# How to Identify Speech When Translating Unpunctuated Poetry

*Timo Baumann*[1], *Burkhard Meyer-Sickendiek*[2], *Hussein Hussein*[2]

[1]*Department of Informatics, Universität Hamburg, Germany*
[2]*Department of Literary Studies, Free University of Berlin, Berlin, Germany*
*baumann@informatik.uni-hamburg.de, {bumesi,hussein}@zedat.fu-berlin.de*

## Abstract

A large proportion of (post)-modern poetry contains no or hardly any punctuation. In our contribution, we will investigate how well punctuation information can be recovered for post-modern poetry based on the information contained in the text and speech of free verse poems. We use the world's largest corpus of spoken (post-)modern poetry from our partner *lyrikline* which contains the corresponding audio recording of each poem as spoken by the original author and features translations for many of the poems. We identify lines that contain a phrase break in the middle of the poetic line, which may already be helpful for philological analysis on one hand, and identify the position of the break in the line on the other hand. We select those poetic lines that contain one or more punctuation characters that typically indicate a phrase break in poetry (`.,;:!?/`) somewhere in the middle (rather than only at the end of the line) as our target class. We train a neural network (bidirectional recurrent neural network (RNN) based on gated recurrent units (GRU) with attention) that combines audio and textual features to identify the punctuation with the goal of applying it to reconstruct them within a corpus of unpunctuated poems. Our results clearly indicate that speech is helpful for recovering the constituency structure of post-modern poetry that is partially obfuscated by missing punctuation.

## 1 Introduction

One of the great challenges in the translation of poetry in general is the translation of poems that do not use any punctuation, a stylistic phenomenon, that can be found very often in modern and post-modern poetry. The problem for the translator is the appropriate recognition of phrase boundaries, as becomes obvious with regards to the poem “Die Farne lappen in den Fluß” (english: The ferns slip into the river) written by the Swiss poetess Ilma Rakusa [1] (emphasis added for clarity):

Die Farne lappen in den Fluß
der trüb durch Böhmen fließt
*das Mückengift verstört*
*die Hunde mit den langen Zungen*
*hecheln durch das Gras*
ein Wanderer liegt
und aus den Schloten
der Papierfabrik
steigt schnelles Gas

The challenge for the interpretation of this poem lies in the adequate identification of the third verb “verstört” (english: to unsettle or confuse). If this is identified as transitive, the sentence

continues from line 3 to line 4. If it is identified as intransitive, the sentence ends with line 3. The spoken rendition of this poem by the original author shows a clear caesura or break after the end of the third line: The phrasing of the sentence ends, so the verb “verstört” has an intransitive meaning. The poem was translated by Andrew Winnard into English as follows [2]:

The ferns slip into the river
mud flowing through Bohemia
*the mosquito bites confuse*
*the long-tongued dogs*
*panting through the grass*
a hiker lying down
and the paper mill's
chimney stacks
breathe out their ragged gas

In the translation – where the translator obviously lacked or chose to ignore the audio information – this meaning is not recognized. This becomes obvious when looking at line 5: the verb “hecheln” (english: to pant) is translated as a participle (panting), which indicates that the translator ignored the fact that “hecheln” is a full verb within this poem, with “the dogs” as subject.

In this paper, we develop a classifier to detect phrase boundaries in the middle of the line. For this purpose, we distinguish two basic forms in a large corpus of auditory poems: a so-called line style (German: Zeilenstil) and a so-called hook style (German: Hakenstil) [3][4, p. 301]. These two terms were coined by the literary scholar Eduard Sievers for Old Germanic and later transferred to Middle High German by the literary scholar Andreas Heusler. Line style means the relationship between sentence and poetic line in which the end of the sentence coincides with the end of the poetic line, so that both periods coincide. The line thus represents a syntactic unit. If line and sentence coincide in every line of the poem, then we are dealing with a continuous line style. The opposite is the hook style, in which the consecutive lines are interlocked by enjambments. So the sentence goes beyond the end of the line, and ends either in the middle of the next line or at its end, or continues to a new line. This breaks up the syntactic unit of the long line and causes the periods to overlap. The result is a rather restless prosodic phrasing: where the line ends, the unfinished sentence continues, where the sentence ends, the line continues. Such an overlapping of the sentence and the long line gives the poem's movement something restless: the end of the line does not provide a rest point because the sentence continues; the end of the sentence does not provide a rest point either because the verse pushes forward. The hook style is thus based on enjambments that continuously allow a unit of sentence or sense to extend beyond the end of a poetic line to the following poetic line.

Extending beyond hook style and line style, and in order to classify spoken poems as found on our partner webpage *lyrikline* (`www.lyrikline.org`), we identify 8 possible variations for the interplay of line and phrasing, as can be seen in Figure 1, based on whether a line starts a new phrase, ends the phrase, and whether it consists of exactly one or of multiple phrases. We assume that postmodern poetry extended these two fundamental styles, i.e. line and hook style, by inventing and using several of these variations. Only in rare cases – such as the early poetry Ezra Pounds or the poems of Expressionism – does the poem consist of exactly one phrase per line (1). In many cases, one phrase can continue beyond the line (lines 2, 3, 6, 7), and consequentially continues from the previous line (3, 4, 7, 8) and, what we will focus on in this paper, a line may be broken into two (or more) phrases (4, 5, 6, 7). This latter phenomenon is called hook style and is particularly relevant for the understanding and translation of post-modern poetry, to avoid the erroneous interpretation of the line break rather than the phrase break as guiding the syntax-based analysis. Regarding our first example of a poetry translation, we can apply this scheme as follows: In the source (German) poem, the nine lines could be identified

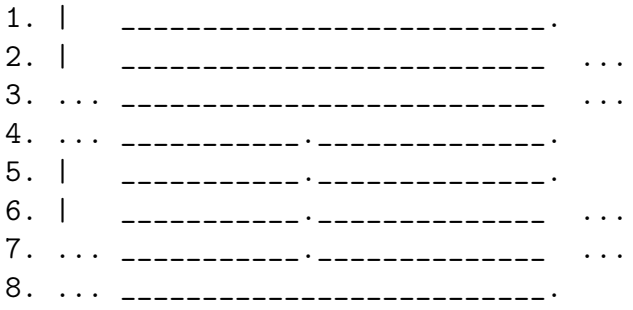


**Figure 1** – The $2^3 = 8$ possible combinations of phrase and line structures in a poem, based on whether the beginning of the line coincides with the beginning of a phrase, whether the end of the line coincides with the end of a phrase, and whether the line is broken into (2 or more) phrases or not. Keep in mind that these differences are based on the interplay between a line and its previous and following line.

as 1-1-1-2-1-2-2-2-1. In the translation or target poem, on the other hand, there is a difference: 1-1-2-2-2-1-2-2-1. Obviously, this difference is caused by the missing punctuation as the fourth line in the poem has been misidentified as type 7 instead of type 2 (and the resulting syntax issue in line 5 has not been noticed).

As a result (and in particular given the line-structure of poetry as described above), it can become difficult to correctly infer the constituency structure of a poem which is highly important for text understanding or machine translation. Many modern and post-modern poets leave out all punctuation in their poems, basically due to stylistic reasons. However, when reading out their poems, it becomes clear from their pronunciation that they nevertheless intend a break. We therefore analyze how well the missing interpretation can be 'recovered' from the prosodic realization of poetry on a large corpus of spoken poetry.

We have previously studied in [5] how to identify *enjambments*, i.e., when a phrase continues beyond the line, as in hook style. In terms of the scheme in Figure 1, enjambments occur in the continuation from line 2 to 3, 3 to 4, 6 to 7, and 7 to 8. We have shown that automatic classification can be trained to perform with similar accuracy as human annotators.

In this paper, we focus on punctuated poetry as the basis for our analysis as this enables us to read off the phrasing from the given punctuation. We expect that our results on punctuated poetry will carry over to unpunctuated poetry with little performance loss. This information would be required (explicitly or implicitly) for a poetic machine translation system that is to successfully translate despite the lack of punctuation in post-modern poetry. We aim at a classification of within-the-line breaks of phrasing (i.e. the lines 4,5,6,7 vs. 1,2,3,8). This task is more challenging as the phrase break can occur anywhere in the line (whereas an enjambment by definition occurs at the end of the line). In addition to classifying whether a line contains a phrase-break, we hence also investigate how we can identify the position of the phrase-break within the line. We contribute (a) an architecture that relates speech and text in such a way that acoustic features can be related to textual material with the goal of detection based on both sequences, (b) an application of this architecture towards the problem of identifying left-out punctuation in post-modern poetry, and (c) an analysis of post-modern spoken poetry towards a deeper understanding of machine translation of such linguistic material.

The remainder of this paper is organized as follows: Section 2 provides an overview about database and pre-processing steps. Section 3 illustrates the modeling of lines for the purpose of identifying the phenomenon in question using neural networks. The experiments and results are described in Section 4 and finally, conclusions and future works are presented in Section 5.

## 2 Data and Pre-processing

We used the data from the webpage of our partner *lyrikline* in the project *Rhythmicalizer* (`www.rhythmicalizer.net`). *Lyrikline* houses contemporary international poetry as texts (original versions and translations) and the corresponding audio files. All the poems are read by the original authors. Altogether there are 230 german-speaking poets (including Germany, Switzerland, and Austria) on the *lyrikline* webpage reading a total of 2,581 poems. A total of 733 German poems are translated into English and are hence particularly interesting for our long-term goal of investigating poetic translations.

Given the nature of modern and post-modern poetry, substantial parts of the corpus are 'unpunctuated' or at least highly 'under-punctuated' as compared to what could be expected. Roughly one third of the poems fall into this category; some other poems do not have any lines with internal phrase-breaks, which we presently discard as well, leaving us with 1,763 poems for analysis that contain lines of all the types 1-8 (see Figure 1). We focus on these punctuated poems and select those lines that contain one or more punctuation characters (`.,;:!?/`) that typically indicate a phrase break in poetry somewhere in the middle (rather than only at the end of the line) as our target class. Our process yields a total of 40,019 lines out of which 16,648 (42 %) contain one or multiple line-internal punctuation characters indicative of a phrase break. In addition, our goal is to identify the position of the break in lines that contain a phrase break. For this, we investigate only those 11,589 lines that contain exactly one line-internal punctuation mark and discard the 5,059 with multiple marks. We record the position (in characters) of the internal punctuation within the poem. We validated the heuristics that punctuation characters indicate phrasing on 30 poems and find it to hold in about 75 % of cases.

A text-speech alignment for the written poems and spoken recordings is implemented. We perform forced-alignment of text and speech for the poems using the text-speech aligner published by [6] which uses a variation of the SailAlign algorithm [7] implemented via Sphinx-4 [8]. The alignments are stored in a format that guarantees the original text to remain unchanged which is important to be able to recreate the exact white-spacing in the poem and would be helpful when adding further annotations to the poem in the future. We extract the line-by-line timing (start of first word and end of last word in the line) for each poetic line. The purpose of this process is to detect the boundaries of poetic lines, since we process every line alone.

## 3 Experiment and Model

We perform our experiments with the data described in Section 2. We did not use text data from other areas, because it does not represent the problem of punctuation of poems. The model must deal effectively with data sparsity, since there are a broad variety and a relatively small number of poems in the experiment. Therefore, we use as few free parameters as possible that need to be optimized during training. For this reason, in textual processing we focused on character-by-character encoding of poetic lines (and using character embedding). The model is not trained using an explicit notion of words. Instead, it may implicitly encode word-level information via the constituting sequences of characters. We use recurrent neural network (RNN)-based models with attention as the basis of our models. Attention (a) may improve the model's representations and hence yield better performance (although some initial testing did not show a large impact), and (b) can be observed during the application of the model and gives an indication of what the model pays attention to, and can be discussed with regards to its philological plausibility. In particular, we train our models on data that contains punctuation within the line which we take as indicators of phrase breaks for the two tasks (for a full system, the two sub-tasks have to be executed in sequence):

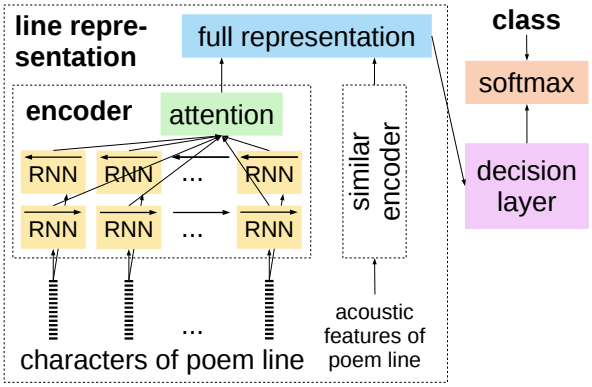


**Figure 2** – Base model for classifying lines with two encoders, one for text and one for audio (as in [5]).

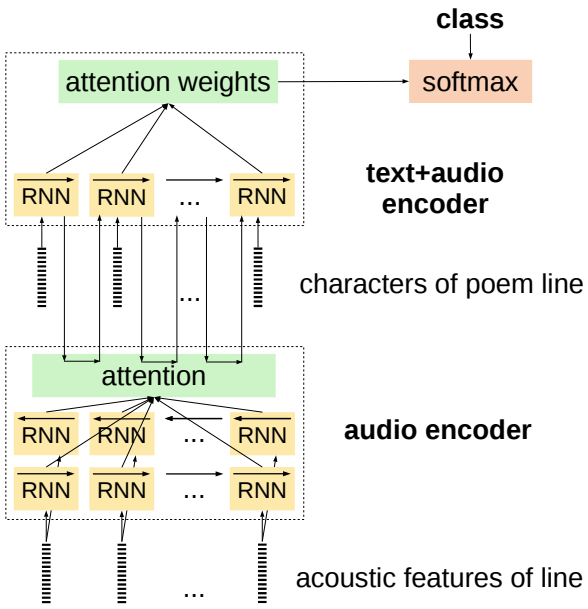


**Figure 3** – Extended model using an *encoder-encoder* model that relates audio to text via inner attention.

- **Task 1:** Identification of lines that contain one or more constituency breaks in the middle of line (types 4,5,6,7 in Figure 1) vs. lines that do not contain such a break (types 1,2,3,8).
- **Task 2:** Identification of the position of the constituency break in those lines that consist of two phrases.

For identifying lines that consist of multiple phrases, we use a model similar to the one previously used for finding enjambment lines [5], encoding both the text and the audio and concatenating these together, as shown in Figure 2, however disregarding the pause features that follow the line. The model uses bi-directional recurrent encoders based on gated recurrent units (GRU) [9] that are followed by *inner attention* [10] to help single in on the most informative portion of the line (and audio), most likely in the vicinity of the phrase break(s). Note that this is similar to the architecture used in [11].

Finding the position of the phrase break in the line is far more challenging, of course, given that the break could be at any position in the line (in contrast to the binary problem of whether there is a break or not). Given the challenging and multifaceted textual material, we expect relatively low performance with a solution that uses the text only, and we expect little help from adding audio information via a separate encoder (as in Figure 2) because the separation will hinder the model to learn from the combination of text and audio. We thus propose an extended *encoder-encoder* model as outlined in Figure 3. In this architecture, we first encode all the audio information (bidirectionally via GRUs, lower block of the figure). The text is encoded via its

**Table 1** – Results of classification for both tasks under various conditions.

| task | condition | correct | f-measure |
|---|---|---|---|
| 1: identification of lines with a break vs. lines without breaks | text-only | 58.4 % | 0.43 |
| | text+audio | 58.3 % | 0.43 |
| | text+audio (via encoder-encoder) | 63.6 % | 0.56 |
| identification of enjambments (from [5]) | text+audio | — | 0.91 |
| 2: identification of break position within the line | text-only | 65.6 % | — |
| | text+audio (via encoder-encoder) | 67.8 % | — |

characters and GRU units which take as additional input the attended-to output from the audio encoder, thus conditioning the text on the way that it is spoken. The attention layer not only uses inner attention but also conditions on the textual RNN's state and we expect that it will be able to learn the relation of acoustic features to textual form, thereby combining textual and (corresponding) acoustic material in the same encoding. We then compute attention weights over the state sequence which we optimize via softmax to yield the one position to which the phrase break should be added.

For both tasks, text is encoded via character-based 20-dimensional embeddings that are learned during the task, and speech via Mel-frequency cepstral coefficients (MFCC) and fundamental frequency variation (FFV) [12] vectors that are aggregated (mean and stddev) for every 100 ms. Our RNNs use GRUs, 2 layers and a 20-dimensional hidden state. Attention is 20-dimensional as well. Obviously, these meta-parameters have not extensively been optimized for the task but have proven to be reasonable defaults in a number of applications in our digital humanities domains. Given the relative sparsity of our data and for better robustness, we perform 10-fold cross-validation. Our network is implemented in DyNet [13].

## 4 Results

The results of our experiments are shown in Table 1. As can be seen in the table, the identification of lines with a constituency break (task 1) is challenging and in particular much more so than the identification of enjambments, i. e., whether no phrase break occurs at the end of the line (as reported in [5]), which yields a very high f-measure of 0.91 although based on manual annotations, whereas the correctness and f-measure by using only text data are 58.4 % and 0.43, respectively. The reasons for this go beyond the manual annotation (which would exclude mistakes such as proposing a phrase-break within a date or time because it contains a full stop or colon) and is most likely based in the fact that the phrase-break can be anywhere in the line vs. is known to be at the end of the line for enjambments. We also find that integrating acoustic information via a second (multi-)encoder does not help identify constituency breaks within the line but that the integration in the *encoder-encoder* model yields a substantial improvement (correctness and f-measure are 63.6 % and 0.56, respectively). Thus, we conclude that the *encoder-encoder* model is successful at relating the textual and acoustic information (which the multi-encoder model is not able to do).

For those lines that hold a constituency break, we looked at the placement of the break within the line (task 2). Our method yields the correct position of the punctuation for more than ⅔ of the lines (correctness is 67.8 % with text and audio features by using *encoder-encoder*). We do not report f-measure for this task as the number of reasonable phrasings differs with every line (as it depends on the number of words in the line). To our surprise, audio information is of relatively little use in this scenario, possibly because the model already tunes in successfully to upper-casing and other textual phrase-break phenomena.

## 5 Conclusion and Future Works

In this paper, we have tackled the problem of identifying the phrasal structure that can be found in post-modern poetry, and we have focused on line-internal constituency breaks, which have been under-studied in the digital humanities so far, as compared to line-based phenomena such as enjambments [5] and the resulting rhythmical classification of poems [14]. We have shown that lines can be identified as containing line-internal breaks with relatively high confidence and that it is possible to also find the position of a break within such a line. It is helpful to not only rely on the textual information but to include acoustic evidence of phrasing, and, in particular for the break placement, to relate the acoustic with the textual information. Our *encoder-encoder* model appears to work well for this task.

The paper offers a new way to use phrasing and audio information within machine translations, based on current discussions within translation theory: Lawrence Venuti criticized the modern call for an utmost "fluent" and "transparent" target language, which erases the foreignness and alterity of the source-text in order to appear almost "readerly", and developed a far more individual idea of translation [15]. To keep the foreignness of the source-text, Barbara Folkart called for rather "writerly" translations preserving the performativity, the rhythm, the sound play, the elliptic phrases, and mostly disfluent patterns of modern and post-modern poetry [16], other theorists like Boase-Beier or Scott have shared this call [17][18]. The translation in such a "writerly" fashion needs an awareness of the poet's phrasing, as Folkart pointed out: "Rhythmical valency is a measure of the way the poem "breathes," overriding its metrical grid (the foot, the line, and even the stanza), phrasing what it is articulating" [16]. Our paper offered first steps towards the identification of such a phrasing by using neural networks. We hope to do future research on this topic with a special focus on audiovisual translations using neural networks, especially with regards to Film-dubbings.

## 6 Acknowledgements

The work is funded by the Volkswagen Foundation in the announcement 'Mixed Methods in the humanities? Funding possibilities for the combination and the interaction of qualitative hermeneutical and digital methods' (funding codes 91926 and 93255).

## References

[1] RAKUSA, I.: *Ein Strich durch alles. Neunzig Neunzeiler*. Suhrkamp Verlag, 1997.

[2] RAKUSA, I.: *A Farewell to Everything: Ninety Nine-liners. Translated by Andrew Shields and Andrew Winnard*. Shearsman Books, 2005.

[3] BEYSCHLAG, S.: *ZEILEN- UND HAKENSTIL. Beiträge zur Geschichte der deutschen Sprache und Literatur (PBB)*, 56, pp. 225–313, 1932.

[4] BURDORF, D., C. FASBENDER, and B. MOENNIGHOFF: *Metzler Lexikon Literatur. Begriffe und Definitionen*. Metzler, 2007.

[5] BAUMANN, T., H. HUSSEIN, and B. MEYER-SICKENDIEK: *Analysing the focus of a hierarchical attention network: The importance of enjambments when classifying post-modern poetry*. In *Proceedings of Interspeech*. Hyderabad, India, 2018. doi:10.21437/Interspeech.2018-2530.

[6] BAUMANN, T., A. KÖHN, and F. HENNIG: *The Spoken Wikipedia Corpus Collection: Harvesting, Alignment and an Application to Hyperlistening*. *Language Resources and Evaluation*, 2018. doi:10.1007/s10579-017-9410-y.

[7] KATSAMANIS, A., M. BLACK, P. G. GEORGIOU, L. GOLDSTEIN, and S. NARAYANAN: *SailAlign: Robust Long Speech-Text Alignment*. In *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*. 2011.

[8] WALKER, W., P. LAMERE, P. KWOK, B. RAJ, R. SINGH, E. GOUVEA, P. WOLF, and J. WOELFEL: *Sphinx-4: A Flexible Open Source Framework for Speech Recognition*. Tech. Rep., Mountain View, CA, USA, 2004.

[9] CHO, K., B. VAN MERRIENBOER, C. GULCEHRE, D. BAHDANAU, F. BOUGARES, H. SCHWENK, and Y. BENGIO: *Learning phrase representations using rnn encoder–decoder for statistical machine translation*. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734. Association for Computational Linguistics, Doha, Qatar, 2014.

[10] LIU, Y., C. SUN, L. LIN, and X. WANG: *Learning Natural Language Inference using Bidirectional LSTM model and Inner-Attention*. *CoRR*, abs/1605.09090, 2016. URL `http://arxiv.org/abs/1605.09090`. `1605.09090`.

[11] BANERJEE, S. and K. TSIOUTSIOULIKLIS: *Relation extraction using multi-encoder lstm network on a distant supervised dataset*. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pp. 235–238. 2018. doi:10.1109/ICSC.2018.00040.

[12] LASKOWSKI, K., M. HELDNER, and J. EDLUND: *The fundamental frequency variation spectrum*. In *Proceedings of FONETIK 2008*. 2008.

[13] NEUBIG, G., C. DYER, Y. GOLDBERG, A. MATTHEWS, W. AMMAR, A. ANASTASOPOULOS, M. BALLESTEROS, D. CHIANG, D. CLOTHIAUX, T. COHN, K. DUH, M. FARUQUI, C. GAN, D. GARRETTE, Y. JI, L. KONG, A. KUNCORO, G. KUMAR, C. MALAVIYA, P. MICHEL, Y. ODA, M. RICHARDSON, N. SAPHRA, S. SWAYAMDIPTA, and P. YIN: *Dynet: The dynamic neural network toolkit*. *arXiv preprint arXiv:1701.03980*, 2017.

[14] BAUMANN, T., H. HUSSEIN, and B. MEYER-SICKENDIEK: *Style Detection for Free Verse Poetry from Text and Speech*. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*. Santa Fe, New-Mexico, USA, 2018.

[15] VENUTI, L.: *The Translator's Invisibility*. Routledge, 1994.

[16] FOLKART, B.: *Second Finding: a Poetics of Translation*. University of Ottawa Press, 2007.

[17] BOASE-BEIER, J.: *A Critical Introduction to Translation Studies*. Continuum, 2011.

[18] SCOTT, C.: *Literary translation and the rediscovery of reading*. Cambridge University Press, 2012.