Vulgaris: Analysis of a Corpus for Middle-Age Varieties of Italian Language

Andrea Zugarini[1,2] and **Matteo Tiezzi**[2] and **Marco Maggini**[2]
[1]DINFO, University of Florence
[2]DIISM, University of Siena
andrea.zugarini@unifi.it, {mtiezzi,maggini}@diism.unisi.it

Abstract

Italian is a Romance language that has its roots in Vulgar Latin. The birth of the modern Italian started in Tuscany around the 14th century, and it is mainly attributed to the works of Dante Alighieri, Francesco Petrarca and Giovanni Boccaccio, who are among the most acclaimed authors of the medieval age in Tuscany. However, Italy has been characterized by a high variety of dialects, which are often loosely related to each other, due to the past fragmentation of the territory. Italian has absorbed influences from many of these dialects, as also from other languages due to dominion of portions of the country by other nations, such as Spain and France. In this work we present Vulgaris, a project aimed at studying a corpus of Italian textual resources from authors of different regions, ranging in a time period between 1200 and 1600. Each composition is associated to its author, and authors are also grouped in families, i.e. sharing similar stylistic/chronological characteristics. Hence, the dataset is not only a valuable resource for studying the diachronic evolution of Italian and the differences between its dialects, but it is also useful to investigate stylistic aspects between single authors. We provide a detailed statistical analysis of the data, and a corpus-driven study in dialectology and diachronic varieties.

1 Introduction

Understanding the evolution of a language is a challenging problem. When a language originated? What are the influences coming from dialects and other languages? These are crucial questions that the study of language evolution aims to face.

Natural Language Processing techniques are powerful tools that can support researchers in the analysis of dialects and diachronic language varieties (Zampieri and Nakov, 2020; Ciobanu and Dinu, 2020). There exists several lines of research that approach the problem of defining distances between languages or varieties. Linguistic phylogenetics (Borin, 2013) aim at determining a rooted tree to describe the evolution of a group of languages or varieties. Trees are built based on the so called *lexicostatistics* technique, that takes into account words with common origin to determine a taxonomic organization of the languages. Language distance approaches instead rely on the distributional hypothesis of words and require cross-lingual corpora. Similarity is based on word co-occurrences (Asgari and Mofrad, 2016; Liu and Cong, 2013), or using perplexity-based methods (Basile et al., 2016; Gamallo et al., 2017; Campos et al., 2018; Campos et al., 2020). Perplexity is estimated from Language Models, typically n-grams LMs of characters, trained on one corpus and evaluated on another variety.

Differently from previous work, we consider Neural Language Models (NLMs) (Bengio et al., 2003; Mikolov et al., 2010), that are more robust estimators well known for their generalization capabilities and currently the state-of-the-art approaches in Language Modeling tasks. There is a vast literature on NLMs. Many works also address the problem of character language modeling (Jozefowicz et al., 2016; Hwang and Sung, 2017) or character-aware LMs (Marra et al., 2018; Kim et al., 2016).

In this work we focus on Italian, a Romance language derived from Vulgar Latin. The uniquely fragmented political situation that occurred in Italy during the middle age makes Italian an extremely

Proceedings of the 7th VarDial Workshop on NLP for Similar Languages, Varieties and Dialects, pages 150–159
Barcelona, Spain (Online), December 13, 2020

variegate and complex case of study, rich of dialects that are still spoken nowadays. We consider a corpus of medieval text collections, with the purpose of easing the research activity on diachronic varieties. Moreover, the dataset can be also a valid resource to study the problem Text Generation in low-data and variegate styles conditions. Similar corpora have been already collected for other languages. Colonia (Zampieri and Becker, 2013), is a Portuguese diachronic dataset of about 5 milion tokens grouped by century in five sub-corpora. In (Campos et al., 2020) they gathered three corpora for English, Spanish and Portuguese.

In summary, the contributions of this paper are: (1) we present a project, *Vulgaris*, that studied a text corpus consisting of vulgar Italian language literary resources, organized in such a way to ease language research, (2) studying the historical and geographical background, the statistical properties of the collected data and its composition and (3) deepening our analysis through a corpus-driven study in dialectology and diachronic varieties exploiting perplexity-based distances. In particular, we introduce Neural Language Models to estimate the perplexity and provide a new indicator, named as Perplexity-based Language Ratio (PLR), to analyse the historical evolution process of the varieties.

The rest of the paper is organized as follows. In Section 2, we describe in detail *Vulgaris*, its composition and we report several statistics on it. Then, in Section 3, we introduce perplexity-based metrics combined with neural language models that are then used to carry out experiments on the diachronic varieties within *Vulgaris*.

2 Vulgaris

The main goal of project *Vulgaris*[1] is the analysis of the diachronic evolution and variance of the vulgar italian language. In order to do so, we collected an heterogeneous literary text corpus, comprehensive of poetry, prose, epistles and correspondence by the most important Italian authors ranging from the dawn of the vulgar language to the Reinassance Age. Henceforth, for compactness, we refer to such data as *Vulgaris*. The dataset represents a fundamental timeframe for the Italian language, including the first steps and diachronic evolutions departing from the Latin language. Moreover, through *Vulgaris* it is possible to gain evidence of the early language fragmentation deriving from the complex historical geo-political context of the Middle Age.

2.1 Hystorical background and families

The earliest years of the 13th century were characterized by a novel and complex civilisation. The rise of medieval Communes, associations among citizens of towns belonging to the same social class, influenced the rise of a novel school of secular thought increasingly unhindered by the religious influences. For these reasons, along with the establishment of the first universities, beside latin literature the vulgar Italian language started to appear in various literary works. The heterogeneous political and geographical context led to a linguistic fragmentation, characterized by various contact points. The first literary evidence of vulgar poetic, which we denote as belonging to the **Archaic text** family, is a collection of verses still connected with religious and moral themes, written in regions of the central Italy, in particular Umbria and Tuscany. Amongst the main authors, we mention *Francesco d'Assisi*. Inspired by this works, in the middle of the century (about 1250) some vulgar authors (e.g. *Jacopone da Todi* et al.), in the same geographical zone, composed several **Laude**, enriching the religious and mystical poetry theme.

The **Northern Didactic poetry** family, flourished in the same years, was influenced by these religious and moral guidelines. We point out *Bonvesin da la Riva* from Milan and *Giacomino da Verona* among the representatives, having the goal to instruct readers about morality, philosophy and doctrine.

The prosperous and thriving Imperial court of Federico II fostered the birth of a **Sicilian School** (1230-1250), where the figure of an angelic woman and the stereotype of love play a central role. This group laid the foundations of modern poetry, introducing a specific metric and organization in *Stanzas*, creating Sonnets and unifying the language lexicon and structure. Among the authors we highlight *Giacomo da Lentini* and *Pier della Vigna*.

[1]The project is available at `https://sailab.diism.unisi.it/vulgaris/`.

With the death of Federico II, the cultural axis moved to Tuscany, thanks to the proliferation of Communes. Differently from the Sicilian School we cannot talk of a unique literary school. Indeed, in several important cities, such as Pisa, Lucca, Arezzo, Siena, other than Florence, which became only afterwards the most important cultural center, emerged themes inspired by the Sicilian similar ones.

The **Northern/Tuscan Courtly poetry** arises from poets belonging to the Sicilian school who moved after the decadence of the Svevian Empire, influencing the themes and style of local authors (*Guittone D'Arezzo*, *Bonagiunta Orbicciani*, *Compiuta Donzella*). In the meanwhile, **Central Italy Didactic poetry** (*Brunetto Latini*) and **Realistic Tuscan poetry** (*Cecco Angiolieri, Folgòre da San Gimignano, Cenne de la Chitarra*) emerged, differentiated by the themes, goal of the poetry, and style. Departing from the literature inspired by court life, a more popular and playful genre, the **Folk and Giullaresca Poetry**, was mainly due to jesters such as *Ruggieri Apuliese*.

Finally, thanks to the influence of Sicilian School and Tuscan poetry, the **Stilnovisti** family (*Guido Guinizzelli*, *Dante Alighieri*, *Guido Cavalcanti*, *Lapo Gianni*, *Gianni Alfani*, *Dino Frescobaldi*, *Cino da Pistoia*) and some authors close to them (**Similar to Stilnovisti** - *Lippo Pasci de Bardi*) evolved and refined the poetry of their predecessors. Metaphors, a noble symbolism and introspection characterize this movement, which was born in Bologna and developed in Florence reaching its climax.

Boccaccio and **Petrarca** compose, together with Dante, the three *Crowns* of Italian literature. Their poetry and prose are inspired by *Dolce Stil Novo*, with an evolution toward a more wordily thematic, rather than spiritual. Their linguistic style is the offspring of an evolved society.

The works developed by these families highly influenced following authors. In particular, **Ariosto** and **Tasso**, at the beginning of the 16th century, were deeply inspired by Petrarca and the Stilnovisti, respectively. However, their different temporal context was reflected in their literary works.

Therefore, the vulgar Italian language, starting from the beginning of the 13th century, became more and more popular amongst various authors, evolving during the following years in several families, which we summarize in Fig.1. The highly fragmented geo-political context gave rise to different schools, groups, communities (depicted in Fig. 1) and hence many language varieties, dialects, that even nowadays are noticeable.

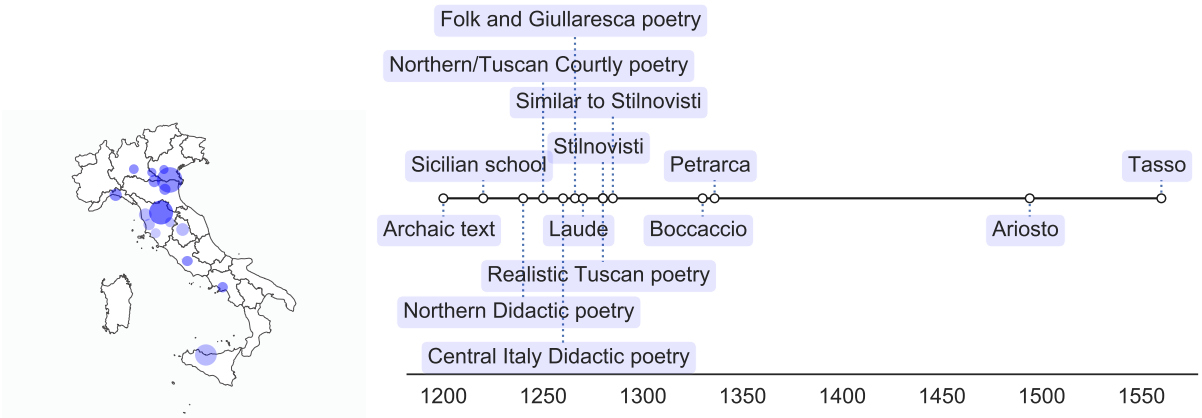

Figure 1: On the left side, a geographical map of the Italian peninsula with the biggest cultural centers marked (the map is attributed to `https://freevectormaps.com/italy/IT-EPS-01-0004`). On the right, a timeline representing the temporal sequence of the different families we described in the main text.

Through the Vulgaris project, we aim to provide a rich resource to analyze the diachronic evolution of the early Italian language, in particular in poetry, prose and correspondence texts.

Family	*Authors*	*#Texts*	*#Poetry*	*#Prose*
Archaic text	Francesco d'Assisi, Ritmo Laurenziano, Ritmo Cassinese	5	5	-
Sicilian School	Giacomo da Lentini, Guido delle Colonne, Pier della Vigna, Pronotaro da Messina	46	46	-
Northern Didactic poetry	Girardo Patecchio Da Cremona, Bonvesin Da La Riva, Giacomino Da Verona, Anonimo Genovese	29	29	-
Northern/Tuscan Courtly poetry	Guittone D'Arezzo, Bonagiunta Orbicciani, Chiaro Davanzati, Monte Andrea Da Firenze	101	101	-
Central Italy Didactic poetry	Brunetto Latini,Garzo, Detto Del Gatto Lupesco Dal Bestiario Moralizzato Di Gubbio	8	8	-
Folk and Giullaresca poetry	Ruggieri Apugliese, Castra Fiorentino, Matazone Da Caligano, Rime Dei Memoriali Bolognesi	23	23	-
Laude	Jacopone Da Todi, Laude Cortonesi, Lauda Dei Servi Della Vergine	41	41	-
Stilnovisti	Guido Guinizzelli, Guido Cavalcanti, Cino da Pistoia, Dante Alighieri, Lapo Gianni	769	704	65
Realistic Tuscan poetry	Rustico Filippi, Cecco Angiolieri, Folgore da San Gimignano, Cenne de la Chitarra	69	69	-
Similar to Stilnovisti	Dante's Friend, Lippo Pasci de' Bardi	709	70	-
Boccaccio	-	1058	296	762
Petrarca	-	872	747	125
Ariosto	-	363	144	219
Tasso	-	3366	1604	1762
Total	-	6820	3887	2933

Table 1: Analysis of the composition of the dataset. We report the families, their most representative authors, total number of provided texts and their distribution in poetry and prose.

2.2 Dataset structure

The examined corpus contains texts retrieved from Biblioteca Italiana[2], a digital library project collecting the most significant texts of the Italian literature, ranging from the Middle Age to the 20th century. The code to retrieve and analyse the data can be found in `https://github.com/sailab-code/vulgaris`. *Vulgaris* provides the following filtered type of information, extracted from the parsed data: **author**, **title**, **collection**, **family**, **type**, **text**. In details, the corpus is composed by **text** produced by 104 **authors** belonging to the 14 **families** described in Section 2. The corpus contains 177 *collections*, consisting of groups of poetry, single poems, personal epistles. Each item of the dataset is a single composition, for instance a poetry, a chapter or a letter. Moreover, we split the resources by the **style** attribute into *poetry* and *prose*, with the latter containing the both the prose and correspondence documents.

The structure of a poetic composition represents an important information in tasks such as Poem Generation (Lau et al., 2018; Zugarini et al., 2019; Zhang and Lapata, 2014). The verse organization of the poetry is encoded by tags denoting each line break `<EOL>` (end of a verse), as well as the end of each stanza `<EOS>`. In the case of prose, only the organization in paragraphs is represented by the tag `<EOS>`.

In Table 1 we report some statistics on the families (first column, ordered by date), including their most representative authors (second column), the total amount of collected texts, divided into poetry and prose (third, fourth and fifth column, respectively). Whilst the older families are underrepresented, families belonging to a later period are mostly characterized by a larger amount of texts. This fact is a good indicator of the diffusion that the Italian language has undergone during this timeline.

The corpus investigated in *Vulgaris* is extremely heterogeneous and composed by 4 million word

[2] `http://bibliotecaitaliana.it/`

occurrences, whose texts have been written by authors from a wide range of geographical regions and time periods, as shown in Figure 1. In Table 2 we summarize some statistics on the total amount of word occurrences, the number of unique words and the average occurrences per word for each text **type**. The total number of words in poetry and prose is almost balanced, whereas their composition is remarkably different. Indeed, poetry has a richer lexicon than prose, containing almost twice unique words.

To depict the contribution of each family to the dataset, Figure 2 reports the total number of word occurrences for each family and the poetry/prose proportion. Once more, these statistics confirm the increasing spread of the Italian language. We can also notice how vulgar spread. Initially, it was mainly used in poetry and only later vulgar prosaic forms appeared. Only 5 out of 14 families contain prose, and, as we can see from the timeline in Fig. 1, they correspond to the latest families.

	Global	*Poetry*	*Prose*
# word occurrences	4090166	1925838	2164328
# unique words	180450	136195	69135
Avg occurrences per word	22.67	14.14	31.31

Table 2: Statistics on words for each text category.

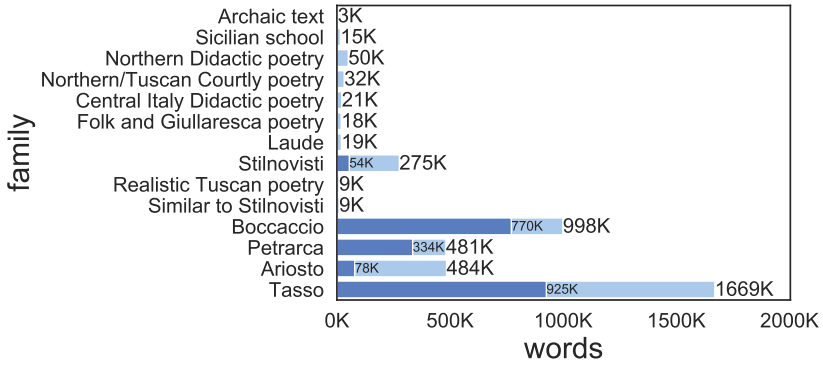

Figure 2: The figure reports the total amount of word occurrences for each family (both in poetry and prose), at the right of each family bin. The darker blue bar denotes the portion of word occurrences in prose texts only.

Finally, in the top row of Fig. 3, we report the average distribution of the text length, in both the styles (i.e, *poetry* on the left and *prose* on the right) among all the families. The bottom row of Fig. 3 shows the average number of words contained in each collection, hence texts having similar characteristics or theme.

3 Analysis of Language Varieties in Vulgaris

Vulgaris texts span over a time period of about four centuries. The diachronic varieties within the dataset are measured in terms of perplexity-based distances, taking into account the different centuries as a reference for the comparison.

3.1 Perplexity-based Language Distance

The quality of a Language Model (LM) is assessed in terms of perplexity. A good estimation by a language model for a given corpus will yield low perplexity values. LMs and perplexity have been already exploited to provide a distance between language corpora (Gamallo et al., 2017), and this approach has been effectively applied for language discrimination and the analysis of historical varieties (Campos et al., 2018; Campos et al., 2020).

Let us consider two language corpora, namely $L1$ and $L2$, and let $\texttt{LM}_{L1}$, $\texttt{LM}_{L2}$ be two language models trained on $L1$ and $L2$, respectively. We can argue that the more the corpora are related to each other,

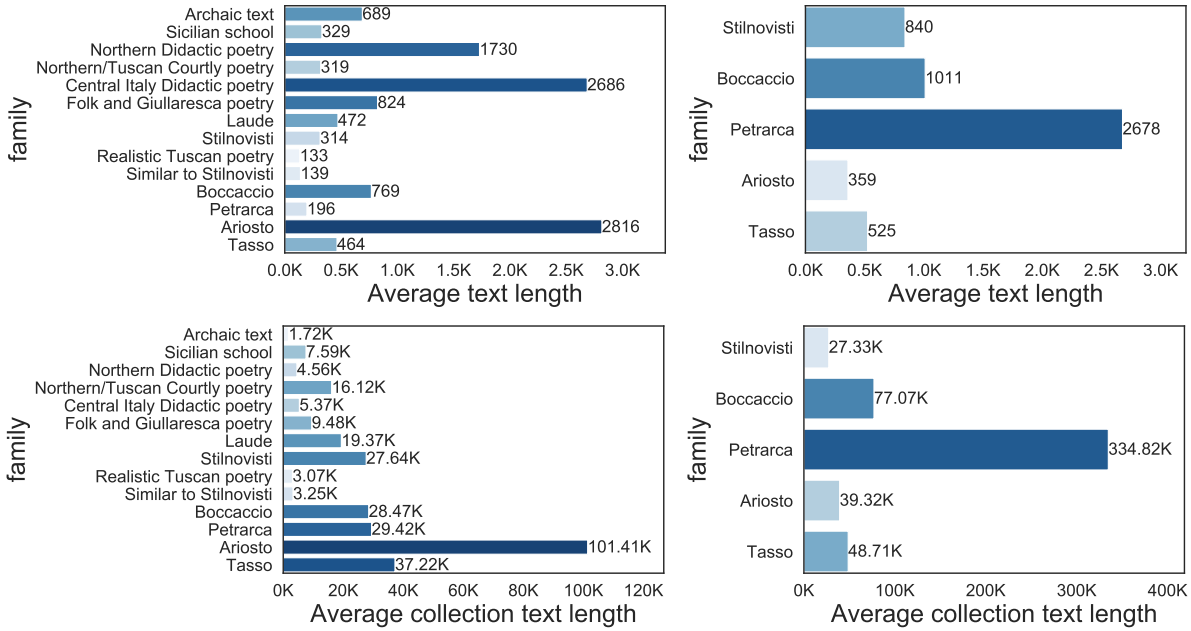

Figure 3: In the top row, the average text length of poetry (left) and prose (right) texts. On the bottom, average number of words contained in each collection in poetry (left) and prose (right).

the more accurate is the estimate provided by the LM trained on one language when evaluated on the other. By denoting the two measures of perplexity as $pp_{L1 \to L2}(L2, \texttt{LM}_{L1})$ and of $pp_{L2 \to L1}(L1, \texttt{LM}_{L2})$, the Perplexity-based Language Distance (PLD) is defined in (Gamallo et al., 2017) as the average of these two values:

$$PLD(L1, L2) = \frac{pp_{L1 \to L2}(L2, \texttt{LM}_{L1}) + pp_{L2 \to L1}(L1, \texttt{LM}_{L2})}{2}. \quad (1)$$

This metric copes with the fact that $pp_{L1 \to L2}(L2, \texttt{LM}_{L1})$ and $pp_{L2 \to L1}(L1, \texttt{LM}_{L2})$ are not symmetric, mostly because LMs are trained and tested on different data distributions. However, the asymmetry in these values can be a good indicator of the language evolution on diachronic/dialect varieties, since it can enlighten either a language compression/simplification or a language expansion over time. Indeed, in the process of language unification, words are reduced, and dialectal expressions are suppressed, thus reducing the overall richness of the language. Hence, we consider also the following Perplexity-based Language Ratio (PLR):

$$PLR(L1, L2) = \frac{pp_{L1 \to L2}(L2, \texttt{LM}_{L1})}{pp_{L2 \to L1}(L1, \texttt{LM}_{L2})}. \quad (2)$$

PLR values greater than 1 indicate that $L1$ is likely to be a more various language than $L2$, whereas values less than 1 are likely to indicate $L2$ as the more complex language.

3.2 Conditional Language Modeling

In this section we briefly describe the structure of the language models that have been exploited in the experimental evaluation. Let us consider a sequence of tokens $\boldsymbol{x} = (x_1, \ldots, x_n)$ from a text corpus. The following description is general for any sequence of tokens $\boldsymbol{x}$ regardless their kind, e.g. words, characters or any token piece. The goal of the LM is to estimate the joint probability $p(\boldsymbol{x})$, that is factorized with the product of conditional probabilities as follows,

$$p(\boldsymbol{x}) = \prod_{i=1}^{m} p(x_i | x_{i-1}, \ldots, x_1), \quad (3)$$

A Neural Language Model (NLM) (Bengio et al., 2003) estimates the conditional probability $p(x_i | x_{i-1}, \ldots, x_1)$ in Equation (3) with a Neural Network. We extend Equation (3) by adding other

Figure 4: Two dimensional t-SNE representation of sentences' state h_T. Different colours indicate different groups, dots for poetry, crosses for prose.

features of the text to condition the NLM. In particular, we leverage the external meta information about author a, family f and kind of composition k (prose or poetry) available in the dataset. Hence Equation (3) becomes:

$$p(\boldsymbol{x}) = \prod_{i=1}^{m} p(x_i | x_{i-1}, \ldots, x_1, a, f, k). \tag{4}$$

We model the distribution in Equation 4 by means of a recurrent neural network. Each token from the vocabulary V of size $|V|$ is associated to a latent embedding $\boldsymbol{e}$ of dimension d. The set of the $|V|$ embeddings are collected in the $|V| \times d$ matrix $\mathbf{E}$. In particular, we consider an LSTM cell (Hochreiter and Schmidhuber, 1997) to model the internal recurrent state of the network. At time t the state $\boldsymbol{h}_t$ is updated as follows,

$$\boldsymbol{h}_t = \texttt{LSTM}(\boldsymbol{e}_t, \boldsymbol{h}_{t-1}) \ , \tag{5}$$

The external features (a, f, k) are concatenated to $\boldsymbol{h}_t$ and then linearly projected into a d-dimensional vector $\boldsymbol{s}_t$:

$$\begin{aligned} \boldsymbol{c}_t &= [\boldsymbol{h}_t \circ \boldsymbol{a} \circ \boldsymbol{f} \circ \boldsymbol{k}], \\ \boldsymbol{s}_t &= W \cdot \boldsymbol{c}_t + b, \end{aligned}$$

where $\circ$ is the concatenation operator, and $\boldsymbol{a}, \boldsymbol{f}, \boldsymbol{k}$ the embedding representations associated to author a, family f and kind k, respectively. The probability distribution $\hat{\boldsymbol{y}}_t$ is the output of a softmax layer sharing the weights of the input embeddings to apply a back-projection of the contextual state $\boldsymbol{s}_t$ into the vocabulary space:

$$\begin{aligned} \boldsymbol{o}_t &= \mathbf{E}^T \cdot \boldsymbol{s}_t, \\ \hat{\boldsymbol{y}}_t &= \texttt{softmax}(\boldsymbol{o}_t). \end{aligned}$$

The PLD and PLR are estimated exploiting this conditional neural language model where input tokens are characters. We chose NLMs over n-grams because of their notorious generalization capabilities. More robust LMs estimation will improve the quality of the PLD and PLR measures. The same kind of architecture is used to build and visualize the sequence representations shown in Fig. 4, learnt from a word-based language model on the entire *Vulgaris*.

3.3 Diachronic Variety

The 14 families of *Vulgaris* are arranged in four language corpora, based on their time periods, as shown in Fig. 1. The first group, referred to as XIII, includes all the families belonging to the 13th century. In this century there are 10 out of 14 families of the dataset, making this language variety the most heterogeneous one, including many authors from different areas of the Italian territory. In the second one (XIV), we consider *Petrarca* and *Boccaccio* families/authors, whereas *Ariosto* and *Tasso* constitute the third and forth corpora, respectively, XV-XVI-1 and XV-XVI-2. Clearly the boundaries are not neat, since the activity of some authors may span across two centuries. From Table 3 we can see that the diachronic corpora are unbalanced. Despite the high number of families and authors, the XIII corpus is the less represented one, followed by XV-XVI-1 that is slightly larger. They both are small compared to XIV and XV-XVI-2. However, XIII is also the dataset with lowest average number of occurrences per word, indicating a high variance of the collection caused by the rich variety of styles and authors.

	XIII	XIV	XV-XVI-1	XV-XVI-2
# words	455583	1480379	484276	1669928
dataset proportion (%)	11.14	36.19	11.84	40.83
# unique words	57343	73530	42594	72369
Avg occurrences per word	7.94	20.13	11.37	23.08

Table 3: Number of words and proportions of the four diachronic groups.

As a first qualitative analysis, we trained a word-based conditional NLM on the entire corpus, using a vocabulary of 50000 words. The final cell state h_T of a text sequence $x = (x_1, \cdots, x_T)$ is projected into a 2-dimensional representation using t-SNE (Maaten and Hinton, 2008). Fig. 4 visualizes the 2-d representation of 2000 examples, colored accordingly to the corpus they belong to, and styled differently in case of prose or poetry works. Prose and poetry are easily discriminated by the NLM. The corpus origin is also captured by the NLM, although not completely, suggesting that the diachronic varieties share a similar structure.

Then, both the PLD and the PLR described in Subsection 3.2 are computed for each pair of corpora. For the character LMs, we consider input character sequences with a maximum length of 50. The state h_t has size 256, with $(\boldsymbol{a}, \boldsymbol{f}, \boldsymbol{k})$ of size 16, 16 and 32, respectively. Special tokens delimiting end of sentence, end of verse and white space are included in the vocabulary of characters. For each $L_i \rightarrow L_j$, the network is trained on 90% of the L_i corpus, whereas the remaining 10% is used for early stopping, and it is finally evaluated on the whole L_j.

	XIII	XIV	XV-XVI-1	XV-XVI-2
XIII	3.90	5.38	5.99	6.08
XIV	5.38	3.52	4.76	4.65
XV-XVI-1	5.99	4.76	3.30	4.47
XV-XVI-2	6.08	4.65	4.47	3.28

Table 4: PLD among pairs of diachronic language varieties.

Results are shown in tables 4 and 5. PLD is lower in diachronic varieties closer in time, as expected. Interestingly enough, PLR highlights a strong asymmetric behaviour on perplexity pairs involving the set XIII. Indeed, while training a language model on a heterogeneous corpus, as it is XIII, makes the LM well performing when testing on simpler varieties, a language model trained on a poorer corpus underperforms when evaluating it on a richer corpus, as XIII.

	XIII	XIV	XV-XVI-1	XV-XVI-2
XIII	1.00	0.81	0.65	0.72
XIV	**1.23**	1.00	0.86	0.95
XV-XVI-1	**1.53**	1.16	1.00	1.14
XV-XVI-2	**1.39**	1.05	0.88	1.00

Table 5: PLR among pairs of diachronic language varieties.

4 Conclusions

In this paper we described *Vulgaris*, a project that analyzes a collection of literary texts covering the production of Italian authors mainly from the middle age. The dataset contains both poetry and prose, and each document is enriched by metadata that provide both information on the text characteristics and structure (the verse and stanza organization for poems and the paragraph splitting for prose). A preliminary analysis on the dataset by means of both simple statistics and perplexity–based measures gives some insights on the main feature of the collection that reflects the complexity and diachronic properties of Italian language in the early stages of its birth.

References

Ehsaneddin Asgari and Mohammad RK Mofrad. 2016. Comparing fifty natural languages and twelve genetic languages using word embedding language divergence (weld) as a quantitative measure of language distance. *arXiv preprint arXiv:1604.08561*.

Pierpaolo Basile, Annalina Caputo, Roberta Luisi, and Giovanni Semeraro. 2016. Diachronic analysis of the italian language exploiting google ngram. *CLiC it*, page 56.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Lars Borin. 2013. The why and how of measuring linguistic differences. *Approaches to measuring linguistic differences, Berlin, Mouton de Gruyter*, pages 3–25.

José Ramom Pichel Campos, Pablo Gamallo, and Iñaki Alegria. 2018. Measuring language distance among historical varieties using perplexity. application to european portuguese. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 145–155.

José Ramom Pichel Campos, Pablo Gamallo Otero, and Iñaki Alegria Loinaz. 2020. Measuring diachronic language distance using perplexity: Application to english, portuguese, and spanish. *Natural Language Engineering*, 26(4):433–454.

Alina Maria Ciobanu and Liviu P Dinu. 2020. Automatic identification and production of related words for historical linguistics. *Computational Linguistics*, 45(4):667–704.

Pablo Gamallo, José Ramom Pichel Campos, and Inaki Alegria. 2017. A perplexity-based method for similar languages discrimination. In *Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects (VarDial)*, pages 109–114.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kyuyeon Hwang and Wonyong Sung. 2017. Character-level language modeling with hierarchical recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 5720–5724. IEEE.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *AAAI*, pages 2741–2749.

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. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1948–1958.

HaiTao Liu and Jin Cong. 2013. Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin*, 58(10):1139–1144.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Giuseppe Marra, Andrea Zugarini, Stefano Melacci, and Marco Maggini. 2018. An unsupervised character-aware neural approach to word and context representation learning. In *International Conference on Artificial Neural Networks*, pages 126–136. Springer.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.

Marcos Zampieri and Martin Becker. 2013. Colonia: Corpus of historical portuguese. *ZSM Studien, Special Volume on Non-Standard Data Sources in Corpus-Based Research*, 5:69–76.

Marcos Zampieri and Preslav Nakov. 2020. *Similar Languages, Varieties, and Dialects: A Computational Perspective*. Cambridge University Press.

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.

Andrea Zugarini, Stefano Melacci, and Marco Maggini. 2019. Neural poetry: Learning to generate poems using syllables. In *International Conference on Artificial Neural Networks*, pages 313–325. Springer.