Metrical Tagging in the Wild: Building and Annotating Poetry Corpora with Rhythmic Features

Anonymous EACL submission

Abstract

One of the main obstacles for many Digital Humanities projects is the low data availability. Especially resources for poetry are scattered across multiple datasets with different standards and variety. Additionally, the automatic annotation of rhythmic features in text, arguably an integral part of poetry research, largely relies on lexical resources that are scarce and typically neglect context. We provide large collections of English and German poetry, including an annotation with part-of-speech, syllable boundaries, and verse measures. To learn these measures, we annotate prosodic features in smaller corpora, aiming at a diverse sample of rhythm in verse, covering the last 400 years. We develop a typology of verse measures and train bilstm-crf models with pre-trained syllable embeddings to learn the stress sequence of these measures, outperforming BERT with word piece tokenizer. In a multi-task setup we find beneficial task relationships particularly in combination with syllable stress (meter and main accents) which in turn benefits from a global line (verse) label.

1 Introduction

Metrical verse, lyric as well as epic, is already common in preliterate cultures (Beissinger, 2012), and the majority of oral literatures to this day is drafted in verse. In order to reconstruct such oral traditions, literary scholars mainly study textual resources (rather than audio) and much rhythmical analysis in poetic verse is accomplished through example-driven textual annotation or rule-based tools (Carper and Attridge, 2020; Kiparsky, 2020; Plecháč, 2020). Fortunately, well-defined constraints and the regularity of metrically bound language aid the prosodic interpretation of a text, i.e., how the piece is supposed to be read and performed.[1] Figure 1 shows examples of the rhythmical variety of some fairly frequent verse measures, including an annotation of main accents and caesura, which for practicality, we call rhythm here.

met="-+|-+|-+|-+|-+|" rhythm="020:0010202:"
(1) Ein **Flüs**tern, **das** in **trü**bem **Schlaf** er**trinkt**.
A whisper that in drowsy sleep.DAT drowns
A whisper that drowns in a drowsy sleep.

met="-+|-+|-+|-+|-+|" rhythm="010002:0102:"
(2) My **love** is **like** to **ice**, and **I** to **fire**:

met="+--+|-+|-+|-+|" rhythm="1002:02010:1:"
(3) **Wish**es to **see**, and **what** he **wish**es, **Spies**:

met="-+|-+|-+|-+|" rhythm="01020001:"
(4) The **win**ter **eve**ning **set**tles **down**

met="+-|+-|+-|+" rhythm="0010102:"
(5) **Walk** the **deck** my **Cap**tain **lies**,

Figure 1: Examples of rhythmically annotated poetry, with meter, feet, main accents & caesura. (1) Georg Trakl, iambic.penta (2) Edmund Spenser, iambic.penta (3) Anne Killigrew, iambic.penta.invert (4) T.S. Eliot, iambic.tetra (5) Walt Whitman, troch.tetra.

While the speech processing community has moved on to end-to-end unsupervised methods to detect and control the overall personal and emotional aspects of speech, including fine-grained features like pitch, tone, speech rate, cadence, and accent (Valle et al., 2020), computational research on prosody in text, including poetry, still largely relies on lexical resources with stress annotation, such as the CMU dictionary (Hopkins and Kiela, 2017; Ghazvininejad et al., 2016), presumes aligned audio (Rosenberg, 2010), or is based on words rather than syllables (Talman et al., 2019; Nenkova et al., 2007). Besides the ill-suitedness of pronunciation dictionaries to model contextual effects of prosody, their creation is laborious, and thus resources only exist for a handful of languages.

Additionally, even though practically every cul-

[1]In contrast, rhythmically varied prose is considerably harder to annotate with prosodic features.

ture has a rich heritage of poetic writing, comprehensive collections of poetry are rare. Most poetry corpora today were collected with the intention of modeling certain poetic features (like rhyme) or consist only of a particular genre (like sonnets).

We present in this work new datasets of annotated verse for a varied sample of around 3500 lines each for German and English. In addition, we collect and automatically annotate large poetry corpora for both languages to advance computational work on literature and rhythm. This may include the analysis and generation of poetry, but also more general work on prosody, or even speech synthesis of creative language.

Our main contributions include:

1. The collection and standardization of heterogenous text sources that span writing of the last 400 years for both English and German, together comprising over 5 million lines of poetry.

2. The annotation of prosodic features in a diverse sample of smaller corpora, including metrical and rhythmical features and a comprehensive set of newly developed rules with a human-in-the-loop approach for the classification of verse measures.

3. The development of preprocessing tools and sequence tagging models to jointly learn the previously annotated features in a multi-task setup, resulting in substantial performance gains over previous work.

2 Related Work

2.1 Annotation of Prosodic Features

Earlier work (Nenkova et al., 2007) already found strong evidence that part-of-speech tags, accent-ratio[2] and local context provide good signals for the prediction of word stress. Subsequently, models like MLP (Agirrezabal et al., 2016) and discriminative sequence models such as conditional random fields (CRFs), LSTMs (Estes and Hench, 2016; Agirrezabal et al., 2019) and later transformer models (Talman et al., 2019) have notably improved the performance to predict the prosodic stress of words and syllables. Unfortunately, most of this work only evaluates model accuracy on syllable or word level, with the exception of Agirrezabal et al. (2019), who achieves 62% line accuracy on English poetry.

A resource with annotation of poetic meter has been missing for the New High German language. However, certain rhythmical patterns have been annotated on other genre (Anttila et al., 2018; Donat, 2010). For Middle High German, Estes and Hench (2016) have annotated a metrical scheme for hybrid meter, and Navarro et al. (2016) annotated hendecasyllabic verse (11 syllables) in Spanish Golden Age sonnets. Agirrezabal et al. (2016, 2019) have used the dataset of Navarro et al. (2016) and the *for-better-for-verse* dataset in modern English. Algee-Hewitt et al. (2014) have annotated 1700 lines of English poetry to evaluate their system. We incorporate the latter two datasets in our work.

2.2 Poetry Corpora

A number of poetry corpora have been used in the nlp community. Work on English has strongly focused on iambic pentameter, e.g., of Shakespeare (Greene et al., 2010) or with broader scope (Jhamtani et al., 2017; Lau et al., 2018; Hopkins and Kiela, 2017). Other work has created corpora of specific genres like sonnets (Ruiz Fabo et al., 2020), limericks (Jhamtani et al., 2019), or Chinese Tang poetry (Zhang and Lapata, 2014). There are further resources with rhyme patterns (Reddy and Knight, 2011; Haider and Kuhn, 2018) or emotion annotation (Haider et al., 2020). Truly large corpora are still hard to find. Parrish (2018) previously provided the poetry in the English Gutenberg collection by filtering single lines with a heuristic (anything that could look like a line), but disregarded the integrity of texts.

2.3 Rhythmic Poetry Generation

Generative models such as (weighted) finite state transducers (WFST) have seen widespread success to analyze and generate 'poetic' language according to rhythmic constraints (Greene et al., 2010; Hopkins and Kiela, 2017; Ghazvininejad et al., 2016). Newer approaches to poetry generation explore jointly learned language models with conditioning (Lau et al., 2018), but still operate on a rather heuristic notion of poetry. We aim to support such research by providing annotated data.

[2] A binomial distribution of how often a word form appears stressed vs. unstressed in a corpus

3 Corpora & Preprocessing

We collect and standardize large poetry corpora for English and German. The English corpus contains around 3 million lines, while the German corpus contains around 2 million lines. Furthermore, we manually annotate prosodic features in around 3.500 lines for each language. To preprocess and augment the data, we train and evaluate models for hyphenation and part-of-speech tagging. The finished corpora and the code for an XML API to process them can be found at https://github.com/anonymous-poetrybot-386

3.1 Large Corpora

In this paper, we present a standardized format to sustainably and interoperably archive poetry in both .json and TEI P5 XML. The .json format is intended for ease of use and speed of processing while retaining a considerable amount of expressiveness. Our XML format is built on top of a "Base Format", the so-called DTA-Basisformat[3] (Haaf et al., 2014) that not only constrains the data to TEI P5 guidelines,[4] but also regarding a stricter relaxNG schema.[5]

3.1.1 A Large English Poetry Corpus

The English poetry corpus contains the entirety of poetry that is available in the English Project Gutenberg (EPG) collection. We firstly collected all files with the metadatum 'poetry' in (temporal) batches with the GutenTag tool, to then parse the entire collection in order to standardize the inconsistent XML annotation of GutenTag and remove duplicates, since EPG contains numerous different editions and issues containing the same material. We also filter out any lines (or tokens) that indicate illustrations, stage directions and the like. We use langdetect[6] 1.0.8 to filter any non-English material.

Parrish (2018) previously provided the poetry in EPG by filtering single lines with a heuristic (anything that could look like a line), not only including prose with line breaks, but also without conserving the integrity of poems but providing a document identifier per line to find its origin. We offer the corpus with intact document segmentation and metadata, still containing over 2.8 million lines.

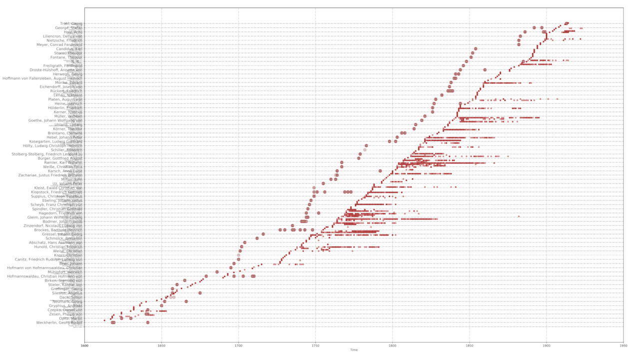

Figure 2: Poems of German Authors over Time. Textgrid (small dots bottom) vs. DTA (large dots top). 1600–1950. Authors not aligned.

3.1.2 A Large German Poetry Corpus

We build a large, comprehensive, and easily searchable resource of New High German poetry by collecting and parsing the bulk of digitized corpora that contain public domain German literature. This includes the German Text Archive (DTA),[7] the Digital Library of Textgrid,[8] and also the German version of Gutenberg.[9]

Each of these text collections is encoded with different conventions and varying degrees of consistency. We implement XML parsers in python[10] to extract each poem with its metadata and fix stanza and line boundaries. The metadata includes the author name, the title of the text, the year it was published, the title and genre of the volume it was published in, and finally, an identifier to retrieve the original source. We perform a cleaning procedure that removes extant XML information, obvious OCR mistakes, and normalize umlauts and special characters in various encodings,[11] particularly in DTA. Finally, we use langdetect[12] 1.0.8 to tag every poem with its language to filter out any poems that are not German (such as Latin or French). The corpus finally contains 2M lines in over 80k poems.

[3] http://www.deutschestextarchiv.de/doku/basisformat/

[4] https://tei-c.org/guidelines/p5/

[5] This schema promotes a somewhat strict layout of both `header` and `text`. Furthermore, it allows us to validate a XML file regarding its correctness. It is also useful for manual annotation with the OxygenXML editor, avoiding parsing errors downstream. Finally, this XML format is compatible with our API code to allow easy access. Examples of the XML and .json annotation schemes can be found in the appendix.

[6] https://pypi.org/project/langdetect/

[7] http://deutschestextarchiv.de

[8] http://textgrid.de

[9] https://www.projekt-gutenberg.org/

[10] with lxml and beautifulsoup

[11] We fix the orthography both on string and bytecode level. We replace the rotunda (U+A75B) and the long s (U+017F), the latter of which is pervasive in DTA.

[12] https://pypi.org/project/langdetect/

	German	EPG64	FORB	PROS
# correct lines	3431	1098	1084	1564
# faulty lines	58	114	49	173

Table 1: Size of manually annotated corpora. Faulty lines denotes the number of lines where the automatic syllabification failed. Correct lines are used for experiments, since only there the annotation aligns.

3.2 Manually Annotated Corpora

With the goal of learning particularly rhythmic features, we found two previously annotated datasets of Modern English poetry. We annotate additional poems for English, and build an annotated dataset for the New High German language from scratch.

3.2.1 English Rhythm Gold Corpus

The English corpus with manual annotation was collected from three sources: (1) The for-better-for-verse (FORB) collection[13] with around 1200 lines which was used by Agirrezabal et al. (2016, 2019). (2) The 1700 lines of poetry against which `prosodic`[14] (Algee-Hewitt et al., 2014) was evaluated (PROS), and finally (3) the 1200 lines in 64 English poems (EPG64) that were previously annotated for aesthetic emotions by Haider et al. (2020). The first two corpora were already annotated for syllable stress. EPG64 is annotated with the same guidelines as the following German corpus. FORB does not contain readily available foot boundaries, and in PROS foot boundaries are occasionally set after each syllable. Additionally, FORB makes use of a `<seg>` tag to indicate syllable boundaries, making it cumbersome to derive the position of a syllable in a word. It also contains two competing annotations, `<met>` and `<real>`.[15]

3.2.2 German Rhythm Gold Corpus

The small German corpus is fairly diverse, considering its size, and covers not only a wide range of different poem lengths and verse measures but also a number of influential German poets of both genders. Haider et al. (2020) previously annotated this corpus for aesthetic emotions. We annotate it for binary syllable prominence, foot boundaries, caesuras, and the main accents of a line. Besides the annotation of poetic features, every poem also has information on the author name, a title, the year of publication, and literature periods. We exclude two Middle High German poems by Walther von der Vogelweide and three poems in free rhythm (by Goethe) that do not allow for a metrical analysis, effectively amounting to 3.489 lines in 153 poems, spanning a time period from 1636 to 1936 CE. This yields a corpus that is somewhat representative for classical New High German poetry, while remaining manageable for manual annotation.

3.3 Preprocessing

Tokenization for both languages is performed with SoMaJo with a more conservative handling of apostrophes to leave words with elided vowels intact (Proisl and Uhrig, 2016).[16] We also train models for hyphenation (syllabification) and part-of-speech tagging.

3.3.1 POS tagging

Since we are dealing with historical data, POS tagger for modern languages are likely to degrade in quality. For English, POS tagging is carried out with the Stanford core-nlp tagger[17]. The tagset follows the convention in the Penn TreeBank. Unfortunately, this tagger is not geared towards historical poetry and consequently fails in a number of cases. We manually correct 50 random tagged lines and determine an F1-score of 92%, where particularly the 'NN' tag is overused as garbage class.

Test	Train				
	TIGER	DTA	DTA+TIG.	Belletr.	Poetry
Poetry	.795	.949	.948	.947	**.953**
Belletristik	.837	**.956**	.954	.955	.955
DTA Zeitung	.793	**.934**	.933	.911	.900
TIGER	**.971**	.928	.958	.929	.913

Table 2: Evaluation of POS taggers across genres. F1-scores.

For German, we use the gold annotation of the TIGER corpus, and pre-tagged sentences from DTA.[18] Both corpora are annotated according to the STTS tagset. We train and test Conditional Random Fields (CRF) with the sklearn crf-suite

[13] https://github.com/manexagirrezabal/for_better_for_verse/tree/master/poems

[14] https://github.com/quadrismegistus/prosodic

[15] `<met>` is the supposedly proper metrical annotation, while `<real>` is an annotation according to a more natural rhythm (with a tendency to accept inversions and stress clashes). We only choose `<real>` when `<met>` doesn't match the syllable count (ca. 200 cases).

[16] Which shows improved accuracy with special characters over NLTK.

[17] https://nlp.stanford.edu/software/tagger.shtml

[18] DTA was previously tagged with TreeTagger and manually corrected afterwards. http://www.deutschestextarchiv.de/doku/pos

across several genres to determine the most robust POS model.[19] See table 2 for an overview of the cross-genre evaluation. We find that training on TIGER is not robust to tag across domains, falling to around .8 F1-score when tested against different genres from DTA. The results suggest that this is mainly due to (historical) orthography, and to a lesser extent due to local syntactic inversions.

3.4 Hyphenation / Syllabification

For our purposes, annotating proper syllable boundaries is paramount. We compare several systems by their ability to draw accurate syllable boundaries. The used systems include *sonoripy*,[20], *Pyphen*[21], *hypheNN*,[22] and a biLSTM-CRF with pretrained character embeddings.[23] These embeddings are trained on the corpora in section 3.1. [24]

	German		English	
	w. acc.	sy. cnt	w. acc.	sy. cnt
SonoriPy	.476	.872	.270	.642
Pyphen	.839	.875	.475	.591
HypheNN	.909	.910	.822	.871
biLSTM-CRF	**.939**	**.978**	**.936**	**.984**

Table 3: Evaluation of Syllabification Systems on Wiktionary (German) and CELEX (English).

To train and test our models, we use CELEX2 for English and extract hyphenation annotation from wiktionary for German.[25] We evaluate our models on 20.000 random held-out words for each language on word accuracy and syllable count. While word accuracy rejects any word with imperfect character boundaries, syllable count is the more important figure to determine the proper length of a line. The biLSTM-CRF performs best for English and does not need any any postprocessing. For German, the model is less practical for poetry however, where over 300 lines were still rejected. We therefore use an ensemble with HypheNN, Pyphen and heuristic corrections for German.

4 Manual Annotation

We manually annotate binary syllable prominence (meter), foot boundaries, caesuras, and main accents of the line for both languages. In addition, we develop an extensive set of regular expressions that capture the verse measure of a line with a human-in-the-loop approach. As noted before, prosodic annotation allows for a certain amount of freedom of expression and (contextual) ambiguity, where several interpretations (performances) can be equally plausible.The eventual quality of annotated data can rest on a multitude of factors, such as the extent of training of annotators, the annotation environment, the choice of categories to annotate (pitch, prominence, contrast, etc.), and the personal preference of subjects (Mo et al., 2008; Kakouros et al., 2016).

See examples (1) and (2) for two annotated lines from Georg Trakl's 'Vorstadt im Föhn', written in regular iambic pentameter. These illustrate the basic annotations. Note that the rhythmic structure of the two lines is fairly different, even though the meter is identical.

met="-+|-+|-+|-+|-+|" free="020:0010202:"
(1) Ein **Flüs**tern, **das** in **trüb**em **Schlaf** er**trinkt**.
A whisper that in drowsy sleep.DAT drowns
A whisper that drowns in a drowsy sleep.

met="-+|-+|-+|-+|-+|" free="0001020:102:"
(2) Die **mit** den **warm**en **Win**den **steigt** und **sinkt**.
That with the warm winds rises and sinks.

Three university students of linguistics/literature were involved in the manual annotation process. They annotated by silent reading of the poetry, largely following an intuitive notion of speech rhythm, as was the modus operandi in related work (Estes and Hench, 2016). The annotators incorporated philological knowledge to recognize instances of poetic license, i.e., knowing how the piece is supposed to be read rather than following intuition. Especially the annotation accuracy of binary syllable stress and foot boundaries benefits from recognizing the schematic consistency of repeated verse measures, license through rhyme, or particular stanza forms (e.g., odes).

[19] As features, we use the word form, the preceeding and following two words and POS tags, and also orthographic information like capitalization, character prefixes and suffixes of length 1, 2, 3 and 4.

[20] https://github.com/alexestes/SonoriPy, https://github.com/henchc/syllabipy

[21] pyphen.org

[22] github.com/msiemens/HypheNN-de

[23] https://github.com/UKPLab/emnlp2017-bilstm-cnn-crf

[24] Syllabipy determines boundaries based on the sonority principle, Pyphen uses the Hunspell dictionaries, and HypheNN is a simple feed forward network that is trained on character windows and whether a boundary occurs in the middle.

[25] For German, wiktionary contains 398.482 hyphenated words, and 130.000 word forms in CELEX. Unfortunately, German CELEX does not have proper umlauts, and models trained on these were not suitable for poetry. For English, wiktionary only contains 5.142 hyphenated words, and 160.000 word forms in CELEX.

	Syllable stress	Syllable feet	Whole Line stress	Whole Line feet
$\mathrm{DE}_{corr.}$	.98	.87	.94	.71
DE_{blind}	.98	.79	.92	.71
EN_{blind}	.94	.95	.87	.88

Table 4: Cohen Kappa Agreement for Binary Stress and Foot Boundaries. Corr. is the agreement of the first version with the corrected version. Blind means that annotators did not see other annotation.

4.1 Annotation Layers & Evaluation

In the following, we describe our annotation layers and evaluation across annotators by calculating Cohen's Kappa. To capture different granularities of correctness, we calculated agreement on syllable annotation (stress), between syllables (for feet or caesura), and on full lines (whether the entire line is correct given a certain feature).

The following example by Percy Blythe Shelley shows the eventual annotation layout that is used for the experiments, including the 'measure' that was derived with regular expressions from the meter line. 'Syll' signifies the position of the syllable in a word.[26] We removed punctuation with a focus on measures.[27]

```
# tok  met ft pos syll csr main smsr    measure     met_line

1  Look  +  .  VB   0   .  1  iambic  i.penta.inv +--+-+-+-+
2  on    -  .  IN   0   .  0  iambic  i.penta.inv +--+-+-+-+
3  my    -  .  PRP$ 0   .  0  iambic  i.penta.inv +--+-+-+-+
4  works +  :  NNS  0   :  2  iambic  i.penta.inv +--+-+-+-+
5  ye    -  .  PRP$ 0   .  0  iambic  i.penta.inv +--+-+-+-+
6  Might +  :  NNP  1   .  1  iambic  i.penta.inv +--+-+-+-+
7  y     -  .  NNP  2   :  0  iambic  i.penta.inv +--+-+-+-+
8  and   +  :  CC   0   .  0  iambic  i.penta.inv +--+-+-+-+
9  de    -  .  VB   1   .  0  iambic  i.penta.inv +--+-+-+-+
10 spair'+  :  VB   2   :  1  iambic  i.penta.inv +--+-+-+-+
```

METER: Binary syllable prominence. In poetry, meter is the basic prosodic structure of a verse or lines in verse. We distinguish metrical structure from rhythmic structure. The underlying (abstract) meter consists of a sequence of beat-bearing units (syllables) that are either prominent or non-prominent, i.e., metrical prominence is conceived of as a binomial categorical variable. Non-prominent beats are attached to prominent ones to build metrical feet (e.g. iambic or trochaic ones). This metrical structure is the scaffold, as it were, for the linguistic rhythm.

We use the term 'binary meter' (met) for a notation of binary syllable prominence (+/-), meaning that a syllable can be either stressed or unstressed. This is annotated bottom up, where first the stress of syllables is determined and then a grouping according to foot boundaries is assigned.

FOOT: A foot is the grouping of metrical syllables. The metre (or measure) of a verse can be described as a sequence of feet, each foot being a specific sequence of syllable types. It is denoted with the pipe symbol (|) in the metrical annotation.

Agreement Meter & Foot: The meter annotation for the German data was first done in a full pass by a graduate student. A second student then started correcting this annotation with frequent discussions with the first author. While on average the agreement scores for all levels of annotation suggested reliable annotation after an initial batch of of 20 German poems, we found that agreement on particular poems was far lower than the average, especially for foot boundaries. Therefore, we corrected the whole set of 153 German poems, and the first author did a final pass. The agreement of this corrected version with the first version is shown in Table 4 in the row $\mathrm{DE}_{corr.}$. To check whether annotators also agree when not exposed to pre-annotated data, a third annotator and the second annotator each annotated 10 diverse German poems from scratch. This is shown in DE_{blind}. For English, annotators 2 and 3 annotated 6 poems blind and then split the corpus.

Notably, agreement on syllables is fairly acceptable, while German feet are a bit problematic. We calculated agreement on all 153 poems and 14 poems had an overall $\kappa < .6$. Close reading revealed that disagreement on poems with κ around .8 is caused by faulty guideline application. Poems with scores lower than $\kappa < .6$ exhibit ambiguous rhythmical structure (multiple annotations are acceptable) and/or schema invariance, where a philological eye considers the whole structure of the poem and a naive annotation approach does not render the intended prosody correctly.

As an example for ambiguous foot boundaries, the following poem, Schiller's 'Bürgschaft', can be set in either *amphibrachic* feet, or as a mixture of *iambic* and *anapaestic* feet. Such conflicting annotations were discussed by Heyse (1827), who finds that in the greek tradition the *anapaest* is preferable, but a 'weak amphibrachic gait' allows for a freer rhythmic composition. This suggests that Schiller was breaking with tradition.

[26] 0 for monosyllaba, otherwise index starting at 1.

[27] Although punctuation is a good signal for pauses in speech

```
(Foot Boundary Ambiguity) Schiller, 'Die Bürgschaft'

(1) met="-+-|-+-|-+-|"
    Ich lasse |    den Freund   dir | als Bürgen, |
(2) met="-+|--+|--+|-"
    Ich las   | se den Freund | dir   als Bürg | en,
Transl.: I leave this friend to you as guarantor

(1) met="-+-|-+-|-+-|"
    Ihn magst   du, | entrinn'   ich, | erwürgen. |
(2) met="-+|--+|--+|-"
    Ihn magst | du,   entrinn' | ich,    erwür | gen.
Transl.: Him you may strangle if I escape.

(1) (amphibrach)
(2) (iambus / anapaest)
```

CAESURA. Caesuras are pauses in speech. While an end-caesura is the norm (to pause at the line break) often there are natural pauses in the middle of a line rather, occasionally the line also runs on without a pause. Caesurae (csra) are denoted with a colon in the rhythm (main accent) annotation.

MAIN ACCENTS. Main accents are intended to reveal a freer rhythm than the rigid metrical structure. These accents are annotated top down, where first the rhythmical group is determined by marking caesura boundaries and then assigning primary accents (2), side accents (1) and weak syllables (0). This annotation differs from meter operating top-down from rhythmic segments to find natural speech rhythm, while meter is bottom-up. Main accents do not need to coincide with meter. Rhythm emerges from the filling of metrical structure with lexical material. Lexical material comes with n-ary degrees of stress, depending on morphological, syntactic, and information structural context; A (practical) ternary notation merely approximates this. Moreover, the linguistic material engenders chunking of metered verse lines according to their syntactic structure. The caesurae are therefore part of the rhythmic structure, the metrical beat does not, by itself, presuppose caesurae.

	Syllable		Whole Line	
	stress	caesura	stress	caesura
DE_{blind}	.84	.92	.59	.89
EN_{blind}	.67	.86	.35	.64

Table 5: Cohen Kappa Agreement for Main Accents and Caesura

Table 5 lists the agreement figures for main accents and caesurae. It shows that caesurae can be fairly reliably detected through silent reading in both languages. On the other hand, agreement on main accents is challenging, especially for non-native speakers of English, as both annotators are native German speakers. Future research should implement proper guidelines with native speakers. Figure 3 shows the confusion of main accents for German. While 0s are quite unambiguous, it is not always clear when to set a primary or side accent.

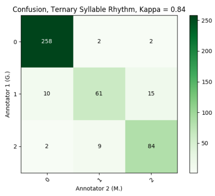

Figure 3: Confusion of German Main Accents

VERSE MEASURES. We develop an extensive set of regular expressions to determine the measure of a line from its metrical annotation. We orient ourselves with the handbook of Knörrich (1971). We implement the usual measures (by feet) like `iambic` (−+), `trochaic` (+−), `dactyls` (+−−), `anapaest` (−−+), `amphibrachs` (−+−), the `alexandrine` and the `dactylic hexameter`, and some ode forms. We implement each measure in different lengths.[28] Table 6 lists the most frequent labels for each language without length, so-called short measure (smsr).

English		German	
freq.	smsr	freq.	smsr
2096	iambic	1976	iambic
490	troch	793	troch
306	anapaest	258	amphibrach
255	amphibrach	206	alexandrine
248	daktyl	76	daktyl
152	hexameter	72	anapaest
91	prosodiakos	26	asklepiade
52	other	17	pherekrateus
35	alexandrine	14	glykoneus

Table 6: Most frequent verse measures in small English and German corpus, without length.

5 Experiments

In the following, we carry out experiments to learn the previously annotated features and determine their degree of informativeness for each other with a multi-task setup.

[28] `di-`, `tri-`, `tetra-`, `penta-`, `hexa-` and `septameter` (by number of stressed syllables). Odes forms include the `asklepiade` (+−+−−+−−+−+). We also annotate inversions whenever the first foot is inverted, e.g., when the first foot in a iambic line is trochaic: (+−−+−+−+), and we allow the insertion of unstressed syllables, making the verse 'relaxed', e.g., (−+−+−−+−+) without changing the meter, and distinguish these from choliambic endings (−+−+−+−−+). We also allow the option of female (unstressed) endings/cadences (−+−+−+−?).

5.1 Prosodic Multi-Task Learning

To learn the previously annotated features, we implement a nominal CRF as baseline (as was used for POS tagging in section 3.3.1), and we use a biLSTM-CRF with pretrained syllable embeddings in a multi-task setup. Instead of char embeddings for syllabification, only now we use pre-trained word2vec syllable embeddings that were trained on the large poetry corpora from section 3.1.

	English		German	
	syll. acc	line acc	syll. acc	line acc
CRF	.922	.478	.941	.553
bilstm-crf	.955	.831	.968	.877
BERT	.850	.371	.932	.498

Table 7: Best Classifiers of each architecture by Languages

We use three layers of size 100 for the LSTM and do the final label prediction with a linear Chain CRF, and apply variable dropout of .25 at both input and output. No extra character encodings are used (as these hurt both speed and accuracy). We also classify meter with BERT,[29] but find that it cannot reach the LSTM performance. We perform a three fold cross validation and average the results with a 80/10/10 split. All results are reported on the test set. See Table 7 for a comparison of best models by architecture. We outperform recent research by a decent margin. Agirrezabal et al. (2019) achieved .62 line accuracy for English on the for-better-for-verse dataset. Our English model achieves .83 line accuracy, most likely due to the pretrained syllable embeddings and more consitent and diverse data.

We attribute the gap of BERT to the LSTM model to the lack of proper syllable representation of BERT. A multilingual BERT model achieves .90 syll. accuracy, inbetween the monolingual models in Table 7. We also experiment with framing the task as document classification, where BERT should learn the proper verse label for a given line. Training on the small datasets only achieves around .22 F1-macro and .42 F1-micro for English. We then tagged 200.000 lines of the large English corpus with a LSTM model and trained BERT on this larger dataset, achieving only .48 F1-macro and .62 F1-micro.

[29](with huggingface code)

5.2 Pairwise Joint Feature Learning:

In Table 8 we investigate the influence of jointly learned additional output features for the classification of out features, running two annotation layers at the same time, and finally run all annotations at the same time. We choose the German dataset here, as the annotation is generally more reliable. Note that POS is on syllable level, but jointly learning the syllable position ('syll in' word) is not beneficial. We find that a global fine-grained verse measure label benefits the meter tagging, while feet strongly benefit from syllable stress (meter, main accent) as might be expected and also benefit from the agglomeration of all other features (+all).

	met	feet	syllin	pos	csra	m.ac
alone	.964	.871	**.952**	.864	.912	.866
+met	-	**.922**	.949	.856	.918	.869
+feet	.961	-	.948	.853	.917	.863
+syllin	.966	.900	-	.860	.919	.867
+pos	.956	.879	**.953**	-	**.924**	**.879**
+csra	.961	.886	.940	.855	-	.868
+m.ac	.964	**.915**	.948	.865	.915	-
+smsr	.965	.884	.942	.854	.918	.868
+fmsr	**.968**	.899	.938	.858	**.926**	.868
+m_line	.966	.882	.937	.853	.919	.868
+all	.967	**.930**	.947	.790	.919	.870

Table 8: Pairwise Joint Feature Learning

The exchange between caesuras and main accents is marginal, but caesuras benefit from POS, syllable position and global measures (in the absence of punctuation), showing that they are integral to poetic rhythm and fairly dependent on syntax.

6 Conclusion

We have created large poetry corpora for English and German to support computational literary studies and annotated prosodic features in smaller corpora. Our evaluation shows that a multitude of features can be annotated through silent reading, even though perceiving free rhythm for non-native speakers is challenging. Finally, we have performed first experiments with a multi-task setup to find beneficial relations between those prosodic features. Learning metrical annotation, including feet and caesurae, largely benefits from a global label, while foot boundaries also benefit from any joint learning with syllable stress and all features alltogether, surpassing the human upper bound.

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, Osaka, Japan. The COLING 2016 Organizing Committee.

Manex Agirrezabal, Iñaki Alegria, and Mans Hulden. 2019. A comparison of feature-based and neural scansion of poetry. *RANLP 2019, arXiv preprint arXiv:1711.00938*.

Mark Algee-Hewitt, Ryan Heuser, Maria Kraxenberger, JD Porter, Jonny Sensenbaugh, and Justin Tackett. 2014. The stanford literary lab transhistorical poetry project phase ii: Metrical form. In *DH*.

Arto Anttila, Timothy Dozat, Daniel Galbraith, and Naomi Shapiro. 2018. Sentence stress in presidential speeches. *Lingbuzz Preprints*.

MH Beissinger. 2012. Oral poetry. *The Princeton encyclopedia of poetry and poetics*, pages 978–81.

Thomas Carper and Derek Attridge. 2020. *Meter and meaning: an introduction to rhythm in poetry*. Routledge.

Sebastian Donat. 2010. *Deskriptive Metrik*. Innsbruch/Wien/Bozen: StudienVerlag.

Alex Estes and Christopher Hench. 2016. Supervised machine learning for hybrid meter. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, pages 1–8.

Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1191.

Erica Greene, Tugba Bodrumlu, and Kevin Knight. 2010. Automatic analysis of rhythmic poetry with applications to generation and translation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 524–533.

Susanne Haaf, Alexander Geyken, and Frank Wiegand. 2014. The dta "base format": A tei subset for the compilation of a large reference corpus of printed text from multiple sources. *Journal of the Text Encoding Initiative*, (8).

Thomas Haider, Steffen Eger, Evgeny Kim, Roman Klinger, and Winfried Menninghaus. 2020. Po-emo: Conceptualization, annotation, and modeling of aesthetic emotions in german and english poetry. *arXiv preprint arXiv:2003.07723*, pages 1652–1663.

Thomas Haider and Jonas Kuhn. 2018. Supervised rhyme detection with siamese recurrent networks. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 81–86.

Johann Christian August Heyse. 1827. Theoretisch-praktische grammatik oder lehrbuch zum reinen und richtigen sprechen, lesen und schreiben der deutschen sprache.

Jack Hopkins and Douwe Kiela. 2017. Automatically generating rhythmic verse with neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 168–178.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence-to-sequence models. *arXiv preprint arXiv:1707.01161*.

Harsh Jhamtani, Sanket Vaibhav Mehta, Jaime Carbonell, and Taylor Berg-Kirkpatrick. 2019. Learning rhyming constraints using structured adversaries. *arXiv preprint arXiv:1909.06743*.

Sofoklis Kakouros, Joris Pelemans, Lyan Verwimp, Patrick Wambacq, and Okko Räsänen. 2016. Analyzing the contribution of top-down lexical and bottom-up acoustic cues in the detection of sentence prominence. *Proceedings Interspeech 2016*, 8:1074–1078.

Paul Kiparsky. 2020. Metered verse. *Annual Review of Linguistics*, 6:25–44.

Otto Knörrich. 1971. *Die deutsche Lyrik der Gegenwart*, volume 401. Kröner.

Jey Han Lau, Trevor Cohn, Timothy Baldwin, Julian Brooke, and Adam Hammond. 2018. Deep-speare: A joint neural model of poetic language, meter and rhyme. *arXiv preprint arXiv:1807.03491*.

Yoonsook Mo, Jennifer Cole, and Eun-Kyung Lee. 2008. Naïve listeners' prominence and boundary perception. *Proc. Speech Prosody, Campinas, Brazil*, pages 735–738.

Borja Navarro, María Ribes Lafoz, and Noelia Sánchez. 2016. Metrical annotation of a large corpus of spanish sonnets: representation, scansion and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4360–4364.

Ani Nenkova, Jason Brenier, Anubha Kothari, Sasha Calhoun, Laura Whitton, David Beaver, and Dan Jurafsky. 2007. To memorize or to predict: Prominence labeling in conversational speech. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 9–16.

Allison Parrish. 2018. A Gutenberg Poetry Corpus.

Petr Plecháč. 2020. Relative contributions of shakespeare and fletcher in henry viii: An analysis based on most frequent words and most frequent rhythmic patterns. *Oxford Journal of Digital Humanities (DSH), arXiv preprint arXiv:1911.05652*.

Thomas Proisl and Peter Uhrig. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin. Association for Computational Linguistics (ACL).

Sravana Reddy and Kevin Knight. 2011. Unsupervised discovery of rhyme schemes. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 77–82.

Andrew Rosenberg. 2010. Autobi-a tool for automatic tobi annotation. In *Eleventh Annual Conference of the International Speech Communication Association*.

Pablo Ruiz Fabo, Helena Bermúdez Sabel, Clara Martínez Cantón, and Elena González-Blanco. 2020. The diachronic spanish sonnet corpus: Tei and linked open data encoding, data distribution, and metrical findings. *Digital Scholarship in the Humanities*.

Aarne Talman, Antti Suni, Hande Celikkanat, Sofoklis Kakouros, Jörg Tiedemann, and Martti Vainio. 2019. Predicting prosodic prominence from text with pre-trained contextualized word representations. *arXiv preprint arXiv:1908.02262*.

Rafael Valle, Kevin Shih, Ryan Prenger, and Bryan Catanzaro. 2020. Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis. *arXiv preprint arXiv:2005.05957*.

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.