EPiC Series in Language and Linguistics
Volume 4, 2019, Pages 70–83
Proceedings of Third Workshop "Computational linguistics and language science"

Genre classification problem: in pursuit of systematics on a big webcorpus

Tatiana Shavrina[1]
[1]NRU Higher School of Economics, Moscow
rybolos@gmail.com

Abstract

This article is devoted to the problem of defining a genre in computer linguistics and searching for parameters that could formalize the concept of a genre. All kinds of existing typologies of genres rely on different types of features, whereas in the practice of NLP, any modern applications are adapted to learning on big data, and therefore - on text features that do not require additional non-automatic markup. Based on such text-internal features, in this article we focus on differentiation of various genres and their grouping on the basis of a similar distribution of features. The description of the contribution of various types of features to the final result and their interpretation are given, and also an analysis of how such features can be used to further adaptation of NLP models is provided. The materials of the "Taiga" corpus with genre annotation are used as experimental data.
Keywords: genre classification, text classification, web corpus, machine learning

1. Looking for systematics in Borges classification

" ... animals are divided into:
(a) those that belong to the Emperor, (b) embalmed ones, (c) those that are trained, (d) suckling pigs, (e) mermaids, (f) fabulous ones, (g) stray dogs, (h) those that are included in this classification, (i) those that tremble as if they were mad, (j) innumerable ones, (k) those drawn with a very fine camel brush, (l) others, (m) those that have just broken a flower vase, (n) those that resemble flies from a distance."
Borges 1999

Text genre is one of the oldest objects of philological and linguistic studies, and one of the worst

G. Wohlgenannt, R. von Waldenfels, S. Toldova, E. Rakhilina, D. Paperno, O. Lyashevskaya, N. Loukachevitch, S.O. Kuznetsov, O. Kultepina, D. Ilvovsky, B. Galitsky, E. Artemova and E. Bolshakova (eds.), CLLS 2018 (EPiC Series in Language and Linguistics, vol. 4), pp. 70–83

formalized as well. Nevertheless, no text in any set can be considered free of genre categorization - therefore it can be regarded as a hidden factor, affecting training and evaluation of NLP applications and biasing the sets. In this paper, we consider genre being one of the factors, changing the distribution of language units on different levels - morphological, syntactic, semantic and so forth. For example, four popular morphological parsers for Russian show different behaviour depending on the genre composition of corpora on which they were trained on: in Table 1 the results of the annotation of the phrase “scholarly husband” are shown - 3 taggers trained on Russian National Corpus tag it as a noun phrase with an adjective or participle, and the forth one, trained on the search queries, tag it as two nouns:

Parser	**Morphological annotation**	**Comment**
MarMoT (Muller e al., 2013)	ucheniy (scholarly) /A,nom,plen,sg,m muzh (husband) /S,nom,sg,m,anim	“scholarly” is an adjective
TreeTagger (Schmid, 1994)	ucheniy/V,nom,partcp,plen,sg,m,ipf,tran,pass muzh/S,nom,sg,m,anim	“scholarly” is a participle
Hunpos (Halacsy et al., 2007)	ucheniy/V,ipf,tran=partcp,pass,praet,m,sg,nom,plen/ muzh/S,m,anim=sg,nom/	“scholarly” is a participle
Mystem (Segalovich, 2003)	ucheniy {ucheniy=S,m,anim=nom,sg} muzh {muzh=S,m,anim=nom,sg}	“scholarly” is a noun

Table 1.

Genre, being, for the most part, the subject of philological, psycholinguistic, typological, pragmatic, cognitive and discursive research, each of the disciplines having their own definitions and features, is little formalized in terms of mathematics and computer linguistics, and is initially defined as "a historically emerging group of literary works united by a combination of formal and content properties". Most of the applied approaches are limited to borrowing a definition from philology or use a simplified definition, using the surrogate of "genres" instead (especially this approach is practiced in case of studying the Internet genres on the materials of web corpora) - prototypical blogs, news, documentation, technical literature, etc., when from the point of view of philology in each of such text objects there can be caught an essay, a lampoon, a poem, an interview or a review.

Alternatively, the following strategy is used: the term "genre" is avoided, and derived analogs are used, for example, "functional text dimensions" (Sharoff, 2018), "registers", "text types" (Biber and Conrad, 2009). Meanwhile, the hypothesis that inspired the present study is that the formalization of the concept of the genre can substantially clarify the nature of the influence of this factor on the different levels of linguistic phenomena, in particular, the distributions of words, word combinations, parts of speech and syntactic relations.

Existing conflict of genre classifications, on the one hand, increases the number of attempts to draw theoretical boundaries between genres, but on the other hand, makes it very difficult to verify hypotheses on corpus data - as (Sharoff, 2018) mentions, there are works like (Adamzik, 1995) with

more than 4000 different genres, including suicide notes, and 2000 genres in (Görlach, 2004) - these lists quickly become obsolete, do not include new forms, such as sms, tweets and so on. There are classifications based on “primitive mental acts”, on the type of attitude between the author and the addressee, on texts being more or less informative and structures (Sharoff, 2018) - but all in all the classifications are based on two approaches to genre features - text-internal and text-external information. Text-external features require manual annotation and are costly to obtain - the train sets in such studies are rather small comparing to the national corpora: but corpus train sets on the contrary traditionally contain less genres with some very general genre information - in (Kessler, Schutze, 1997) for English there are 6, there are 7 in (Lee, Myaeng, 2002) for Korean, 10 in (Stamatatos et al., 2000) for Greek, etc. - but at least, this way we can have some more evidence on genres and groups of genres - on the English material it was shown (Biber, 1988) that the frequency of using passive voice and the distribution of pronouns depend on the genre of the text.

But if based on text-internal features only, we can enlarge our train set very effectively and control the number of features we extract - and these are the features influence the behavior of language models used in real NLP-tools. The aim of this paper is to provide a more detailed research on genre classification on a bigger set of texts with a larger genre set, regarding only text-internal features frequently used in NLP applications and analyzing how they are distributed depending on the genre or a group of genres.

2 Applicational methodology

During our research, we have considered different groups of features, frequently used in NLP-modeling, some used in previous genre classification experiments, and some being introduced for the first time:

2.1 Lexical features

- lemmata of words, unigrams
- word forms, unigrams
- n-grammes of lemmas - from 1 to 5 grams
- n-gram word forms - from 1 to 5 grams

2.2 Morphological features

- pos-tag n-grams - from 1 to 5 grams
- full morphotag n-grams - from 1 to 5 grams

2.3 Syntactic features

- frequency of syntactic connections - from 1 to 5 grams
- properties of syntactical graphs - for the graph of syntactic dependencies itself and for the graph of successive links: the number of vertices, the number of edges, the radius of the graph, the diameter, the mean clustering coefficient, transitivity, density, pearson correlation coefficient - means of all the sentences for each text.

2.4 Readability features

- The complexity of the text is generally accepted readability metrics: Flesch-Kindcaid

Reading Ease, Flesch-Kincaid Grade Level, Gunning-Fog Score, Coleman-Liau Index, SMOG Index, Automated Readability Index (counted using library[1])

2.5 Text length and word length features

- average length of text in tokens, average word length in characters, number of paragraphs, number of sentences, presence of hashtag, smilies, lexical diversity - the ratio of unique word forms to the total number of word forms in the text, the number of capital letters

2.6 Symbolic features

- symbolic n-grams - from 1 to 5 grams

The author of this work believes that from the point of view of the applied approach, it is possible to choose a slightly different approach to the selection of features: not to reduce the potential set of labels using new terminology, avoiding the use of the word "genre", not to build yet another hierarchy on a small the number of artificial interpreted features, but to refer exclusively to the internal features of the text that do not require additional annotation. The lexico-grammatical features of the text and their distribution, on the one hand, make research completely based on a specific language, on the other hand, they make it possible to investigate the real grouping of texts according to their forms and means that have developed in the language. It seems unessential that such characteristics can always be easily interpretable in classification problems, their importance is somehow difficult to explain, however, they work.

The selection of attributes in the work is caused by the following motivation: the signs that are somehow involved in various applied natural language processing models-various n-grammes, lexical dictionaries and character sequences, or those of the signs that can be extracted from the text and theoretically included in a preliminary assessment of how to configure the model's hyperparameters.

All the experiments in this work were carried out on the materials of the open webcorpora "Taiga" (Shavrina, 2018) with morphological and syntactic annotation . The corpus includes 5 billion words of Russian from different genre segments of Internet texts: contemporary fiction, poetry, social media, news, subtitles for films and "the rest" - a collection of thematic magazines.
The corpus is marked with the help of its own unicode script[2], the tokenizer of the sentences from the nltk[3] library, morphological processing and parsing by UDpipe parser (Straka et al., 2017). Tagset of the annotation is universal dependencies 2.0[4]. Separately for prose and for poetry, a randomized sample of 1000 texts was compiled for each genre - 39 genres for prose and 39 genres for poetry - experiments are carried out on two separate sets (for full lists of genres see appendix 1 and 2). On the total volume of the corpus, this collection excludes the influence of the stylistics of individual prolific authors.

All the results of the experiments are obtained on the SVM classifier with weighted classes for uniformity and comparability with figures in the work on genre and stylistic classification on signs of rhetorical structures (Galitsky et al., 2016).

1 https://github.com/mmautner/readability
2 https://github.com/TatianaShavrina/taiga/blob/master/tagging_pipeline/unify.py
3 http://www.nltk.org/api/nltk.tokenize.html
4 http://universaldependencies.org/

3. Primarily results

3.1 Symbolic features

Symbolic n-grams from 1 to 5 show a different quality for prose and poetry with the f-measure of 40% in prose and 60% in poetry. The following genres are distinguished for prosaic texts with the best quality, which differs in accuracy or completeness from the general value: detectives (accuracy 0.76), children's creativity and stories about children (accuracy 100%), literary criticism (recall 0.72), mysticism (accuracy 0.82) and musical and film reviews (accuracy 0.86).

Worst of all, they were able to stand out on such signs as bikes (accuracy 0.17), horrors (completeness 0.03), cyberpunk, critical articles and satirical articles (zero quality).

In the case of poetic texts, such genres as poetic translations (accuracy 0.90) and verses in other languages (recall 0.97), as well as libretto (accuracy 0.94) and sonnets, canons, rondos (accuracy 0.86), and acrostics (accuracy 0.89).

Special types of lyrics could be defined - philosophical lyrics (accuracy 0.27), as well as civic and urban lyrics (0.31 for both genres) - it was also noted that different types of lyric poetry mix strongly with each other during the classification.

3.2 Lexical features

One of the strongest attributes for classification is for prose quality on lemmas in the region of 48% f-measures, and for poetry - 50% (unigrams and 1-5-gram word forms, unigrams and 1-5 grammes of lemmas) - on lexical grounds it is possible to distinguish, on the one hand, the most homogeneous types of genres (for example, anecdotes, film reviews, fairy tales for prose (recall 0.73, 0.70, 0.75 respectively)), as well as genres with the most different from other texts lexical composition - for example, children's creativity, cyberpunk, translations and literature in other languages for prose, sonnets, canzones, rondos, fables for poetry.

3.3 Morphological features

Best features - full morphotags and tags of parts of speech, their n-grams from 1 to 5 - for prose give quality in the area of 37% f-measure, and for poetry - 36%

Poetry: the most well-defined genres on such signs are verses with the influence of other languages: translations of songs, verses in other languages and poetic translations (recall of 0.62, 0.64 and 0.79 respectively). Also foreign forms of literature are well distinguished: the West: sonnets, canzones, rondos, essays and articles. In the genres of the east: rubai, hokku, the tank with a median recall (0.40), but a sufficiently high accuracy - 0.60 with a median in the 0.30 region.

In general, more conversational and format-free genres show the quality of the definition below the median - for example, in "chanson" the completeness of the definition is 0.23 and the accuracy is 0.18, and in civil lyric poetry, children's poems, love lyrics and philosophical lyrics, the accuracy and recall fluctuates around 0.12 -0.15.

Prose: the best are thematically separate genres - anecdotes, stories about children and children's prose (recall 0.56, 0.49, precision 0.46,), as well as cyberpunk and erotic prose (accuracy 0.74 and 0.44 respectively). The stories, novels and miniatures are worst of all, mixing with each other.

3.4 Syntactic features

Best features - n-gram of syntax marks from 1 to 5 - for prose give quality in the region of 28% f-measure, for poetry - 40% of f-measure.

Poetry: vers libre and white verse are predictably the best ones (accuracy 51% and 57% respectively), as well as sonnets, canzones, rondos, author's songs, translations of songs and poetic translations (non-standard word order - accuracy 60%, 58% , 56% and 69%). Prose: Nevertheless, such genres as religious texts stand out well (such as non-standard word order gives 46% accuracy), cyberpunk (non-

standard word sequences - 66% accuracy), literary translations (non-standard word order - 61% accuracy) .

Also stated: features of the averaged sentence syntax in the text - on average gives only 8% f-measures, and define children's creativity, mysticism, essays and articles.

3.5 Readability

All 6 metrics give for prose quality in the region of 11% f-measure, and for poetry - 9%

With admittance of the low quality of the general classification, the best quality was given to the genres "child creativity" and "prose in other languages" on prose data, which really implies some polarity of the complexity of the perception of the text. For verse data, translations and verses in other languages, as well as genre such as aphorisms (prototypically short ones) and essays and articles (prototypically long ones) were also singled out.

3.6 Text length

All the features for prose give 12% f-measure, and 11% for poetry.

The quality of classification on such signs is close to zero in most genres, but several genres have acrostics (recall 32%), aphorisms (74% recall), essays and articles and plays, miniatures - really stand out on such grounds. Surprising was the selection of the genres "east: rubai, hokku, tank" and "west: sonnets, canzones, rondos" (recall of 24% for both), distinguishable from such genres, more characteristic of the Russian tradition.

3.7 Results with no annotation

The features of syntax graphs surprisingly produced a good result of classification and clustering of the webcorpora texts without genre tags - texts from 4 sources of the Taiga corpus - news, twitter, literary magazines and popular science magazine NPlus1 - were taken of 10 000 texts each. In 4 classes, the classification accuracy was 88% of the f-measure, with a large mixture of news classes and popular science texts. The results of clustering on these characteristics can be seen in Figure 1.

Precision: 0.88 Recall: 0.88 F1 Measure: 0.87 Accuracy: 0.88

	precision	recall	f1-measure	examples
Fiction	0.98	0.98	0.98	3302
News	0.76	0.90	0.82	3332
NPlus1	0.67	0.41	0.51	1619
Twitter	1.00	1.00	1.00	3297

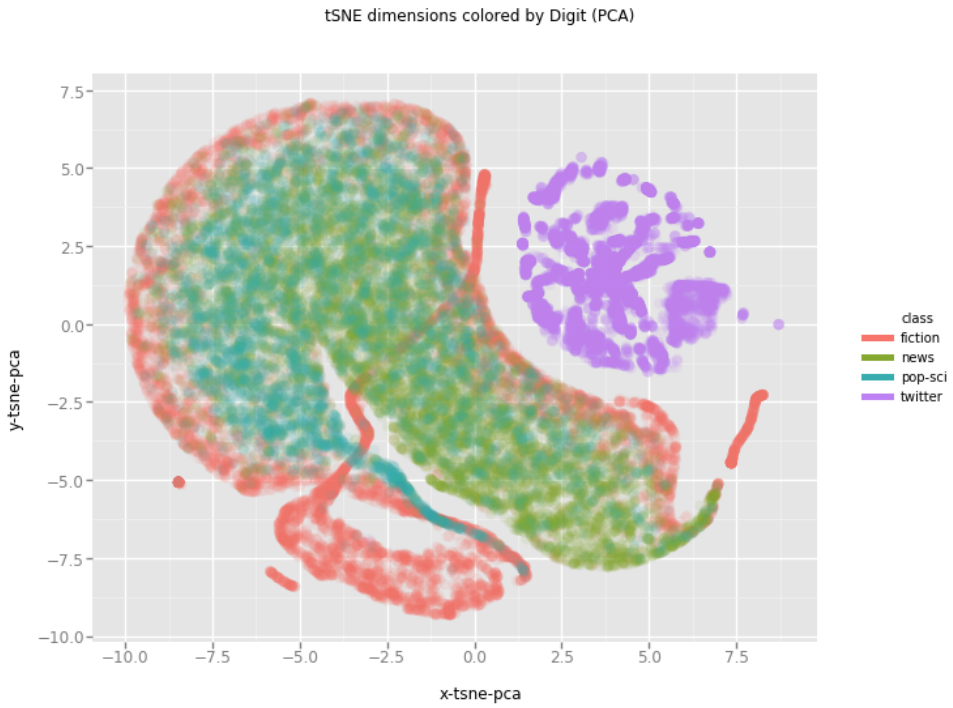

Figure 1 - Clustering Internet sources on the features of syntax graphs using the nearest-neighbor method, bringing it into a two-dimensional form using t-SNE.

4. Discussion (moderate one)

According to the results of experiments with classification, it can be concluded that not all genres are well separated on the basis of the attributes used in context models for the purposes of automatic text analysis.

Separate genres from the sample, both poetic and prosaic, stand out worse than others: for poetic genres, it is primarily poetry without a rubric, combining texts of very different forms and themes, as well as various types of lyrics that are not easily distinguishable; for prose - stories, novels and tales that do not seem to have the characteristic distinguishing features, as well as the class of genres combined into a group of "large forms" - cycles of poetry, since the poem from the cycle is difficult to separate from independent ones, as well as poems and plays that mix with each other. At the same time, the children's sections in prose and poetry were well separated, as well as song and music genres, literary translations, complex analytical texts (essays and articles, literary criticism).
In general, it is possible to single out such specific genres that are well distinguished on certain grounds:

- symbolic features: thematic genres - detectives, mysticism, literary criticism, children's creativity and stories about children, literary criticism, music and film reviews, as well as lyrical genres, various translations, texts in other languages;
- lexical features: give a sufficiently high quality of classification, while it is rather low in macroclasses and large samples - anecdotes, film reviews, fairy tales and genres that have the most lexical composition - children's creativity, cyberpunk, translations and

literature in other languages prose, sonnet, canzone, rondo, fables;

- morphological features: translations of songs, verses in other languages and poetic translations, foreign forms of literature - western: sonnets, canons, rondos, essays and articles, east: rubai, hokku, tanka, anecdotes, children's literature, as well as cyberpunk and erotic prose;
- syntactic features: religious texts, cyberpunk, literary translations, vers libre and white verse, sonnets, canzones, rondo, author's song, translations of songs and musical genres;
- features of the length of the text: acrostics, aphorisms, essays and articles and plays, miniatures, "large forms";
- readability: children's creativity, literature for children, essays and articles, prose in other languages.

One possible solution may be to combine the most similar to each other on different characters of genres, and adjust the models to a small group, and not under all sorts of classifications of varying degrees of fragmentation from the N genres.

The task becomes reducible to the task of clustering - for applied needs it is possible to make corrections for a cluster of similar genres. We introduce the definition of such a group of genres:

The genre cluster is a group of texts of various genres, united on the basis of similar, generalizing these genres, internal signs of the text.

As a result of a number of experiments on the characteristics described above, it was found that on the samples of prosaic and poetic texts, they are best grouped by genre into 4-5 genre clusters.
So, on lexical signs on poetic texts it is possible to reveal clusters, in which they fall most often:

1. musical creativity, parodies and humor, translations and verses in other languages, solid forms
2. children's sections, parodies and humor
3. lyric, solid forms, large forms
4. free forms and prose, lyrics, parodies and humor

We can use this information directly when teaching models, for example, morphological and syntactic analysis, and the work of transfer learning. Examples of such a setting are given in the table 2 - this analysis shows the degree of flexibility with which those or other algorithms can be applied on different genres, and also potentially lead to an improvement in the quality of processing texts of different genre clusters.

Feature	Feature importance	applications	Comment
symbol n-grams	When classifying shows the best quality on a sample of genres, and also on a sample with groups of genres.	used for morphological analysis on neural network parsers analyzing a character chain. In the usual case of this type of model, the length of the n-gram is a hyperparameter, and therefore it is better to adjust it separately to adapt to different groups of genres.	The length of an n-gram is specified in the model parameter, and can reach up to a hundred, whereas in the experiments above, n-grams of length no more than 5 are considered. With increasing the length of the n-gram sequence, it is logical to assume an increase in the spread of genre differences due to more unique chains.
word n-grams	When classifying shows the second quality after the symbolic n-gram on a sample of genres, as well as on a sample with groups of genres, while being the most easily interpreted and widely used characteristics	Used in all the main methods of morphological markup, (CRF, HMM, recurrent neural networks) as they are the most elementary unit of the language sequence.	Affect the results of all models, as 1) form a dictionary of possible word forms and lemmas. 2) for homonymous words, genre-differentiated n-grammes and word vectors make it possible to better remove ambiguity.
morphotag n-grams	Show a good quality of classification on a large sample of genres and can affect both the result of the removal of homonymy by a morphological parser and the construction of a tree by a syntactic parser. The result of the work of the dictionary systems will not be affected.	CRF, Viterbi algorithm, HMM, all syntactic parsers using the result of morphological analysis	Formulate the statistics used when sampling the local and global maximum when removing homonymy in morphology.

syntactic features	Show an unsatisfactory quality on the classification of the total sample of genres, but the larger groups of genres that are similar in syntactic properties distinguish them very well.	All used properties of the graph are not used directly, however some of them, as well as the graph structure itself, are used in the algorithms of parsing.	Can be used as a hyperparameter for selection
text readability	Show an unsatisfactory quality on the classification on the total volume of the sample genres, but the most contrasting genres, children's literature, essays and philosophical texts from all the others, share well.	Are not used directly in the technologies of morphological and syntactic analysis, however information about the number of syllables is used in parts of morphological parsers guessing unknown words.	
text length features	Show unsatisfactory quality on the classification of the total volume of the genres of the sample, but well distinguish the most different genres - with short and long texts	Can influence the models in which the global maximum is selected on a string of words, and the length of the string has a value - for example, in CRF.	

Table 2

Further work will be aimed at adapting models of morphological and syntactic analysis to different genre samples - the problem of automatic and syntactic analysis is not solved, and the genre factor and parameters correlating with it, such as morphological n-gram frequencies, lexical composition, sentence and word lengths, are important features that can be relied upon for 1) differentiating the training sample 2) adapting the parameter hyperparameters 3) adding new genre-differentiating information to the model - depending on the type of algorithm used.
While this work was written, several works were published using the third method: for example, in work (Sagot and Alonso, 2017) for the improvement of morphological parsing by means of a bidirectional recurrent neural network on symbolic sequences, information was added from the dictionary with embedded words , and in work (Sorokin, 2018) information from language models was used to improve the performance of a unidirectional recurrent neural network in the problem of

morphological analysis.

5. Discussion (strong one)

On all the features that we use when modeling a language - words, symbols, word chains, word structures - genres are not generally distinguished with sufficient quality, or separate groups of genres from the total mass are singled out. This is shown in a large and randomized sample, devoid of the author's stylistic features. It may be advisable to introduce a new definition for such a phenomenon, abandoning the term "genre" in the applied approach, and in applied works to use a more formalized concept, for example, a genre cluster. We define the genre cluster as a group of texts of different genres, united on the basis of similar, generalizing these genres, text-internal features.

6. Conclusion

In the present work, a study of the influence of text attributes on the genre classification on the material of a large corpus is made.

As part of the classification experiments, it was shown that the greatest contribution to the classification of genres and groups of genres of prose and verse text is given by the following features: n-grams of word forms, n-grams of morphological tags and tags of syntactic relations.

Dedicated features are able to interpretably separate clusters of web corpus texts with no additional annotation - texts of blogs, news, popular science journals, prose. In this case, the texts themselves with genre tags are best clustered into the following groups:

- by lexical features - parodies and humor and children's sections, solid and large forms and cycles of poems, a variety of lyrics, literary translations;
- by morphological features - translations, foreign forms of literature - west: sonnets, canons, rondo, east: rubai, hokku, tanka, as well as essays and articles, genres of "large forms", anecdotes, children's literature;
- by syntactic features - "music and song " genres, including translations, parodies and humor and children's sections, a variety of lyrics.

These results make it possible to form a more precise definition of the genre from an applied point of view: the author introduces the term "genre cluster" - this is a group of texts that show similar behavior based on the highlighted text-internal attributes. For example, on the data obtained from the web corpus, "ordinary" news and news collected from a popular scientific resource can be referred to one genre cluster based on syntactic features. As shown by the classification experiments in this paper, it is not possible to identify most of the individual genres with the help of exclusively extracted from the text - it works only with certain genres that are most distinct in form and content. In this direction, research can be continued, including with the help of the corpus used in this work (Shavrina, 2018).

As a by-product of the work, frequency dictionaries for various genres have been obtained, showing a significant difference in lexical frequencies in these groups of texts, now available at https://github.com/TatianaShavrina/on_genres.

7. References

Adamzik, Kirsten (1995) Textsorten - Texttypologie: eine kommentierte Bibliographie. Münster : Nodus-Publ., 1995. - 301 S.

Benoıt Sagot and Hector Martınez Alonso (2017) Improving neural tagging with lexical information. Proceedings of the 15th International Conference on Parsing Technologies, pages 25–31,Pisa, Italy; September 20–22, 2017.c©2017 Association for Computational Linguistics

Biber, D. (1988). Variations Across Speech and Writing. Cambridge University Press.

Biber, D. and Conrad, S. (2009). Register, genre, and style. Cambridge University Press.

Borges, Jorge Luis (1999), "John Wilkins' Analytical Language", in Weinberger, Eliot, Selected nonfictions, Eliot Weinberger, transl., Penguin Books, p. 231, ISBN 0-14-029011-7. The essay was originally published as "El idioma analítico de John Wilkins", La Nación (in Spanish), Argentina, 8 February 1942, and republished in Otras inquisiciones

Galitsky B. A., Ilvovsky D. A., Chernyak E. L., Kuznetsov S. O. (2016) Style and Genre Classification by Means of Deep Textual Parsing. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2016”. Moscow, June 1–4, 2016

Görlach, M. (2004). Text types and the history of English . Walter de Gruyter.

Peter Halacsy, Andras Kornai, and Csaba Oravecz. 2007.Poster paper: Hunpos – an open source trigram tagger. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 209–212, Prague, Czech Republic, June. Association for Computational Linguistics.

Kessler, Brett, Geoffrey Nunberg, and Hinrich Schutze. 1997. Automatic detection of text genre. In Proceedings of 35th Annual Meeting, pages 32-38. Association for Computational Linguistics

Lee Y.-B., Myaeng S.H. (2002) Text genre classification with genre-revealing and subject-revealing features. In Proc. of SIGIR, pages 145–150, 2002.

Thomas Müller, Helmut Schmid and Hinrich Schütze (2013) Efficient Higher-Order CRFs for Morphological Tagging. EMNLP

Schmid, G. (1994). TreeTagger—A language-independent part-of-speech tagger . Available at http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html

Segalovich I. (2003), A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications, Las Vegas, Nevada, USA.

Sharoff, S. (2018). Functional Text Dimensions for annotation of Web corpora. Corpora,31:2.

Shavrina T. (2018) Differential Approach to Webcorpus Construction. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018”. Moscow, June 1–4, 2018

Shavrina T., Shapovalova O. (2017) TO THE METHODOLOGY OF CORPUS CONSTRUCTION FOR MACHINE LEARNING: «TAIGA» SYNTAX TREE CORPUS AND PARSER. in proc. of “CORPORA2017”, international conference , Saint-Petersbourg, 2017.

Sorokin A. (2018) Improving neural morphological tagging using language models . Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018”. Moscow, 2018

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. Computational Linguistics, 26(4), 471–495.

Milan Straka and Jana Straková. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, August 2017.

Acknowledgements

The author expresses sincere gratitude to Olga Lyashevskaya and Sergey Sharoff for fruitful comments and fundamental suggestions. Special thanks are also expressed to students of NRU HSE who acted as annotators of the corpus - due to their work a further differentiated genre-specific evaluation of morphological models will become possible.

Appendix

Appendix 1 - Genre of prose

{"miniatures": "Small forms", \
"short stories": "Small forms", \
"stories": "Small forms", \
"reports": "Small forms", \
"stories": "Large forms," \
"novels": "Large forms", \
"dramaturgy": "Genre works", \
"detectives": "Genre works", \
"adventure": "Genre works", \
"Fantasy": "Genre works", \
"Fantasy": "Genre works", \
"horrors": "Genre works", \
"cyberpunk": "Genre works", \
"erotic prose": "Genre works", \
"humorous prose": "Humor", \
"ironic prose": "Humor", \
"feuilletons": "Humor", \
"anecdotes": "Humor", \
"bikes": "Humor", \
"history and politics": "Essays and articles", \
"Literary criticism": "Essays and articles", \
"natural science": "Essays and articles," \
"journalism": "Essays and articles," \
"philosophy": "Essays and articles," \
"religion": "Essays and articles," \
"mysticism": "Essays and articles," \
"Memoirs": "Essays and articles," \
"critical articles": "Literary criticism," \
"literary reviews": "Literary criticism," \
"musical and film reviews": "Literary criticism," \
"literature for children": "Children's sections", \
"stories about children": "Children's sections", \
"fairy tales": "Children's sections", \
"children's creativity": "Children's sections", \
"Poems": "Poetry," \
"poems in prose": "Poetry", \
"literary translations": "Translations and prose in other languages," \

"Prose in Other Languages": "Translations and Prose in Other Languages"}

Appendix 2 — Poetry genres

{'love lyrics': 'lyrics',
'civil lyricism': 'lyric',
'landscape lyrics': 'lyric',
'urban lyrics': 'lyrics',
'religious lyrics': 'lyrics',
'philosophical lyrics': 'lyric',
'mysticism and esoterics': 'lyric',
'cycles of poetry': 'Large forms',
'poems': 'Large forms',
'plays': 'Large forms',
'west: sonnets, canzons, rondos': 'Solid forms',
'Uncategorized': 'Poems without a rubric',
'east: rubai, hokku, tank': 'Solid forms',
'acrostics': 'Solid forms', 'vers libre': 'Free forms and prose',
'white and free verse': 'Free Forms and Prose',
'poems in prose': 'Free forms and prose',
'prose miniatures': 'Free forms and prose',
'essays and articles': 'Free Forms and Prose',
'aphorisms': 'Free Forms and Prose',
'parodies': 'Parodies and humor',
'imitation': 'Parodies and humor',
'comic verses': 'Parodies and humor',
'ironic verses': 'Parodies and humor',
'satirical verses': 'Parodies and humor',
'fables': 'Parodies and humor', 'poems for children': 'Children's sections',
'children's creativity': 'Children's sections',
'author's song': 'Musical creativity',
'pop song': 'Musical creativity',
'Russian rock': 'Musical creativity',
'libretto': 'Musical creativity',
'chanson': 'Musical creativity',
'translations of songs': 'Musical creativity',
'rework of songs': 'Musical creativity',
'poetic translations': 'Translations and verses in other languages',
'verses in other languages': 'Translations and verses in other languages',
'translations of songs': 'Translations and verses in other languages',
'Uncategorized': 'Verses Uncategorized'}