Style Detection for Free Verse Poetry from Text and Speech

Timo Baumann
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, USA
tbaumann@cs.cmu.edu

Hussein Hussein and **Burkhard Meyer-Sickendiek**
Department of Literary Studies
Free University of Berlin
Berlin, Germany
{hussein,bumesi}@zedat.fu-berlin.de

Abstract

Modern and post-modern *free verse poems* feature a large and complex variety in their poetic prosodies that falls along a continuum from a more fluent to a more disfluent and choppy style. As the poets of modernism overcame rhyme and meter, they oriented themselves in these two opposing directions, creating a *free verse spectrum* that calls for new analyses of prosodic forms. We present a method, grounded in philological analysis and current research on cognitive (dis)fluency, for automatically analyzing this spectrum. We define and relate six classes of poetic styles (ranging from *parlando* to *lettristic decomposition*) by their gradual differentiation. Based on this discussion, we present a model for automatic prosodic classification of spoken free verse poetry that uses deep hierarchical attention networks to integrate the source text and audio and predict the assigned class. We evaluate our model on a large corpus of German author-read post-modern poetry and find that classes can reliably be differentiated, reaching a weighted f-measure of 0.73, when combining textual and phonetic evidence. In our further analyses, we validate the model's decision-making process, the philologically hypothesized continuum of fluency and investigate the relative importance of various features.

1 Introduction

One of the most important explanations for modern art is the theory of aesthetic pleasure, which claims that the fluency of cognitive processing is the cause for the positive effect of aesthetic experience (Topolinski and Strack, 2009). Similarly, cognitive research on fluency showed that people rate stimuli that are processed more easily higher (Belke et al., 2010). On the other hand, many modern artists like Picasso or Schönberg complicated the processability of their works using processes of abstraction in order to prevent such automated or fluid forms of art comprehensibility. Regarding this development, Bullot and Reber introduced the term *disfluency* as an artistic strategy to bring more analytical forms of art experience to the fore (Bullot and Reber, 2013). Other researchers suggested that disfluency prompts people to process information more carefully, deeply, and on a higher level of abstraction (Smith and Smith, 2006).

We present a computational analysis that tests these two trends in the current discussion on art experience by focusing on the prosodic feature of (dis)fluency in modern and post-modern poetry. We assume that the rhythmical quality of modern poetry is "not properly measurable by the rules of traditional verse" (Wesling, 1996), since modern poetry often rejects the traditional metrical verse. Most modern and post-modern poetry uses rhythmical features beyond the metrical forms, preferring the imitation of naturalness of everyday language and its very fluent speech prosody. At the same time, modern poetry uses disfluent processes to introduce the kinds of obstructions to ease consumption as outlined above. Since Dadaism, many avantgarde poets developed certain kinds of disfluencies like line-breaks as well as segments and "micro-particles" known from sound poetry. In this paper, we offer a classification of modern and post-modern poetry along the continuum of fluency, going back to a similar idea developed in Eleanor Berrys theory of a 'Free Verse Spectrum' (Berry, 1997).

Proceedings of the 27th International Conference on Computational Linguistics, pages 1929–1940
Santa Fe, New Mexico, USA, August 20-26, 2018.

We develop a method to identify poetic features that relate to literary prosodic classes and with a special regard to the modelling of prosodic (dis)fluency, hence a form of style detection. Style detection has been a long-running topic in literary study: Tools for style analysis like *Metricalizer* (Bobenhausen, 2011) analyze metrical patterns given in a poem's text (see also (Agirrezabal et al., 2016) for more recent work on English metrical poems) and *Sparsar* (Delmonte and Prati, 2014) have used similar analyses to aid speech synthesis for metrical poems. Style modeling has also been used to automatically generate poems such as limericks (Manurung et al., 2000) or, more recently, traditional Chinese poetry (Zhang and Lapata, 2014; Wang et al., 2016). The above approaches feature poetic styles with very restrictive patterns far from the breadth of free verse. Focusing on contemporary American poetry, Kaplan and Blei (2007) analyze and visualize (textual) features of poems with respect to their poetic properties and Kao and Jurafsky (2015) estimate whether poems were written by professionals or amateurs. All the above works have used only the textual form of the poem. In contrast, music genre classification frequently relies on both audio and lyrics. Particularly close to our work is Tsaptsinos (2017) who uses a hierarchical attention network (Yang et al., 2016), which, however, is using just the lyrics. In our work, we extend the hierarchical network to comprise both speech and text in order to differentiate poetic styles that are primarily differentiated by their recitation.

In our own prior work, we have tried to differentiate classes of post-modern poetry using conventional feature extraction-based classification approaches. In (Hussein et al., 2018a), we havedifferentiated two enjambment-dominated styles using simple features from pausing and POS tagging and achieved an f-measure of 0.69; in (Hussein et al., 2018b), we differentiate *parlando* and *variable foot* styles with similar features, again reaching an f-measure of 0.69.

2 Classifying the 'Free Verse Spectrum'

At least 80 per cent of modern and post-modern poems have neither rhyme nor metrical schemes such as iambic or trochaic meter. Does this, however, mean that they lack any rhythmical features? According to the theory of free verse prosody, the opposite is true. Modern poets like the Imagists (Silkin, 1997; Cooper, 1998; Beyers, 2001), the Black Mountain poets (Golding, 1981; Steele, 1990; Silkin, 1997; Berry, 1997; Finch, 2000), as well as European poets before and after the second world war (Meyer-Sickendiek, 2012; Lüdtke et al., 2014) developed a post-metrical idea of prosody that employs rhythmical features of everyday language, prose, and musical styles including jazz and hip hop. In parts, they intended to create a fluent, in parts a disfluent prosody, for example in Dadaistic poetry. Donald Wesling (1971) offered a number of examples to illustrate the stylistic range of free verse poetry, focusing on the typical line arrangements in modern poems: (1) "Whitmanic", referring to Walt Whitman's adaptation of "the biblical verset and syntax" in "end-stopped lines . . . with boundaries so often equivalent to those of larger units of grammar," which Wesling sees as "constitut[ing] the precomposition or matrix of free verse in English"; (2) "line-sentences," as developed by Ezra Pound in *Cathay* on the basis of Ernest Fenollosa's theories of the sentence, in turn derived from the study of Chinese; (3) dismemberment of the line, whereby the line becomes "ground to the figures of its smaller units," and, as a sub-category, spatial dismemberment of the line by indentation, as William Carlos Williams does in his triadic line verse; (4) systematic enjambment (breaking a sentence or phrase into two lines), whereby the lines are "figures on the ground of the larger unit, the stanza"; (5) dismemberment with enjambment of the line, such that "the middle units on the rank scale engage in a protean series of identity shifts as between figure and ground" (Wesling, 1971). As can be seen, these five classes imply a continuous development from more to less fluent styles using an increase in dismemberments and enjambments.

Based on these examples, Eleanor Berry (1997) called for the investigation of poetry with regards to their features in 'the multidimensional space of free verse' on the basis of five 'axes' of form: (1) line-length, including extent of variability in length; (2) line-integrity, as determined by intralinear features as well as line-divisions; (3) line-grouping, whether stichic (ungrouped), in verse paragraphs, in stanzas, or dispersed on the page; (4) sensory basis of the verse form, whether aural, visual, or both; and (5) semantic function, that is, the relation of the verse form to the semantic aspect of the text, characterized in such terms as organic, iconic, and abstract (Berry, 1997). Berry used these five axes in order to classify

the spectrum of free verse in modern and post-modern poetry.

In this paper, we add prosody to this aforementioned free verse spectrum and establish a gradual one-dimensional continuum, whose two poles are denoted by the terms fluent and disfluent. We illustrate this prosodic spectrum by ranking six different poetic styles respectively prosodic patterns within the free verse spectrum, starting with the most fluent one: (a) The **parlando** pattern was coined by the German poet Gottfried Benn, who created a colon (word group)-based line grouping as in (3) above, but ignored the gap to the run-on-line – i.e. the part after the enjambment – by a fluent reading not emphasizing the enjambment. This fluency was typical for his conversational idiom. The second poetic pattern, (b) the **variable foot** is is identical to the "triadic line verse" (3 above) invented by W. C. Williams (Cushman, 1985). Like the parlando, the variable foot uses a "soft enjambment", but the poet now emphasizes the gap to the run-on-line. As long as this run-on-line also occurs between each singular colon of the poem, i.e. the noun and verbal phrases, this gap does not really affect the flow of the stanza and the poem still sounds quite natural.

In the third pattern, the (c) **unemphasized enjambment**, the poet now creates a more disfluent, choppy style by using the so-called "hard enjambments" that interrupts the reading flow of the poem. This occurs when the enjambment runs across stanzas; separates articles or adjectives from their nouns or splits a word across a line. Finally, the (d) **gestic rhythm** even emphasizes these hard enjambments, which makes the poem sound way more disfluent than in the two previous patterns. Bertolt Brecht coined this technique by calling it a "gestic rhythm", preventing the ear from gliding past the message. Gestic rhythm is any rhythm that causes some difficulty in listening to it.

Even more radical kinds of poetic disfluency have been developed in modern "sound poetry" by dadaistic poets like Hugo Ball and Kurt Schwitters or concrete poets like Ernst Jandl and Oskar Pastior. Within the genre of sound poetry, there are two main patterns: the (e) **syllabic decomposition** and the (f) **lettristic decomposition**, the last and most disfluent pattern. A typical example for syllabic decomposition in sound poetry is the *Ursonate* [The Sonata in Primal Speech] by Dadaist Kurt Schwitters, which begins with "Fümms bö wö tää zää Uu." A typical example for the most disfluent lettristic decomposition can be found in Ernst Jandl's *schtzngrmm*, which is presented in Figure 2 (b) below.

Given this spectrum of free verse poetry, we can add a simple hierarchy of linguistic units to clarify the range from fluency to disfluency. In a grammatical hierarchy, letters are the smallest units and they combine to form morphemes, which combine to form words, to groups, to clauses and finally sentences. Dominating units differ between the different styles along the fluency continuum, from lettristic decompositions to parlandos which are dominated by the largest units. In addition, the recitation style itself can be less or more fluent, and the formation and emphasis of the enjambments at the end of each line is indicative of the fluency of the style.

3 Modeling the Prosody of Spoken Free Verse Poetry

In this section, we describe our model, which is inspired by Yang et al. (2016), as well as our high-level decisions for modeling. Poetry, in particular post-modern poetry, is challenging material for computational modeling and statistical natural language processing. The very purpose of art (and post-modern poetry in particular) is to stand out and to defy or re-define rules, making generalization difficult. Poems, as compared to normal language use, contain unusual words (or no words at all in the case of decompositions, see above), and there is generally only very little data available as compared to most other domains. The automatic alignment of text and audio in spoken poetry is non-trivial (in particular for decompositions in more abstract poetry) and important clues may be contained not only in how the textual material is spoken, but also in the gaps between textual material, such as extra white-space or the pausing between the lines of a poem.

Given the broad variety of the poems in combination with their relatively small number (see below), our model must deal well with *data sparsity*, i.e. use as few free parameters as possible that need to be optimized during training. For this reason, we decide to focus our textual processing on character-by-character encoding of the lines in the poem (and using character embeddings). We use a bidirectional recurrent neural network (RNN, using gated recurrent unit (GRU) cells (Cho et al., 2014)) which encodes

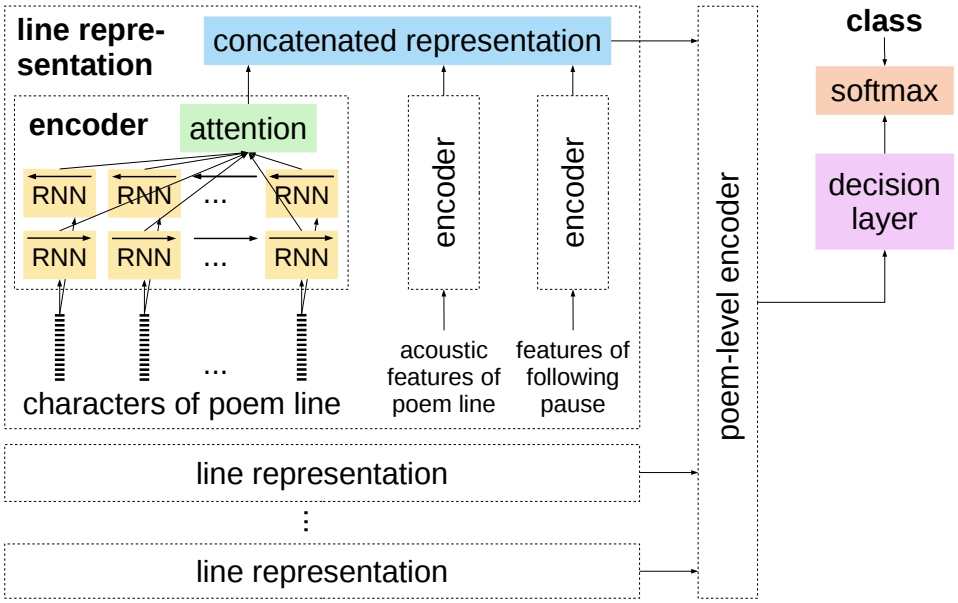


Figure 1: Full model for poetry style detection: each line is encoded character-by-character by a recurrent neural network (using GRU cells) with attention. Acoustic features of each line, as well as of the pause following up to the next line, are encoded similarly. Per-line representations are concatenated and passed to a poem-level encoder. The final decision layer optimizes for the poem's class.

the sequence of characters into a multi-dimensional representation that, although it is not directly applicable to human interpretation, is trained to be optimal towards differentiating the prosodic classes.

Our 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, e.g. for speech recognition (Hannun et al., 2014; Graves and Jaitly, 2014), which no more explicitly models 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, e.g. for indentation, to create justified paragraphs, or other uses, nor special characters.

We encode the speech (after feature extraction) in a similar way in order to capture the notion of fluency of speech delivery in the author's recitation. As for the text, we use speech line-by-line so that the model may synchronize what it 'hears' and what it 'reads'. A more fine-grained alignment of text and speech would be much harder to produce and might plausibly fail for sound poetry (although it would certainly be desirable). Finally, in order to differentiate the reading of enjambments, we also encode the pauses in-between lines, which may also contain important information about breathing, pausing, in-breath, etc.

For a model to be a suitable and acceptable tool for (digital) humanistic research, it should provide insight into its decision making process, as our primary goal is not so much the automatic classification of poetry but to learn about and better understand poetic styles. To satisfy this requirement of inspectability of the decision making process (at least to some extent), we implement a notion of *inner attention* (Liu et al., 2016) that is to learn how to combine the sequential states of each line's encodings (text, audio, and final pause) to a representation that is best suited towards our training objective. 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 wrt. its philological plausibility.

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. While our network is similar to those of Yang et al. (2016) and Tsaptsinos (2017), we base ours on characters instead of words as textual input and include the audio stream into the analysis via

Table 1: Descriptive characteristics of the data used in the experiments.

		poems	lines	characters	audio
lyrikline: German subcorpus		2392	61849	2025484	52 h
	parlando	34	1435	44323	67 min
	variable foot	34	878	23684	39 min
	unemphasized enjambment	36	1090	33178	48 min
	gestic rhythm	33	897	27741	44 min
	syllabic decomposition	21	540	12390	26 min
	lettristic decomposition	17	684	10007	31 min
deutschestextarchiv.de		—	34291	996714	—

additional encoders.

Our model is implemented in *dyNet* (Neubig et al., 2017) and the software to replicate our experiments is available at `https://bitbucket.org/timobaumann/deeplyrik` in order to foster research on free verse poetry. Our implementation is flexible towards the number of classes to distinguish and can be configured to ignore some of the features (cmp. Table 3). We make use of this flexibility and perform a range of classification experiments in Section 5, once we have described our data in more detail.

4 Data and Setup

We collected German poems available on the webpage *lyrikline*[1] and classified 175 of a total of ~2400 German poems into the six prosodic classes defined above, with a special focus on poets known for the use of such prosodic patterns. Manual classification of the data was carried out by the third author, a philologist and literary scholar and following criteria from the literature as outlined above. We also collected the corresponding audio recording of each poem as spoken by the original author, yielding 52 h of audio. Checking manually, we found some poems tagged as German that actually were not (<1 %) and discarded these from further processing. For pre-training (see below), we also downloaded from the German Text Archive (Geyken et al., 2011)[2] additional text-only poetry published between 1800 and 1930 (and hence most likely not post-modern; more recent poems are hard to find as they are typically still copyrighted).

Some key descriptive statistics of the poems as assigned to their classes are reported in Table 1. In order to check whether poetic classes can already be singled out based on their length (in lines, characters, or audio duration) alone, we checked for significant deviations from the overall corpus. For none of the classes, the poems' durations, or number of lines significantly differ from the average poem in the corpus (two-tailed t-tests, $p > .05$ for all tests). However, variable foot poems, as well as syllabic and lettristic decompositions have significantly fewer characters than average poems.

4.1 Preprocessing

We perform forced-alignment of text and speech for the poems in our six classes using the text-speech aligner published by Baumann et al. (2018) which uses a variation of the SailAlign algorithm (Katsamanis et al., 2011) implemented via Sphinx-4 (Walker et al., 2004). 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 (e.g. parts of speech, syntax or semantics) to the poem in the future.

We extract the line-by-line timing (start of first and end of last word of the line) for each line. Forced alignment of poetry is far from trivial and often individual words cannot be aligned. Lettristic and syllabic decompositions, being a form of sound poetry, are notoriously hard to align automatically and we resorted to manual alignment of those lines that could not be aligned automatically.

[1] `http://www.lyrikline.org`

[2] `http://www.deutschestextarchiv.de`

We extract Mel-frequency cepstral coefficients (MFCC) for every 10 milliseconds of the audio signal as well as fundamental frequency variation (Laskowski et al., 2008, FFV) vectors, which are a continuous representation of the speaker's pitch. We z-normalize all feature dimensions. In order to not overwhelm the model with acoustic sequence information, and given that relevant speech phenomena are typically much longer than 10 milliseconds, we compute the mean and standard deviation of 10 consecutive frames for every feature.

4.2 Pre-training

The manually classified corpus is small and hence the quality of intermediate representations is limited by data sparsity. As an example: a strongly distinctive characteristic of syllabic and lettristic decomposition is the presence of repetitive consonant-vowel sequences in the poem (which occur frequently in syllabic decompositions). Yet, in the two-class problem of distinguishing the sub-types of decompositions, it is hard to infer the differentiation of characters into consonants and vowels from only 38 example poems (which feature a wealth of other characteristics). It would be similarly unreasonable to ask a student of literary studies to learn to differentiate poetic styles based on just a few examples in a language and writing script unknown to them. We mitigate this problem by using pre-training methods (Erhan et al., 2010) to help bootstrap the generation of reasonable intermediate representations (i.e., we teach the model some notion of poetic language as a foundation before teaching it to differentiate styles).

We pre-train the character embeddings and the line encoder using a recurrent autoencoder that aims to build a representation of the line that best allows it to re-create the original line (using combined costs of both forward and backward decoding as training objective); in other words: we ask our model to memorize poetic lines but given its limited memory it has to learn an abstraction of each line that helps it to remember the line. We train this on the whole poetry corpus including poems collected from deutschestextarchiv.de.

We pre-train our representations for the acoustic features of each line similarly to the textual pre-training in that we train a recurrent autoencoder that aims to re-create the original line-by-line features, as well as the length of the acoustic stimulus. Re-creating the length of the original stimulus is particularly important as this feature is directly relevant for measuring the pause between two lines and is otherwise only a very indirect objective in pre-training. Given that line-by-line alignments are only available for the 175 poems that were manually classified, we pre-train the acoustic representations on inter-pausal units for the line representations (pausal units for between-line representations) detected using voice activity detection.

It would also be desirable to pre-train the model's poem-level encoding (i.e. not just teach it about lines in a poem by having it memorize lines, but also teach how lines are combined into a poem). Unfortunately, line-by-line text-audio alignments are not available for the full corpus and hence we are limited to either pre-train based on textual information only (reported as 'text-only' in Section 5) or to use audio information but not use pre-trained poem-level information.

4.2.1 Training Procedure

Even when using pre-trained internal representations, only 175 training instances are too few for training the deep model towards the classification objective. However, poems typically display their structural properties on the vast majority of the lines they are composed of. We hence split training into two steps by first training a decision network that learns to classify individual lines of the poem in order to adapt the pre-trained network. While we here ignore the run-on-line in the case of enjambements, we do include the pausing information to model enjambments at least partially. Coming back to Figure 1, we first leave out the poem-level encoding and directly pass each line representation to a line-by-line decision layer.

Afterwards, we replace the line-by-line decision layer with the poem-level encoder and final decision layer and train towards the per-poem decisions based on the parameters estimated before. Thus, the final model is able to steer its attention mechanism towards the important lines and can learn to sacrifice the initially trained per-line optimization for the overall per-poem optimization.

For all classification experiments reported below, we perform 15 training epochs and use a dropout probability of 0.3 (Srivastava et al., 2014) to reduce overfitting. Each encoder is two layers deep and

Table 2: Per-class f-measures as well as the confusion matrix for the six-class classifier.

	f-measure	parlando	var. foot	unemph. enj.	gestic	syll. dec.	lettr. dec.
parlando	0.83	30	2	2			
variable foot	0.60	3	20	6	5		
unemph. enj.	0.71	2	4	27	3		
gestic rhythm	0.68		6	5	21		1
syllabic dec.	0.81	2	1			17	1
lettristic dec.	0.77	1				4	12

Table 3: Results (weighted f-measure) for reduced feature sets; $^{\text{NS}}$ indicates that the classifier does not perform significantly above chance level.

	all features	no pause	text-only
all six classes	0.73	0.66	0.47
parlando vs. variable foot	0.85	0.85	0.65
unemphasized enjambment vs. gestic rhythm	0.78	0.66	0.57^{NS}
syllabic vs. lettristic dec.	0.82	0.92	0.82

has a 20-dimensional state. Our character embeddings are 20-dimensional as well as are the attention representations.

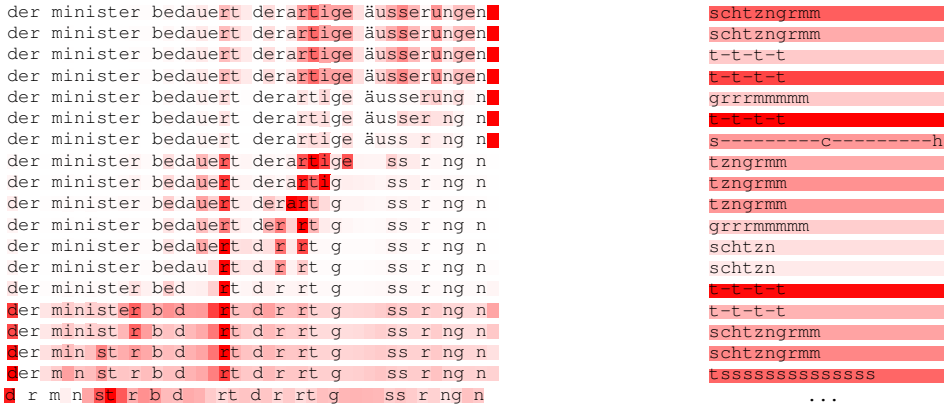
5 Classification Experiments

We train a classifier to distinguish the six classes of poetic style with all features (text, speech, and pause) using pre-processing and pre-training as described in the previous section; given the little available data, we use 25-fold cross-validation (5 poems per test fold). We report the per-class f-measures and confusion matrix in Table 2. The average f-measure (weighted by class size) is 0.73, indicating that it is indeed possible to distinguish the postulated prosodic classes based on text, speech, and pauses and using a deep neural model.

Analyzing the confusion matrix (in which we ordered the classes by their postulated fluency), we furthermore find that most misclassifications are clustered near the central diagonal. We take this as an indicator that classes close to each other on the fluency continuum are more easily misclassified with each other. In contrast, none of the very fluent classes is taken for a decomposition and only rarely are decompositions misclassified as member of the fluent classes. The picture is less clear for the three classes of variable foot, unemphasized enjambment and gestic rhythm. Given that all of these styles feature enjambments of some sort, they might be somewhat harder to differentiate, or the limited performance might point to the importance of higher-level information in the manual classification which our model is unable to pick up.

In Table 3 we show the weighted f-measure (as the most descriptive single performance metric) for the full six-class classifiers, as well as for binary classifiers that we train to differentiate (a) the two most fluent styles which are still regarded as relatively fluent to read and listen to; (b) the two styles that build on enjambments and mostly differ by whether these enjambments are unemphasized or emphasized as in gestic rhythm, and (c) the two types of decomposition which differ quite strongly in their textual appearance. We train each of these classifiers for the full feature set, we leave out pause information and only use audio during speech, and finally use on the textual information.

The results in Table 3 indicate that indeed, (a) parlando and variable foot can hardly be differentiated based on textual features alone but only with access to the recitation style (yet do not require pausing information). (b) Pause information is crucial to differentiate the enjambment-dominated styles which differ mostly between the lines. In fact, these styles cannot be differentiated at all based on textual information alone, in which case the classifier does not significantly outperform the chance level. Finally,



(a) Jaap Blonk: Der Minister I

(b) Ernst Jandl: schtzngrmm

Figure 2: Visualization of attention in two lettristic poems: (a) attention to characters within the line, (b) attention to lines (including the audio) in the poem.

the (c) decomposition styles can already be differentiated based on textual information. Access to the recitation style improves classification performance but adding the pausal features shrinks this advantage, presumably because the pausing information is irrelevant to distinguish the classes and confuses the classifier during training. We also find that the neural model substantially outperforms our previous feature-based approaches (Hussein et al., 2018a; Hussein et al., 2018b), at least when using the full feature sets.

6 Model Analysis

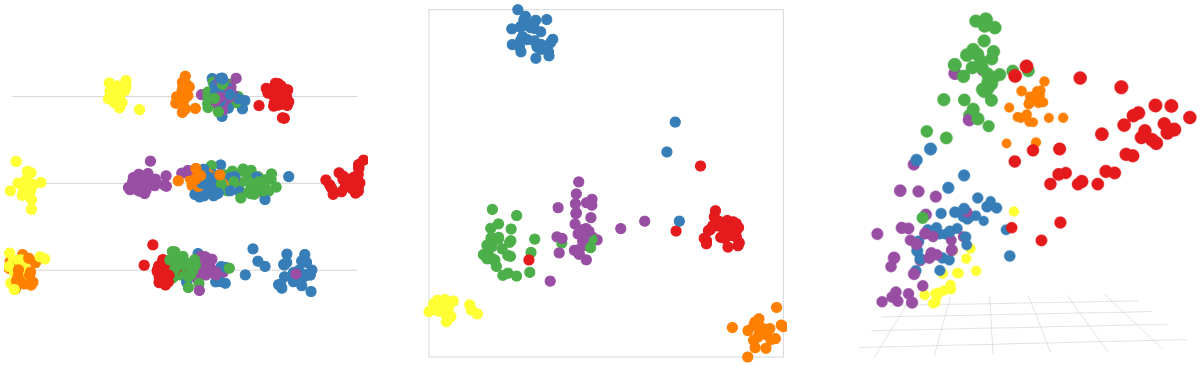
In order to gain additional insight into the classification performance and to investigate whether the model is observing the philologically 'correct' features of poems, we perform further analyses on the attention sub-model, as well as by analyzing the poem-level representations in a lower dimensional space.

Attention models can provide insight into the decision making of the model. In particular, they allow to analyze the weighing of parts of the input in the decision making process. This enables us to test whether the model has actually learned to classify similarly to how a human might.

We plot attention results for two exemplary poems in Figure 2: As can be seen, the poem in Figure 2 (a) repeats the same line over and over which gradually decomposes from the right. This leads to a diagonal frontier in the poem and the model attends to this frontier, at least to some extent. Another poem by the same author (which decomposes from the left) shows a similar pattern. The poem in Figure 2 (b) is about trench[3] warfare during World War II and onomatopoeically mimics the noises of war (leaving out all vowels in doing so). The model singles out those lines in the poem that deviate most from fluent language – both textually and acoustically. The decision layer trained in line-by-line training can be used for further analysis. For example, in the poem in Figure 2 (a), the line-by-line classification is wrong for some of the first few lines, mirroring the fact that decomposition only gradually sets in.

We have postulated in Section 2 that the six classes of poetic style can be ordered along a (dis)fluency continuum, with *parlando* being the most fluent (and similar to normal spoken language) and *lettristic decomposition* the least. While we have provided external evidence that the classifier does differentiate the fluency classes with the confusion matrix presented in Section 5, we wonder if the internal poem representation also reflects the ordering in this continuum. In order to visualize the internal representations of the poetic model, we train another classifier which injects a low-dimensional *bottleneck layer* immediately before the final decision layer, thus forcing the model to represent every poem as a point in low-dimensional space (we use 1-3 dimensions). We then plot each poem's representation as a point and color-code classes by color. Resulting plots are shown in Figure 3: on the left, we show the results of three 1D mappings resulting from different random initializations of the model. It can be seen that the first model represents the classes in the order of fluency but fails to differentiate the enjambment-based

[3]German: 'Schützengraben' which is reduced to 'schtzngrmm' by Jandl.



(a) three trials of linear mapping (b) 2D-mapping (c) Projection of 3D-mapping[4]

Figure 3: Visualization of low-dimensional mapping of the poems contained in the six classes (parlando: red, variable foot: blue, unemphasized enjambment: green, gestic rhythm: purple, syllabic decomposition: orange, lettristic decomposition: yellow).

classes. The second mapping better differentiates the classes but more strongly deviates from the fluency continuum, which the third mapping ignores alltogether.

The center and right of the figure show 2D and 3D mappings[4]. As can be seen, the classes are mapped into separate areas of the representation space and cluster nicely (except for a few outliers). However, their ordering only partially corresponds to the fluency continuum. For the 3D mapping[4] we find differentiations in the representations that could be interpreted as aspects of fluency modeling: (a) the naturalness of the text (with both types of decomposition and their generally non-word content set apart from the other poems in one dimension) and (b) auditory vs. textual fluency, with enjambment and syllabic decomposition being auditorily similarly fluent to *parlando* but textually not.

7 Conclusion

With our analysis, we have captured the spectrum of free verse prosody in modern poetry, using computational techniques, along the fluency/disfluency continuum which also plays a central role in the discussion on the cognitive processing of aesthetic artifacts; to the best of our knowledge, we are the first to do so. We have trained classifiers that integrate for each line the textual information, the spoken recitation, as well as the pausing information, and integrate information across the lines within the poem. We deem the overall classification performance as high (although classification is not our primary goal). In addition, we have trained classifiers for sub-problems and using reduced feature sets and the results obtained in these experiments support the expectations based on philological understanding.

Our classifiers provide some insight into the decision making process via the attention mechanism and the possibility to map the internal state into lower-dimensional spaces for visualization. We find that our model indeed seems to internally re-create some notion of fluency. However, we also find that the mappings to lower dimensions do not fully support the claim of one single dimension of fluency. In particular in 3D-mapping, a more complex picture evolves. This may be due to shortcomings in the model, the underlying data and annotated classes. It may also question our initial hypothesis calling for a further refinement of the fluency theory.

In our future work, we intend to analyze the whole *lyrikline* corpus (and beyond) in order to gain insights about this broad sample of post-modern poetry. We hope to semi-automatically find additional poems that belong into one of our classes (and could be added as further training material after manual validation). We also intend to analyze the corpus for clusters of outliers from our current classification in order to determine further and refine the existing classes using an iterative "human in the loop" appoach (Baumann and Meyer-Sickendiek, 2016). Our goal is the mutual benefit of the model (which requires human input) and the philological expert who will be able to quickly scan, analyze and browse vastly larger collections of poetry than has been possible in the past.

[4]An interactive version of the 3-D plot is available at `https://timobaumann.bitbucket.io/colingfreeversepoetry/`.

Our model so far is still far from optimal (beyond the fact that we have not searched meta-parameters that yield best classification results) and we want to point out some aspects of future work: enjambments are linked to syntactic characteristics and we could integrate syntactic information (such as part-of-speech (POS) tags) or try to pre-train our character encoding to decode the POS tag sequence. It would be helpful if we could inject philological insight about what to pay attention to into the training process (and extract it out of the application process) in order to increase the philological validity of our model. With regards to the neural network, we could link the text and speech streams using connectionist temporal classification (Graves et al., 2006) for it to relate auditory to textual information in more detail, and we could connect the (sequential) line and pause encodings of audio in order for the model to better normalize out speaker specific (but not style-specific) characteristics.

Acknowledgements

This work is funded by Volkswagen Foundation in the programme 'Mixed Methods in the Humanities.' We thank Vasu Sharma for valuable suggestions on neural network processing, and Arne Köhn for valuable comments on a draft of the paper, as well as the anonymous reviewers for their insightful comments.

References

Manex Agirrezabal, Iñaki Alegria, and Mans Hulden. 2016. Machine learning for metrical analysis of english poetry. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 772–781. The COLING 2016 Organizing Committee.

Timo Baumann and Burkhard Meyer-Sickendiek. 2016. Large-scale analysis of spoken free-verse poetry. In *Proceedings of Language Technology Resources and Tools for Digital Humanities (LT4DH)*, Osaka, Japan, December.

Timo Baumann, Arne Köhn, and Felix Hennig. 2018. The Spoken Wikipedia Corpus Collection: Harvesting, Alignment and an Application to Hyperlistening. *Language Resources and Evaluation*.

Benno Belke, Helmut Leder, Tilo Strohbach, and Claus Christian Carbon. 2010. Cognitive fluency: High-level processing dynamics in art appreciation. *Psychology of Aesthetics, Creativity, and the Arts*, 4(4):214–222.

E. Berry. 1997. The Free Verse Spectrum. *College English*, 59(8):873–897.

C. Beyers. 2001. *A History of Free Verse*. University of Arkansas Press.

Klemens Bobenhausen. 2011. The Metricalizer – automated metrical markup of German poetry. *Current Trends in Metrical Analysis, Bern: Peter Lang*, pages 119–131.

NJ Bullot and R. Reber. 2013. The Artful Mind Meets Art History: Toward a Psycho-Historical Framework for the Science of Art Appreciation. *Behavioral and Brain Sciences*, 36(2):123–137.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. 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)*, pages 1724–1734, Doha, Qatar, October. Association for Computational Linguistics.

G. B. Cooper. 1998. *Mysterious Music: Rhythm and Free Verse*. Stanford University Press.

S. Cushman. 1985. *William Carlos Williams and the Meaning of Measure*. Yale Studies in English. New Haven and London.

Rodolfo Delmonte and Anton Maria Prati. 2014. Sparsar: An expressive poetry reader. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 73–76, Gothenburg, Sweden, April. ACL.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. 2010. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(Feb):625–660.

A. Finch. 2000. *The Ghost of Meter: Culture and Prosody in American Free Verse*. University of Michigan Press.

Alexander Geyken, Susanne Haaf, Bryan Jurish, Matthias Schulz, Jakob Steinmann, Christian Thomas, and Frank Wiegand. 2011. Das deutsche textarchiv: Vom historischen korpus zum aktiven archiv. *Digitale Wissenschaft*, page 157.

A. Golding. 1981. Charles Olson's Metrical Thicket: Toward a Theory of Free-Verse Prosody. *Language and Style*, 14:64–78.

Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *International Conference on Machine Learning*, pages 1764–1772.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376. ACM.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Hussein Hussein, Burkhard Meyer-Sickendiek, and Timo Baumann. 2018a. Automatic detection of enjambment in german readout poetry. In *Proceedings of Speech Prosody*, Poznán, Poland, June.

Hussein Hussein, Burkhard Meyer-Sickendiek, and Timo Baumann. 2018b. Tonality in language: The "generative theory of tonal music" as a framework for prosodic analysis of poetry. In *6th International Symposium on Tonal Aspects of Language (TAL)*, Berlin, Germany, June.

Justine T Kao and Dan Jurafsky. 2015. A computational analysis of poetic style. *LiLT (Linguistic Issues in Language Technology)*, 12.

David M Kaplan and David M Blei. 2007. A computational approach to style in american poetry. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 553–558. IEEE.

Athanasios Katsamanis, Matthew Black, Panayiotis G Georgiou, Louis Goldstein, and S Narayanan. 2011. SailAlign: Robust long speech-text alignment. In *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*.

Kornel Laskowski, Mattias Heldner, and Jens Edlund. 2008. The fundamental frequency variation spectrum. In *Proceedings of FONETIK 2008*.

Yang Liu, Chengjie Sun, Lei Lin, and Xiaolong Wang. 2016. Learning natural language inference using bidirectional LSTM model and inner-attention. *CoRR*, abs/1605.09090.

J. Lüdtke, B. Meyer-Sickendieck, and A. M. Jacobs. 2014. Immersing in the stillness of an early morning: Test ing the mood empathy hypothesis of poetry reception. *Psychology of Aesthetics, Creativity, and the Arts*, 8(3):363–377.

Hisar Manurung, Graeme Ritchie, and Henry Thompson. 2000. Towards a computational model of poetry generation. In *Proceedings of the AISB'00 Symposium on Creative and Cultural Aspects and Applications of AI and Cognitive Science*.

B. Meyer-Sickendiek. 2012. *Lyrisches Gespür: Vom geheimen Sensorium moderner Poesie*. Fink, Wilhelm.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980*.

J. Silkin. 1997. *The Life of Metrical and Free Verse in Twentieth-Century Poetry*. Palgrave Macmillan UK.

L. F. Smith and J. K. Smith. 2006. The nature and growth of aesthetic fluency. In *New directions in aesthetics, creativity, and the psychology of art, ed. P. Locher, C. Martindale, L. Dorfman, V. Petrov and D. Leontiev*, page 47–58.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

T. Steele. 1990. *Missing Measures: Modern Poetry and the Revolt Against Meter*. University of Arkansas Press.

S. Topolinski and F Strack. 2009. The architecture of intuition: Fluency and affect determine intuitive judgments of semantic and visual coherence, and of grammaticality in artificial grammar learning. *Journal of Experimental Psychology*, 1(138):39–63.

Alexandros Tsaptsinos. 2017. Lyrics-based music genre classification using a hierarchical attention network. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)*, pages 694–701.

W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel. 2004. Sphinx-4: A Flexible Open Source Framework for Speech Recognition. Technical report, Mountain View, CA, USA, November.

Zhe Wang, Wei He, Hua Wu, Haiyang Wu, Wei Li, Haifeng Wang, and Enhong Chen. 2016. Chinese poetry generation with planning based neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1051–1060. The COLING 2016 Organizing Committee.

D. Wesling. 1971. The Prosodies of Free Verse. In Reuben A. Brower, editor, *Twentieth-Century Literature in Retrospect*, pages 155–187. Cambridge, MA: Harvard University Press.

D. Wesling. 1996. *The Scissors of Meter: Grammetrics and Reading*. University of Michigan Press.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

Xingxing Zhang and Mirella Lapata. 2014. Chinese poetry generation with recurrent neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680, Doha, Qatar, October. Association for Computational Linguistics.