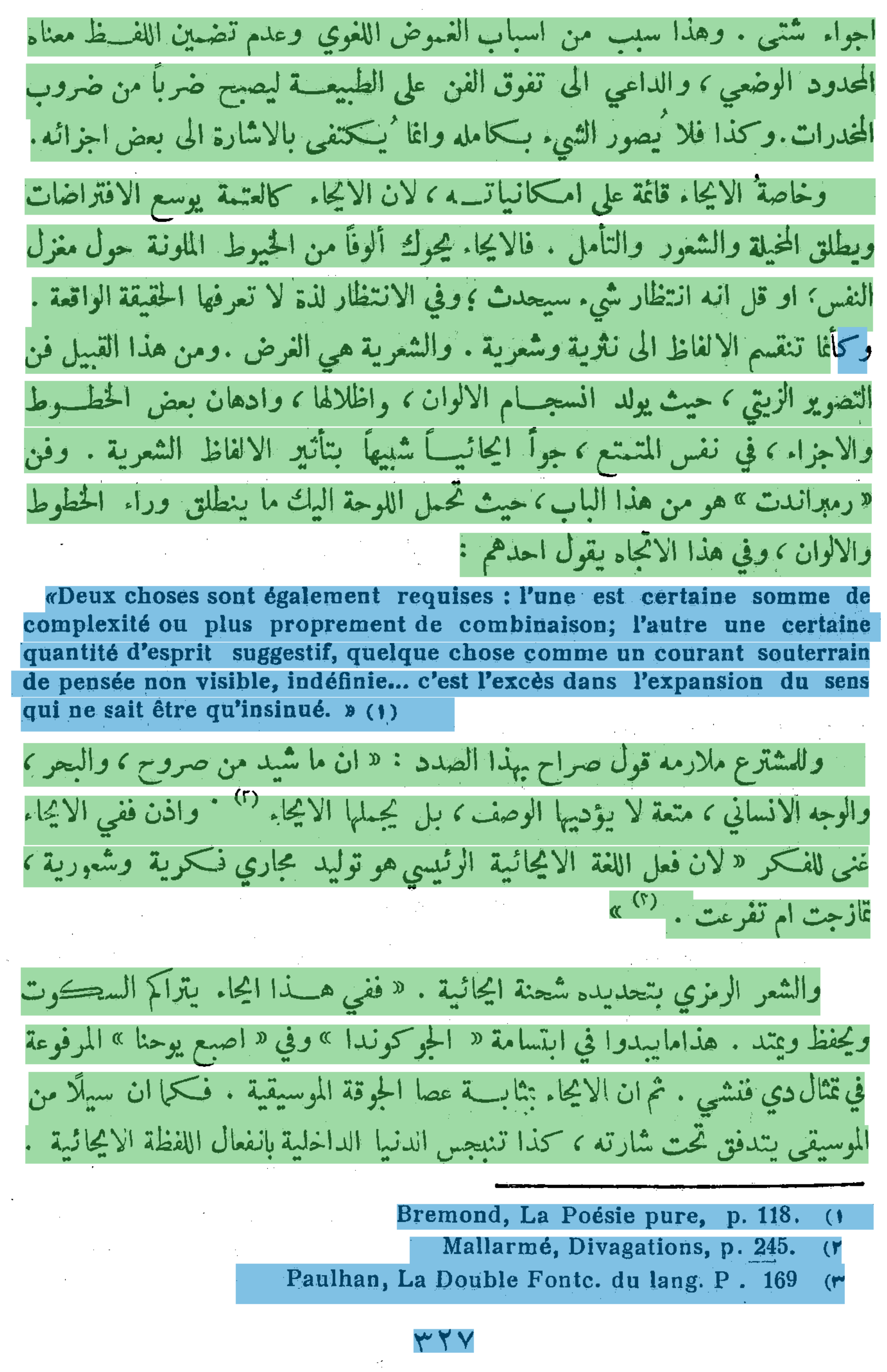
Script and Emphasis Detection using Recurrent Neural Networks

Benjamin Kiessling (EPHE, Université PSL & Leipzig University)
Daniel Kinitz, Christoph Gümmer & Parivash Mashhadi (Bibliotheca Arabica, Saxon Academy of Sciences)

Introduction

Parallel texts, mixed Fraktur-Antiqua printing, dictionaries, and library catalogs are of particular interest to much Digital Humanities research and often contain multiple scripts and semantic markup. With the increased availability of Optical Character Recognition software at least in part accessible to the determined DH scholar robust script and text emphasis detection methods are of special importance for effective digitization of these works.



اجواء شتى . وهذا سبب من اسباب الغموض اللغوي وعدم تضمين اللفـظ معناه
المحدود الوضعي ، والداعي الى تفوق الفن على الطبيعـة ليصبح ضرباً من ضروب
المخدرات. وكذا فلا يُصور الشيء بكامله وانما يُكتفى بالاشارة الى بعض اجزائه.

وخاصةُ الايحاء قائمة على امكانياتـه ، لان الايحاء كالعتمة يوسع الافتراضات
ويطلق المخيلة والشعور والتأمل . فالايحاء يحوك ألوفاً من الخيوط الملونة حول مغزل
النفس؛ او قل انه انتظار شيء سيحدث ؛ وفي الانتظار لذة لا تعرفها الحقيقة الواقعة .
وكأنما تنقسم الالفاظ الى نثرية وشعرية . والشعرية هي الغرض . ومن هذا القبيل فن
التصوير الزيتي ، حيث يولد انسجـام الالوان ، واظلالها ، وادهان بعض الخطـوط
والاجزاء ، في نفس المتمتع ، جواً ايحائيـاً شبيهاً بتأثير الالفاظ الشعرية . وفن
« رمبراندت » هو من هذا الباب ، حيث تحمل اللوحة اليك ما ينطلق وراء الخطوط
والالوان ، وفي هذا الاتجاه يقول احدهم :

«Deux choses sont également requises : l'une est certaine somme de complexité ou plus proprement de combinaison; l'autre une certaine quantité d'esprit suggestif, quelque chose comme un courant souterrain de pensée non visible, indéfinie... c'est l'excès dans l'expansion du sens qui ne sait être qu'insinué. » (١)

وللمشترع ملارمه قول صراح بهذا الصدد : « ان ما شيد من صروح ، والبحر ،
والوجه الانساني ، متعة لا يؤديها الوصف ، بل يحملها الايحاء (٢) . واذن ففي الايحاء
غنى للفكر « لان فعل اللغة الايحائية الرئيسي هو توليد مجاري فكرية وشعورية ،
تمازجت ام تفرعت . (٣) »

والشعر الرمزي بتحديده شحنة ايحائية . « ففي هـذا الايحاء يتراكم السكوت
ويحفظ ويمتد . هذاما يبدوا في ابتسامة « الجوكوندا » وفي « اصبع يوحنا » المرفوعة
في تمثال دي فنشي . ثم ان الايحاء بمثابـة عصا الجوقة الموسيقية . فكما ان سيلاً من
الموسيقى يتدفق تحت شارته ، كذا تنبجس الدنيا الداخلية بانفعال اللفظة الايحائية .

Bremond, La Poésie pure, p. 118. (١
Mallarmé, Divagations, p. 245. (٢
Paulhan, La Double Fontc. du lang. P . 169 (٣

٣٢٧

Figure 1: Script recognition on French-Arabic sample page

State of the art neural sequence-to-sequence models have largely supplanted older character-based methods for Optical Character Recognition. While neural methods have generally higher accuracy and decreased requirements on training data annotation depth, script and text emphasis detection approaches do not translate well to recent OCR engines.

Bibliotheca Arabica aims to gain new insights into Arabic literature from 1150 to 1850 CE by analysing the production, transmission, and reception of texts. The basis of this research are **~500, mostly multilingual, library manuscript catalogs with extensive semantic markup** in the form of text emphasis.

Related work

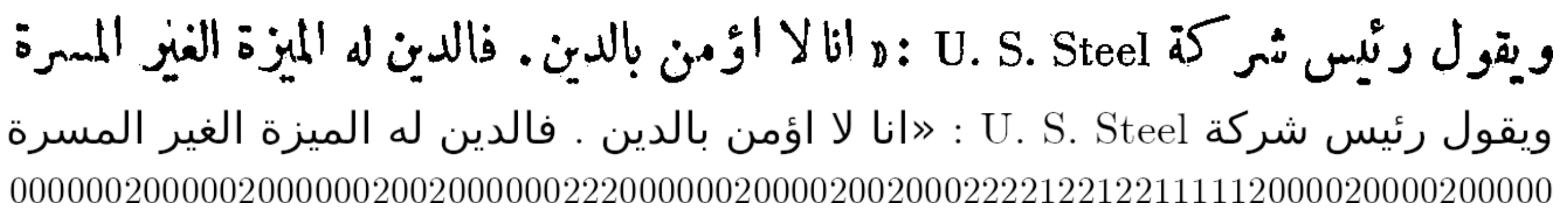
Past approaches to segmentation-less multilingual OCR have focused on building combined models capable of recognizing multiple scripts [3]. The irregularity of early modern printing and large number of typefaces result in **character accuracy below 95% for mixed-font models** even on mono-script texts [5], necessitates menial training data acquisition and retraining of these large models. Direct reuse of training data is also regularly prevented by a mixture of wanted and undesirable scripts.

The method described in [1] labeling whole lines using a recurrent neural network is inappropriate for many humanities texts because of extensive intra-line script switching. [4] published a conceptually simpler approach without feature extraction directly classifying character script using an LSTM network. A refined version of the latter method is the basis of our script detection system.

RNNs for Script and Emphasis Detection

Script Detection

The system treats script detection as a **segmentation-less sequence classification** problem. Instead of unique output labels per grapheme, all **code points of a particular script are assigned the same label** (figure 2) and the network is trained to generate the correct sequence of script labels using the CTC loss function [2]. On the face CTC is unsuitable for this task, as it includes no mechanism to ensure temporal alignment between graphemes in the input sequence and output activations; fortunately the network's activations are fairly close to their corresponding location in the input image. The output sequence is then used to split the line into single-script runs that can be classified with monoscriptual recognition models.



ويقول رئيس شركة U. S. Steel : «انا لا اؤمن بالدين. فالدين له الميزة الغير المسرة

ويقول رئيس شركة U. S. Steel : «انا لا اؤمن بالدين . فالدين له الميزة الغير المسرة

000000200000200000020020000002220000000200002002000222122122111112000020000200000

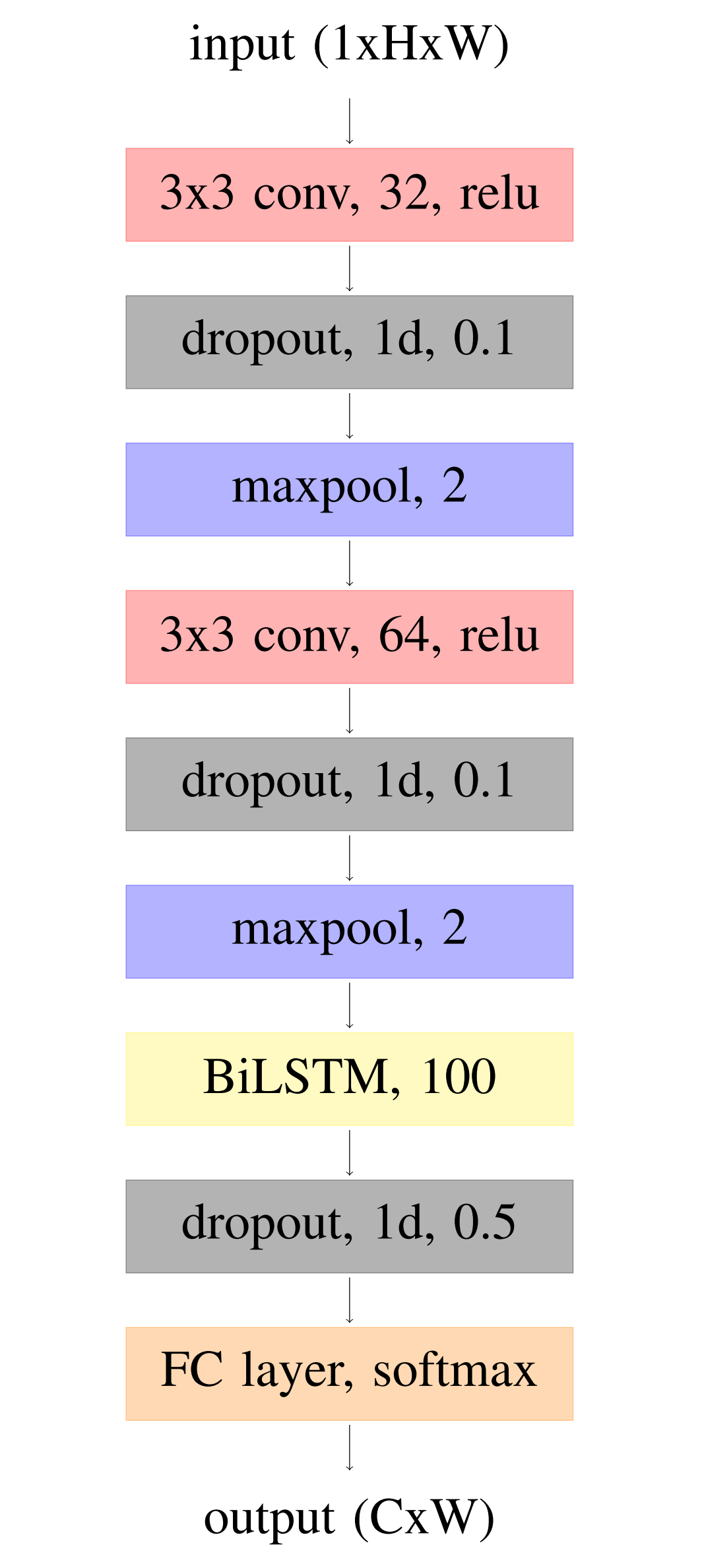
Figure 2: Modified ground truth (top: original line, middle: transcription, bottom: assigned script classes)

Script classes are derived from the Unicode script property in conjunction with ISO 15924 identifiers. **Graphemes occuring in multiple scripts such as numerals and punctuation are assigned a separate common class.** Merging these during post-processing based on their surrounding script increases classification robustness on non-body text such as page numbers and tables. Classification of bidirectional text is supported by applying the Unicode BiDi algorithm before script assignment.

Two additional post-processing steps are performed. First individual runs are replaced by the whole line bounding box if only a single script remains after common/inherited merging. A second step increases accuracy through a user-supplied whitelist of allowed scripts, merging invalid runs into the surrounding context after filtering for common confusions.

Emphasis Recognition

We evaluated two methods of encoding two common text emphasis methods for recognition by a neural network. Initially, **italicized and text components with increased letter spacing** were marked up with special **start and stop markers**. While the results of the training were promising, obtaining the amount of training data needed to get the network to reliably place both markers was infeasible for our target documents. Creating **separate labels for italicized/spaced graphemes** and training for these, remedied the marker placement issue with a sufficiently small amount of training data.



input (1xHxW)
3x3 conv, 32, relu
dropout, 1d, 0.1
maxpool, 2
3x3 conv, 64, relu
dropout, 1d, 0.1
maxpool, 2
BiLSTM, 100
dropout, 1d, 0.5
FC layer, softmax
output (CxW)

Figure 3: Network architecture (H: sequence height, W: sequence length, C: alphabet size)

Architecture

Both the script detection and emphasis recognition share a common network architecture of a small **convolutional block followed by a bidirectional LSTM layer** and a final linear projection with softmax activation as shown in figure 3. The network is trained with **CTC loss and single-sample stochastic gradient descent with momentum** (learning rate: 0.0001, momentum: 0.9). **Early stopping** is used to terminate training. The system is implemented as part of the kraken OCR engine.

The networks operate on binarized whole lines. **Baselines and line height are normalized** using a slightly modified version of the centerline normalizer implemented in the OCRopus system. Labels and locations are extracted from the $C \times W$ output matrix with a simple **greedy decoder**.

Results

Dataset

We repurposed publicly available non-synthetic training data for recognition models to build a corpus of **85000** script-annotated line images containing **Arabic, Cyrillic, polytonic Greek, Hebrew, Latin, Fraktur, and (western) Syriac** text. The majority of text lines contain only a single non-common script although there are mixed lines for all scripts in the corpus. The exact distribution of code points is shown in table 4. 850 randomly selected lines are separated from the corpus as a test set.

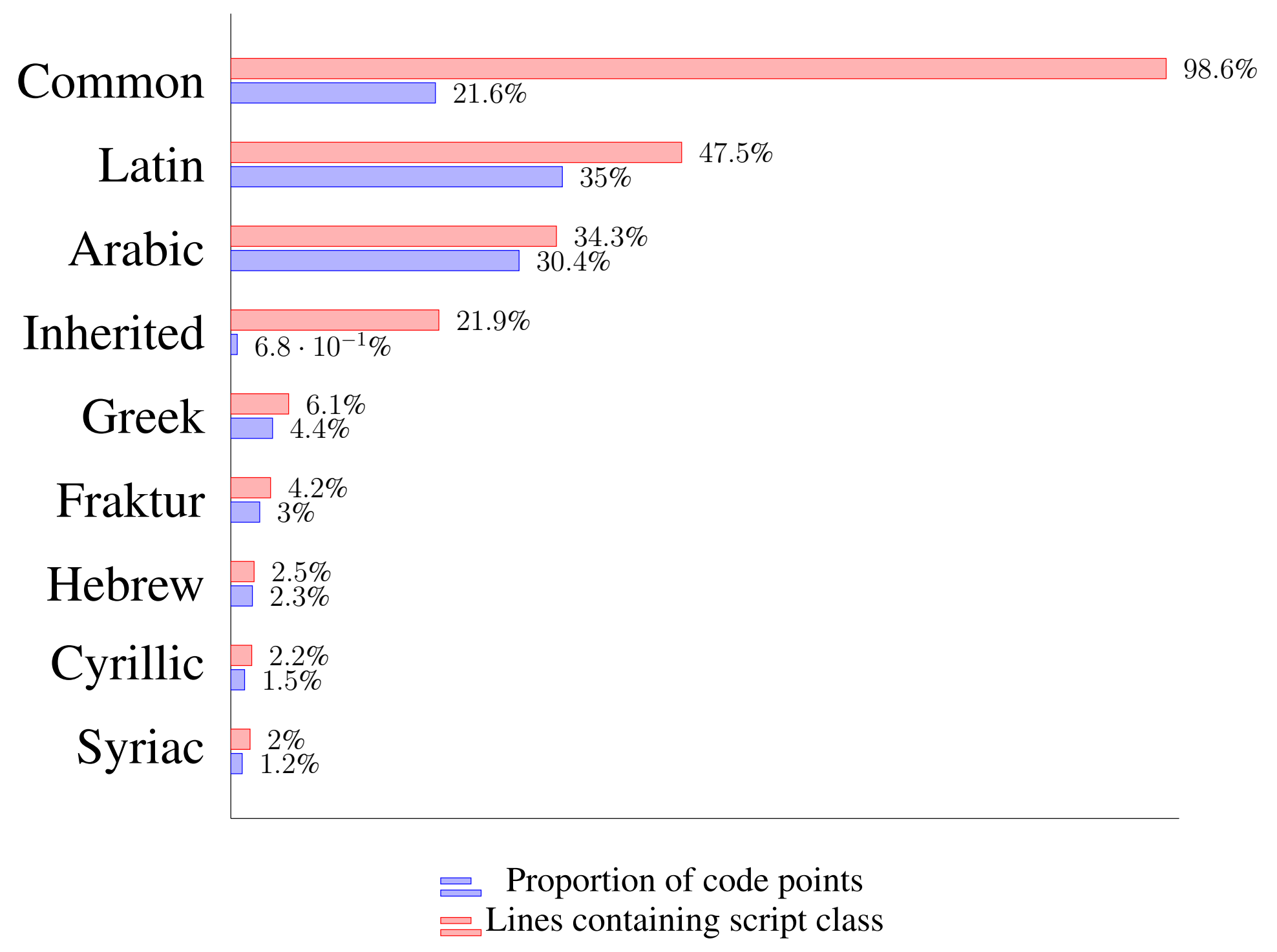


Figure 4: Script detection training data distribution

Emphasis recognition was evaluated on an **English and romanized Arabic catalog on a set of 350 transcribed lines**. An additional 50 line transcriptions were used as a test set. Overall 220 lines contain some kind of text emphasis.

Script Detection

The fully trained network yielded a **character accuracy of 94.62%** on the test set. Output for a French-Arabic bilingual sample page can be seen in figure 1. The misclassification of Eastern Arabic numerals as Latin text is caused by the transcription as Latin Arabic numerals.

Emphasis Recognition

The average character accuracy of the trained model over 10 runs is **99.3%** ($\sigma = 0.16$) with **95.38% on cursive and text with increased spacing** ($\sigma = 1.46$). When using only emphasized text accuracy as the stopping criterium mean accuracy rises to **99.03%** ($\sigma = 0.28$).

References

[1] Yasuhisa Fujii et al. "Sequence-to-Label Script Identification for Multilingual OCR". In: *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*. Vol. 1. IEEE. 2017, pp. 161–168.

[2] Alex Graves et al. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks". In: *Proceedings of the 23rd international conference on Machine learning*. ACM. 2006, pp. 369–376.

[3] Adnan Ul-Hasan and Thomas M Breuel. "Can we build language-independent OCR using LSTM networks?" In: *Proceedings of the 4th International Workshop on Multilingual OCR*. ACM. 2013, p. 9.

[4] Adnan Ul-Hasan et al. "A sequence learning approach for multiple script identification". In: *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*. IEEE. 2015, pp. 1046–1050.

[5] Uwe Springmann, Florian Fink, and Klaus U Schulz. "Automatic quality evaluation and (semi-) automatic improvement of mixed models for OCR on historical documents". In: *CoRR. URL: https://pdfs. semanticscholar. org/b8b7/c369a01164b289ed7c41ef33ce1b74e0fb1f. pdf* (2016).