How to Identify Elliptical Poems within a Digital Corpus of Auditory Poetry

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

[1]*Department of Literary Studies, Free University of Berlin, Berlin, Germany*
[2]*Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA*
hussein@zedat.fu-berlin.de, bumesi@zedat.fu-berlin.de, tbaumann@cs.cmu.edu

Abstract: Ellipses denote the omission of one or more grammatically necessary phrases. In this paper, we will demonstrate how to identify such ellipses as a rhythmical pattern in modern and postmodern free verse poetry by using data from *lyrikline* which contain the corresponding audio recording of each poem as spoken by the original author. We present a feature engineering approach based on literary analysis as well as a neural networks based approach for the identification of ellipses within the lines of a poem. A contrast class to the ellipsis is defined from poems consisting of complete and correct sentences. The feature-based approach used features derived from a parser such as verb, comma, and sentence ending punctuation. The classifier of neural networks is trained on the line level to integrate the textual information, the spoken recitation, and the pause information between lines, and to integrate information across the lines within the poem. A statistic analysis of poet's gender showed that 65% of all elliptical poems were written by female poets. The best results, calculated by the weighted F-measure, for the classification of ellipsis with the contrast class is 0.94 with the neural networks based approach. The best results for classification of elliptical lines is 0.62 with the feature-based approach.

1 Introduction

Ellipses denote the omission of one or more grammatically necessary phrases or words, the effect of which is supplied by context, for example: “She loves him and he her”. Their preform of such an ellipsis is the asyndetic sequence like in the sentence “Ich mag lieber Lyrik, du lieber Prosa” (english: I like poetry better, you prose), which would be complete: “Ich mag lieber Lyrik, während du lieber Prosa magst” (english: I like poetry better whereas you like prose better). In the second case, the two subsections are strung together syndetically and related by the conjunction “während” (english: whereas). In the first case, however, ranked asyndetically because the connective conjunction is missing. According to Alexander Polykarpov [1], asyndetic propositional structures often occur in the spontaneous spoken language, while syndetic propositions are more common in written language. Traditionally thought of as a grammatical trope (i.e. it operates on sentence level syntax), the term ellipsis is often applied more broadly today to refer to missing and assumed parts of stories, arguments and trains of thought. In this way, Otto Lorenz [2] identified the elliptical lines in poems of authors such as Friedrich Hölderlin, Rainer Maria Rilke, or Paul Celan as “indexical signs” for something which cannot be identified and represented in language, for example, the “transcendence of God” (Hölderlin), the “excess of personal memory” (Rilke), or the “incomprehensible suffering” (Celan). The elliptic spelling of these authors serves this lyrical silence. At the same time, it refers to the secrecy that must or at least can be articulated by the reader afterwards. Lorenz describes this interaction as deictic-elliptical writing. Ellipses may have the effect of a writer's taking the reader into his/her confidence through shared knowledge of a missing word or words.

However, we are pursuing a genuinely syntactic idea of the ellipsis. So we only pay attention to those parts that are missing in the sentence, without supplementing this absence with a higher meaning, as Lorenz has done by combining deixis and ellipsis. To give a first example, we will refer to Friederike Mayröcker's poetry. The literary scholar Beda Allemann identified the elliptical expressions in Mayröcker's poetry by the term "association fugue" [3, pp. 317], claiming that individual elliptical sentences, sentence remnants and words are combined by non-causal motivated connections. As an example, Mayröcker's poem "Was brauchst du" [4] (english: what do you need) is quoted below. Adding the missing phrases and punctuation in parentheses to the original version of the poem, it is easier to locate and identify the ellipses in the complementary text.

"was brauchst du? **(Du brauchst)** einen Baum **(und)** ein Haus **(, um)** zu
ermessen wie groß **(oder)** wie klein das Leben als Mensch **(ist.)**
wie groß **(oder)** wie klein **(ist das Leben,)** wenn du aufblickst zur Krone
(, oder wenn du) dich verlierst in grüner üppiger Schönheit **(.)**
wie groß **(oder)** wie klein **(ist das.)** bedenkst du **(,)** wie kurz
dein Leben **(ist?)** vergleichst du es mit dem Leben der Bäume **(?)**
du brauchst einen Baum **(und)** du brauchst ein Haus
(, du brauchst) keines für dich allein **(, du brauchst)** nur einen Winkel **(und)** ein Dach
(, um dort) zu sitzen **(,)** zu denken **(,)** zu schlafen **(,)** zu träumen
(,) zu schreiben **(und)** zu schweigen **(. Oder um)** zu sehen **(, z. B.)** den Freund**(,)**
die Gestirne **(,)** das Gras **(,)** die Blume **(und)** den Himmel"

The poem was translated by Rosmarie Waldrop to english as follows:

"what do you need? **(You need)** a tree **(and)** a house **(in order)** to
gauge how great **(or)** how small **(is)** our human life **(.)**
how great **(or)** how small **(is life)** when you look up to the top of the tree
and get lost in the lush luxuriant green**(.)**
how great **(or)** how small **(is it.)** when you think **(about)** how short
your life **(is,)** compared with the life of trees**(.)**
you need a tree **(and)** you need a house
(You do) not **(need them)** all for yourself **(,you)** just **(need)** a corner **(and)** a roof
to sit**(,)** to think**(,)** to sleep**(,)** to dream**(,)**
to write **(and)** to be silent **(. Or)** to see your friend**(,)**
the stars**(, the)** grass**(, the)** flower **(and the)** sky"

We will focus on poems read out by the original authors. We assume that such elliptic poems may indicate a certain alienation of the spoken language, as Nick Piombino's theory of "Aural Ellipsis" [5] indicated. Piombino uses the term "aural ellipsis" to denote the use of sounds beyond their usual meaning, as found in modern and postmodern poems. The term "aural ellipsis" defines effects of indeterminacy in the writing, reading and listening of contemporary poetry, which also means the effect of the "acoustic ellipsis" in auditory poems.

In this paper, we develop a method to identify poetic features that relate to literary prosodic classes. This method is compared with an approach based on neural networks for the classification of ellipsis and contrast class as well as for the identification of each line in the poem whether it is an elliptical line or not. The paper is organized as follows: Section 2 provides an overview about database. Section 3 reviews the feature engineering (rule-based) as well as neural networks (NNs) based approach. The experimental results are described in Section 4. Finally, conclusions and future works are presented in Section 5.

2 Database

The data used in the project *Rhythmicalizer* (`www.rhythmicalizer.net`) is from our partner *lyrikline* (`www.lyrikline.org`). Lyrikline was initiated by the Literaturwerkstatt Berlin and 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 *lyrikline* reading a total of $2,543$ poems. In this work, we have selected only a small amount of data. In total, there were 121 poems by 42 different poets. We used two different classes: poems that are highly elliptical as well as poems based on syntactically regular sentences. The first group is the ellipsis and the second group was our contrast class. The contrast class contains poems consisting of complete and correct sentences. In the "elliptical" group, there were 56 poems, written by 26 different authors. In the contrast corpus, there were 65 poems by 21 different authors: 5 authors were thus represented in both classes. The number of samples or poetic lines in the ellipsis and the contrast class is $1,845$ and $2,090$, respectively.

2.1 Manual Annotation

In the manual annotation of ellipsis, only the poetic lines where a missing phrase would have to be added were identified as being elliptical. We did not mark those lines as elliptical lines where "only" the punctuation was missing. We marked the lines by adding a (yes) or (no) as "elliptical" or "not elliptical" poetic lines, respectively. For example, the following poem "Junge Hunde" (english: Young pups) which was written by lyricist Marcel Beyer and translated by Hans-Christian Oeser and Gabriel Rosenstock.

"Ach, die Gutgebügelten, junge (no)
Luden am Nebentisch, trinken (no)
das Panzerpils und tauschen (no)
Herrenheftchen. Einer wirft beim (no)
Aufstehn die Flasche Bier um, (no)
schmiert dann, nach und nach, ein (no)
ganzes Paket Tempos über den (no)
Plastiksitz. Kurzschnitte, und (no)
Pomade. Totes Büffet. Im Jungsklo, (yes)
schöne Teile. Ein alter Glatzkopf (yes)
zupft sich etwas von den Lippen. (no)
Humer-Bursche. Geschlipst. Gewienert. (yes)
Junge Hunde. Fickriges Blau. Im (yes)
Klappergang, Wien West, verschwitzte (yes)
Gürteltiere. Rauch schneller, Lude. (no)
Ohneservice. Vielleicht Ein- bis (yes)
Zweihundert, in Randbezirken, (yes)
Arbeiterbeisln, Siebzehnter. Trotte (yes)
im Regen, aus einem offenen Fenster, (no)
obere Etage, Dampf. (yes)"

"Ah, the well-creased, young (no)
pimps at the next table, slurping (no)
Panzer pils and swapping (no)
porno mags. One of them gets up, (no)
knocks over his bottle of beer, (no)
then, bit by bit, mops up with a whole (no)
packet of kleenex across the (no)
plastic seat. Crew cuts and (no)
brilliantine. Dead buffet. In the jacks (yes)
nice bits. An old slap-head (yes)
plucks something from his lips. (no)
HUMER chap. Betied. Polished. (yes)
Young pups. Randy blue. Clip- (yes)
clopping along. Vienna West, sweaty (yes)
armadillos. Smoke faster, pimp. (no)
Topless service. Maybe one to (yes)
two hundred, on the outskirts, (yes)
workers' dives, seventeenth district. (yes)
Trotting in the rain, from an open window, (no)
upper floor, steam. (yes)"

For the 56 poems in our corpus that are dominated by ellipsis, the students of literary studies during the "Plotting Poetry" symposium [6] annotated each line whether it was an ellipsis or not (or whether they faced a severe difficulty in taking this decision). The annotations of ellipsis are corrected by the philological scholar of our project (second author). From a total of $1,845$

annotated lines, 89 lines labeled as unclear and are excluded from the analysis below. The number of remaining lines is $1,054$ and 703 for the annotations (yes) and (no), respectively.

3 Classification Approaches

For the classification task of poems, we extracted different features in order to measure the influence of various modeling parameters on classification performance. Two approaches are developed. The first one based on traditional feature extraction and classification with machine learning algorithms. The second approach uses a neural networks. The following tools are used for the analysis and feature extraction:

1. **Text-Speech Aligner**: The first step required is to create a text-speech alignment for the written poems and spoken recordings. We perform forced-alignment of text and speech for the poems using the text-speech aligner published by [7] which uses a variation of the SailAlign algorithm [8] implemented via Sphinx-4 [9]. We extract the line-by-line timing (start of first word and end of last word in the line) for each line. The forced alignment of text and audio in spoken poetry is non-trivial and often individual words or lines cannot be aligned. Therefore, the automatic extracted alignment information are manually corrected by the first author.

2. **Parser**: Poems are processed using a statistical parser in order to add syntactic features. The Stanford parser [10] is used to parse the written text of poems. The parser used the Stuttgart-Tübingen-TagSet (STTS) table [11]. The main problems in poem parsing are the absence of punctuation in some cases (11 poems in the ellipsis and 2 poems in the contrast class are without punctuation), writing with special characters, or the writing of the whole text in lowercase or some words in uppercase. Therefore, the recognition of sentence boundaries by the parser is difficult. In addition, elliptical sentences within a poem often run on to the next line and go beyond the line boundary. Such unconnected syntactic elements result from the dissolution of poetic lines, caused by the so-called enjambments. Each sentence of a poem is structured as a parse tree, which is an ordered, rooted tree that represents the syntactic structure according to some context-free grammar. Within each line, there is one root node, containing of two (or more) branch nodes, the nominal phrase and the verbal phrase.

3.1 Feature Engineering-Based Approach

In order to recognize the ellipses, we had to proceed line by line, even if there were run-onlines (enjambments) within a poem. The most important indicator for an ellipsis was the absence of a verb within a complete sentence or half-sentence. We used parser informations based on a number of abbreviations of words' Part-of-Speech (PoS). We focused on the following verbs: finite verbs (VVFIN), imperative verbs (VVIMP), auxiliary verbs (VAFIN), auxiliary imperative verbs (VAIMP), and finite modal verbs (VMFIN). We also had to identify the punctuation marks for the detection of ellipses, cause the complete sentences in lines can be identified by the sentence ending punctuations (. ? ! ; :), and the clauses by the comma. Therefore, all punctuation marks are detected in every poetic line. Two types of conjunctions are identified: subordinate conjunction with sentence (KOUS) and coordinating conjunction (KON). We also identified the following types of nouns: normal noun (NN) and own name (NE). However, parsers cannot yet distinguish between nominative and accusative, so the most important indicator for a complete sentence was the verb. The features are detected as follows: If the poetic line contains one or more verbs, a value of one is added to the feature vector. The same process is implemented for noun, comma, sentence ending punctuation, and conjunction. Three feature sets are utilized: **A** (2 features): verb, sentence ending punctuation; **B** (3 features): verb, comma, sentence ending

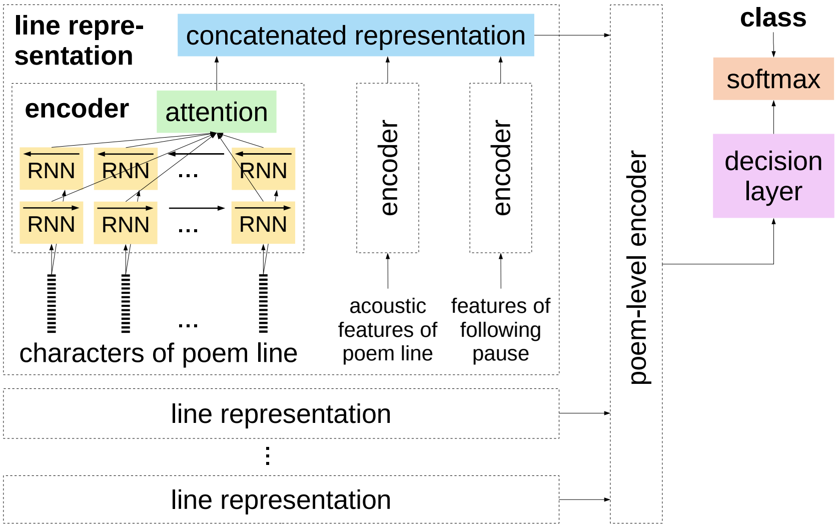

Figure 1 – Full model for poetry style detection using neural networks.

punctuation; **C** (5 features): verb, noun, comma, sentence ending punctuation, conjunction. Several machine learning algorithms in the WEKA data mining toolkit [12] are selected in the classification process: **IBk**: the Instance-Based (IB) classifier with a number of (k) neighbors is the K-nearest neighbours (KNN) classifier using the euclidean distance and 1-nearest neighbour [13]; **RandomForest**: The classifier of random forest consists of several uncorrelated decision trees [14]; **J48**: The J48 algorithm used to generate a pruned or unpruned decision tree [15].

3.2 Neural Networks-Based Approach

We describe in this section the approach based on neural networks for classification of prosodic styles [16, 17]. The model must deal well with *data sparsity*, since there are a broad variety and relatively a small number of poems. Therefore, we used a few free parameters as possible that need to be optimized during training. For this reason, we focused by textual processing on character-by-character encoding of poetic lines (and using character embeddings). The textual information, and the spoken recitation on line level as well as the pause information between lines are utilized. We use a bidirectional recurrent neural network (RNN, using gated recurrent unit (GRU) cells [18]) which encodes the sequence of characters into a multi-dimensional representation that is trained to be optimal towards differentiating the prosodic classes. Pre-training with additional data from German Text Archive [19] is implemented. The model is not trained using an explicit notion of words. Instead, it may implicitly encode word-level information (such as parts of speech) via the constituting sequences of characters. This is in line with recent work on end-to-end learning, for example, in speech recognition [20], which no more explicitly model of phonemes nor words, but directly transfers audio features to character streams. While processing on the word level might allow our model to build a better higher-level understanding of the poem's meaning, this semantic information would likely not help in style differentiation. In addition, word representations would not capture the usage of whitespace, example for indentation, to create justified paragraphs, or other uses, nor special characters. We combine the line-by-line representations using a poem-level encoder which is fed to a decision layer and a final softmax to determine the poem's class, yielding the hierarchical attention network as shown in Figure 1.

Table 1 – Classification results (weighted f-measure) of the ellipsis versus the contrast class using feature- and NNs-based approaches.

feature engineering and classifier			representation learning and NNs		
A	B	C	text-only	text+speech	text+speech+pause
0.57	0.62	0.62	0.81	0.94	0.93

Table 2 – Classification results (weighted f-measure) for poems that are dominated by ellipsis on line level using feature- and NNs-based approaches.

feature engineering and classifier			representation learning and NNs		
A	B	C	text-only	text+speech	text+speech+pause
0.57	0.62	0.60	0.53	0.55	0.52

4 Results

A statistic about the gender of authors shows that 65% of all elliptical writers were female (17 female authors from a total 26 in the "elliptical" group). In the contrast group, 29% of writers were female (6 female authors from a total 21). Of course, one can not explain this striking difference by a lack of language skills or a feminine tendency to silence. A better explanation for this phenomenon offers the aforementioned idea of "aural ellipsis" coined by Piombino. By this he means linguistic contractions or omissions that, when sounded, cause the listener to fill in the gaps with their 'inner experience.' This is caused by effect that we hear something parallel which is otherwise mysteriously inaccessible. The aural ellipses opens spaces for invention on the part of the listener: "This opening or freeing of forms of focusing in turn makes possible an intensified collaborative sharing (between a poet and listeners at a reading, for example) in the effort of organizing otherwise anomalous, disparate and incommunicable perceptions into patterns of meaning that can be further articulated, refined, and better understood, in an ongoing process" [5, pp. 57].

We use the approaches described in the previous section in order to differentiate the ellipsis from the contrast class. The results, calculated by the weighted F-measure, for the classification of the ellipsis versus the contrast class using feature engineered and NNs-based approach are presented in Table 1. The feature based approach yielded best results for both feature vectors (B and C) with a F-measure of 0.62. We get the best results with the NNs-based approach by using text and speech features of poetic lines (F-measure is 0.94). The difference between both approaches is very large, but we must to mention that the feature based approach based only on textual information from the parser.

Table 2 shows the classification results for the poems that are dominated by ellipsis and annotated manually whether each line is elliptical or not (see section 2.1). The best results by feature- and NNs-based approach are 0.62 and 0.55, respectively. This indicated that the identification of verbs and punctuation marks plays an important task by the recognition of elliptical poetic lines. The three machine learning classification algorithms (IBk, RandomForest, and J48) yielded the same results in feature based approach for results in Table 1 and 2.

5 Conclusion and Future Works

In our study, we examined the largest corpus of spoken poems currently available from *lyrikline* using a feature- and NNs-based approach. Both approaches utilized for the classification of ellipsis versus a contrast class of poems with complete and correct sentences as well as for the identification of elliptical lines in poems dominated by ellipsis. The features (verb, noun, comma, sentence ending, and conjunction) extracted from parser (based on text data only) utilized in the feature-based approach. The NNs-based approach used textual information, speech data of poetic lines, and pause information between lines. The NNs-based approach yielded best results for classification of ellipsis with the contrast class. The identification of elliptical lines in poems dominated by ellipsis was better using the feature based approach. With regard to German-language poems, we found that “aural ellipsis” (Piombino) were used in particular by female poets like Friederike Mayröcker, Marie-Luise-Kaschnitz, Ilma Rakusa, Doris Runge, Ulrike Draesner, Anja Utler, Elke Erb, Ginka Steinwachs or Isabeella Baeumer. We assume that these poets used the ellipsis to break with the lyrics of the 1970s, which often attempted to portray everyday phenomena in syntactically complete descriptions. Already the antigrammatic notation of Helmut Heißenbüttel served the attempt to replace the regular grammar with ellipses, missing verbs or a string of nouns without syntactical connection. The experiencing instance of the lyric ‘I’ was meant to disappear behind this new language. Many female poets followed this example, even radicalized it, as our study showed.

The feature based approach for the classification of poetic lines in the poems that are dominated by ellipsis using parser information yielded better results than NNs-based approach. Therefore, we want to add the parser features into the neural networks approach in order to improve the classification of poetic styles. A further step would now be to analyse translations of such elliptical poems: Where are these ellipses really kept within the target language, and where does the translation tend to correct elliptical sentences in order to clarify the poems meaning?

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’.

References

[1] POLIKARPOW, A.: *Zum Problem der asyndetischen Subordination in der Syntax der gesprochenen deutschen Sprache. Deutsche Sprache*, 2/96, pp. 154–168, 1996.

[2] LORENZ, O.: *Schweigen in der Dichtung: Hölderlin, Rilke, Celan. Studien zur Poetik deiktisch-elliptischer Schreibweise*. 1989.

[3] ALLEMANN, B.: *Experimentelle Dichtung in Österreich. Neue Rundschau*, (2), pp. 317–325, 1967.

[4] MAYRÖCKER, F.: *Notizen auf einem Kamel: Gedichte 1991-1996*. Suhrkamp, 1996.

[5] PIOMBINO, N.: *The Aural Ellipsis and the Nature of Listening in Contemporary Poetry. Close Listening: Poetry and the Performed Word, ed. C. Bernstein (Oxford University Press)*, pp. 53–72, 1998.

[6] MEYER-SICKENDIEK, B. and A.-S. BORIES: *Plotting Poetry II: Bringing Deep Learning to Computational Poetry Analysis*. Available on `https://www.geisteswissenschaften.fu-berlin.de/v/rhythmicalizer/termine/Unser-Symposium_-_Plotting-Poetry_II--Bringing-Deep-Learning-to-Computational-Poetry-Analysis_.html`, 2018. Last accessed at 25. January 2019.

[7] 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.

[8] 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.

[9] 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.

[10] RAFFERTY, A. N. and C. D. MANNING: *Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines*. In *Proceedings of the Workshop on Parsing German*, PaGe '08, pp. 40–46. Association for Computational Linguistics, Stroudsburg, PA, USA, 2008.

[11] SCHILLER, A., T. TEUFEL, C. STÖCKER, and C. THIELEN: *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und größes Tagset)*. Available on `http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-1999.pdf`, 1999. Last accessed at 25. January 2019.

[12] HALL, M., E. FRANK, G. HOLMES, B. PFAHRINGER, P. REUTEMANN, and I. H. WITTEN: *The WEKA Data Mining Software: An Update*. *SIGKDD Explorations*, 11(1), pp. 10–18, 2009. doi:10.1145/1656274.1656278.

[13] AHA, D., D. KIBLER, and M. ALBERT: *Instance-Based Learning Algorithms*. *Machine Learning*, 6, pp. 37–66, 1991.

[14] BREIMAN, L.: *Random Forests*. *Machine Learning*, 45(1), pp. 5–32, 2001.

[15] QUINLAN, R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[16] 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.

[17] 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 *Proc. Interspeech 2018*, pp. 2162–2166. 2018.

[18] 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.

[19] GEYKEN, A., S. HAAF, B. JURISH, M. SCHULZ, J. STEINMANN, C. THOMAS, and F. WIEGAND: *Das deutsche textarchiv: Vom historischen korpus zum aktiven archiv*. *Digitale Wissenschaft*, p. 157, 2011.

[20] GRAVES, A. and N. JAITLY: *Towards end-to-end speech recognition with recurrent neural networks*. In *International Conference on Machine Learning*, pp. 1764–1772. 2014.