing of the remainder.

The presented clusterings demonstrate that these stenographic texts which, as mentioned above, were once thought to be potentially impossible to transcribe, can indeed be processed by handwritten text recognition and document image analysis approaches.

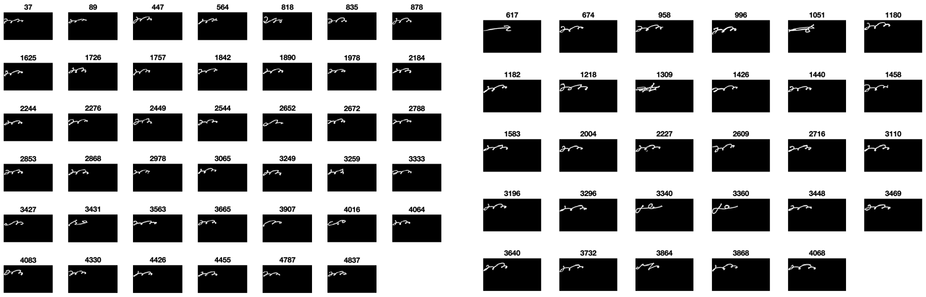
In the future, we intend to expand this application of HTR techniques with the goal of providing (semi-) automatic approaches for the fast and precise transcription of the more than 600 remaining manuscripts in this corpus.



(a) Sub cluster Atlas

(b) Sub cluster Heatmaps

Fig. 4: Example of sub clustering of one of the clusters in Fig. 3 and the corresponding heatmaps. In the second level individual words start to appear.



(a) Sub cluster 1

(b) Sub cluster 2

Fig. 5: Words (mainly 'jonatan') obtained from the two first clusters in Fig. 4. Here the number above the word is the word identification number, automatically assigned in the extraction process. Note: no. 37 is one example for the stenographic spelling of the word 'jonatan' in Fig. 2 (b)

Acknowledgements

This work is supported by the Riksbankens Jubileumsfond (RJ) (Dnr P19-0103:1).

References

1. Andersen, J.: Denna dagen ett liv. En biografi ver Astrid Lindgren, p. 180. Norstedts, Stockholm (2014)
2. Benato, B.C., Gomes, J.F., Telea, A.C., Falco, A.X.: Semi-automatic data annotation guided by feature space projection. Pattern Recognition **109**, 107612 (2021). https://doi.org/https://doi.org/10.1016/j.patcog.2020.107612, http://www.sciencedirect.com/science/article/pii/S0031320320304155
3. Bernard, J., Hutter, M., Zeppelzauer, M., Fellner, D., Sedlmair, M.: Comparing visual-interactive labeling with active learning: An experimental study. IEEE Transactions on Visualization and Computer Graphics **24**(1), 298–308 (2018). https://doi.org/10.1109/TVCG.2017.2744818
4. Bohlund, K.: Den okndа Astrid Lindgren, p. 173. Astrid Lindgren text, Stockholm (2018)
5. Bowles, H.: Dickens and the Stenographic Mind. Oxford University Press (2018)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 886–893 vol. 1 (2005). https://doi.org/10.1109/CVPR.2005.177
7. Fischer, A., Frinken, V., Fornés, A., Bunke, H.: Transcription alignment of latin manuscripts using hidden markov models. In: Proceedings of the 2011 Workshop on Historical Document Imaging and Processing. p. 2936. HIP '11, Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/2037342.2037348, https://doi.org/10.1145/2037342.2037348
8. Hast, A., Marchetti, A.: Improved illumination correction that preserves medium sized objects. Machine Graphics & Vision **23**(1/2), 3–20 (2014)
9. Hast, A., Mårtensson, L., Vats, E., Heil, R.: Creating an atlas over handwritten script signs. In: DHN 2019, March 6–8, Copenhagen, Denmark (2019)
10. LeCun, Y., Cortes, C., Burges, C.: Mnist handwritten digit database. AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist **2** (2010)
11. Leedham, C.G.: Computer acquisition and recognition of Pitman's handwritten shorthand. Ph.D. thesis, University of Southampton (1985)
12. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(86), 2579–2605 (2008), http://jmlr.org/papers/v9/vandermaaten08a.html
13. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction (2020)
14. Rothfeder, J., Manmatha, R., Rath, T.M.: Aligning transcripts to automatically segmented handwritten manuscripts. In: Bunke, H., Spitz, A.L. (eds.) Document Analysis Systems VII. pp. 84–95. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
15. Törnqvist, L.: Man tar vanliga ord: att läsa om Astrid Lindgren, p. 185f. Salikon, Liding (2015)

16. Trnqvist, L.: Rapport: Astrid lindgrens arkiv nya forskningsmjligheter. Barnboken **34**(2) (Nov 2011). https://doi.org/https://doi.org/10.14811/clr.v34i2.127, https://www.barnboken.net/index.php/clr/article/view/127
17. Vats, E., Hast, A., Singh, P.: Automatic document image binarization using bayesian optimization. In: Proceedings of the 4th International Workshop on Historical Document Imaging and Processing. p. 8994. HIP2017, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3151509.3151520, https://doi.org/10.1145/3151509.3151520
18. Wilkinson, T., Nettelblad, C.: Bootstrapping weakly supervised segmentation-free word spotting through hmm-based alignment. In: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 49–54 (2020), https://doi.org/10.1109/ICFHR2020.2020.00020
19. Zhai, F., Fan, Y., Verma, T., Sinha, R., Klakow, D.: A dataset and a novel neural approach for optical gregg shorthand recognition. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) Text, Speech, and Dialogue. pp. 222–230. Springer International Publishing, Cham (2018)