Exploration of register-dependent lexical semantics using word embeddings

Andrey Kutuzov
University of Oslo
Norway
andreku@ifi.uio.no

Elizaveta Kuzmenko
National Research University
Higher School of Economics
lizaku77@gmail.com

Anna Marakasova
National Research University
Higher School of Economics
anya.tiva@gmail.com

Abstract

We present an approach to detect differences in lexical semantics across English language registers, using word embedding models from distributional semantics paradigm. Models trained on register-specific subcorpora of the BNC corpus are employed to compare lists of nearest associates for particular words and draw conclusions about their semantic shifts depending on register in which they are used. The models are evaluated on the task of register classification with the help of the deep inverse regression approach.

Additionally, we present a demo web service featuring most of the described models and allowing to explore word meanings in different English registers and to detect register affiliation for arbitrary texts. The code for the service can be easily adapted to any set of underlying models.

1 Introduction

The phenomenon of language registers has long attracted the attention of linguists as well as specialists in other humanities areas. It is closely related to the issues of how humans produce and interpret texts. This paper and the accompanying demo web service intend to make exploration of these questions more accurate and data-driven.

We present an approach to track how words change their dominant meaning depending on the particular text register in which they are used. A typical example is the English word '*cup*' denoting mostly mugs in the fiction texts, but switching its primary meaning to the championship prize in the news texts. To find this (often rather slight) meaning shifts we use prediction-based distributional semantics models. In particular, we employ the *Continuous Bag of Words* model (Mikolov et al., 2013), trained on the register-separated subcorpora of the British National Corpus (BNC).

We evaluate this approach and show how it can be used to detect semantic shifts that cannot be discovered with traditional frequency-based corpus analysis methods. In addition, a similar algorithm is used to efficiently detect the most probable register of arbitrary text. We also present a convenient web service demonstrating the aforementioned techniques for English language material.

The paper is organized as follows. In Section 2 we discuss the related work, describe the notions of language registers and genres in more details, and put our work in the academic context. Section 3 presents our data and methods and provides examples of register-dependent semantic shifts. Section 4 is devoted to the algorithm of extracting register affiliation for texts based on the models we used in the previous section. It also evaluates the models. In Section 5 we conclude and draw main directions for the future work.

2 Related work

The notion of *language register* is often used as synonymous to *genre* or *style*. In any of these manifestations, it is a set of loosely defined categories of texts depending on both form and content (Chandler, 1997). It is very important that registers are not topics, which are defined on solely **internal** criteria (in

Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH),
pages 26–34, Osaka, Japan, December 11-17 2016.

other words, they describe what the text is about). Meanwhile, registers are defined on the basis of **external** criteria (namely, they depend on who produces the text, for whom and under what circumstances).

Any system that classifies texts into 'registers' or 'genres' is unavoidably subjective: the distinction between registers is more of a convention than of some objective properties of texts. Notwithstanding this fact, this separation can be a useful tool in humanities as it is inherently relative to typical communicative situations and human behaviour within them, as well as the issues of the influence of language on the society. On the other hand, one can try to extract some objective linguistic phenomena associated with this or that register (cf., for example, (Biber, 1995)). Our work described in this paper also aims to facilitate the process of linguistic description of differences between registers and (to a lesser extent) genres.

It is important to note that we do not try to redefine existing register or genre classifications. Instead, we take them as is (see the thorough discussion in Section 3). We presuppose that the register separation introduced during compiling large and well-established corpora, like the BNC, makes at least some sense and represents linguistic reality to at least some extent. However, the same experiments can be performed with any other register classification scheme.

Our main source of data is the British National Corpus. The foundations of text classification in the BNC are described in (Lee, 2001). It states, *inter alia*, that there is no such thing as the 'general English language'. Instead, there exist several 'Englishes' (sublanguages), depending on the communicative situations. They can be called *registers*, whereas *genres* in this approach are their subdivisions: for example, sport column is a genre within news register.

(Lee, 2001) emphasizes that linguistic specificity is associated with registers, not genres; this is confirmed in NLP research, see, for example, (Dell'Orletta et al., 2013). We use this hypothesis in our work, as we are interested in differences of word meanings, which are manifested on the register level. This is the reason for us to compare macro-registers in the BNC (like *news*, *academic* and *fiction*) instead of comparing particular genres (like *humanities academic* texts and *fiction drama* texts). Of course, another reason is that the BNC is not a very large corpus and subcorpora consisting of texts belonging only to particular genres would be too small to train reasonable distributional models.

The size of the subcorpora is important, as we share the idea expressed in the seminal work of (Kilgarriff, 1997): the sets of word senses are absolutely dependent on the corpora we use to study them (and the purposes of the study). It essentially means that word meanings are abstractions over clusters of word usages. The word usage in corpus linguistics is traditionally defined through the frequencies of the word co-occurrences with other words (typical contexts). This is where distributional models come into play.

In natural language processing, distributional models, based on the foundational idea of 'meaning as context', are now one of the primary tools for semantic-related tasks. They stem from the so called *distributional hypothesis*, which states that co-occurrence statistics (word co-occurrence distributions) extracted from large enough natural language corpus can in some way represent the 'meaning' of words as perceived by humans (Firth, 1957). More formally, with a given *training corpus*, each word is represented as a vector of frequencies for this word occurring together with other linguistic entities (its contexts). These vectors are located in a *semantic space* with all possible contexts or semantic features as dimensions. Vector models of distributional semantics or vector space models (VSMs) are well established in the field of computational linguistics and have been studied for decades; see the review in (Turney et al., 2010), among others.

Recently, a particular type of these models has become very popular, namely, the *prediction models* that utilize artificial neural networks, introduced in (Bengio et al., 2003) and (Mikolov et al., 2013) and employed in a wide-spread *word2vec* software. Prediction models directly learn dense vectors (*embeddings*) which maximize the similarity between contextual neighbours found in the data, while minimizing the similarity for unseen contexts.

Initial vectors are generated randomly (the target vector size is set at the beginning of the training process, typically hundreds of components) and then gradually converge to optimal values, as we move through the training corpus with a sliding window. To this end, predictive models employ machine learning and consider each training instance as a prediction problem: one wants to predict the current

word with the help of its contexts or vice versa. The outcome of the prediction for the given training example (context window) determines whether we change the current word vector (embedding) and in what direction.

To sum it up, prediction models learn meaningful vector representations of words in the training corpus by stochastically trying to predict a focus word from its context neighbours, and slightly adjusting vectors in case of mistake. These models are simple, fast and efficient in various NLP tasks; for evaluation, see (Baroni et al., 2014), among others.

Interestingly, this approach is absolutely compatible with (Lee, 2001) idea that registers exist independently of text-level structures. Thus, one can extract linguistic features of different registers without paying attention to particular texts, treating all of them as one large register-marked corpus. This is exactly what we do with the register-separated subcorpora of the BNC. The details of this are described in the next section.

There have been several studies dedicated to the task of automatic text register (or genre) identification; most of them were inspired by (Kessler et al., 1997). To the best of our knowledge, none of the studies made use of distributional semantics. The algorithms were mainly based on simple word and text statistics; see, for example, (Lee and Myaeng, 2002), (Amasyal and Diri, 2006). In some of the works ((Stamatatos et al., 2000), (zu Eissen and Stein, 2004)), apart from statistical features (word count, sentence count, character per word count, punctuation marks count), more complex linguistic features were employed: morpho-syntactic (passive count, nominalization count, relative clauses count and other frequencies of different syntactic categories) and lexical (type-token ratio).

More resent research, for example, (Biber and Egbert, 2015), also demonstrates the advantages of incorporating lexico-grammatical characteristics into text types prediction. We attempt to move it even further and use state-of-the art word embedding models in order to explore linguistic specificity of language registers.

3 Distributional approach to meaning across registers

(Lee, 2001) states that a text belonging to a particular register is '*the instantiation of a conventionalised, functional configuration of language tied to certain broad societal situations, that is, variety according to use*'. In a sense, we also follow the (Chandler, 1997) concept of 'ideal reader' for each register. Ideal reader shares semantic structures with the text producer. This is reflected in the meaning of the words used in producing register-specific texts. These meanings can be different from the so-called *core meaning* of the word. Note that the concept of 'core meaning' is disputable itself, but with a certain degree of confidence we can say that this is the meaning stored in the dictionaries *or* the one given by distributional models trained on a full balanced and representative corpus (for example, the whole BNC).

As stated above, distributional models compute meaning by analysing word co-occurrences. The trained model can represent semantics of a given word as a sequence of its 'nearest associates': words closest to the key word by the cosine similarity of their vectors (embeddings). Our work is based on the presupposition that in the models trained on register-specific subcorpora, the register-specific words would feature meaningfully different sets of nearest associates.

3.1 Data preparation

The data used for our experiments was compiled from the British National Corpus: it is freely accessible, comparatively large, well-balanced, and, therefore, supposed to be representative for the respective language. What is even more important is that the BNC features well-developed register and genre annotation. Since our aim is to discover semantic shifts associated with the switching from one register to another, we train several distributional semantic models on the texts of particular registers.

The BNC provides eight 'text type' categories: *academic writing, published fiction, published non-fiction, news and journalism, other published writing, unpublished writing, conversation, other spoken*. Although it might be an apparent way to split the BNC into subcorpora representing different registers, that is not something we were looking for. The share of each text type is highly disproportionate ('*news and journals*' contains 9.56% of all the corpus tokens, while '*published non-fiction*' – 24.58%; '*other*

published writing' appears to be quite a large category (18.26%) that contains text of different registers). This disproportion would, for certain, affect the quality of distributional models.

Therefore, we have developed our own split into language registers based on the BNC genre classification[1] (Lee, 2001). The classification provides us not only with token and sentence counts for each genre, but also with "macro-genres" categories. For instance, the following genres of academic writings – 'humanities arts', 'medicine', 'nature science', 'politics law education', 'social science', 'technical engineering' – are encoded as '*W ac:humanities arts', 'W ac:medicine', 'W ac:nat science', 'W ac:polit law edu', 'W ac:soc science', 'W ac:tech engin*' respectively. This encoding let us form an 'academic' register out of the texts of the mentioned genres. Thus, we have obtained the split which is more balanced in terms of token counts and which, we hope, is more reliable. The registers are listed below:

1. academic texts (15 632 085 tokens);
2. fiction texts (15 950 682 tokens);
3. newspaper articles (14 214 484 tokens);
4. non-fiction and non-academic texts (18 307 605 tokens);
5. spoken texts (17 451 494 tokens).

The '*academic texts*' register comprises research papers and articles from the various fields of study: social and political science, law, education, natural science, medicine, humanities, arts and computer science. The same topics are covered in '*non-academic*' texts, the difference is that their intended audience is non-professional. '*Fiction*' texts include mainly prose, but also a small amount of drama and poetry. As for the '*news*' register, it consists of the following text types: report (the prevalent one), commerce, social, sport, science, arts. The remaining texts are of various genres (religion, parliamentary, email, advertisement, administrative etc.) and fall into miscellaneous category. As they do not represent any particular register, we exclude them from consideration.

These subcorpora were pre-processed, replacing each token with its lemma and PoS tag ($loved \rightarrow love\_VERB$). Data on lemmas and PoS affiliation was extracted from the BNC mark-up. We also removed all functional words and one-word sentences.

3.2 Training distributional models

Then, CBOW models (Mikolov et al., 2013) were trained on each of the subcorpora and on the whole BNC corpus. We used the standard set of hyperparameters for training, excluding the selection of prediction material: hierarchical softmax was employed, instead of more widely used negative sampling, as this makes it easier to implement text classification via deep inverse regression (see Section 4 for more details). Another important decision to make was the size of sliding window (how many words to the right and to the left of the focus word to consider). It is known from previous work that larger windows tend to generate more 'associative' models ('*cup*' is semantically close to '*coffee*'), while narrower windows favour more 'functional' models ('*cup*' is semantically close to '*mug*'). It is not immediately clear what mode is better for our task. That's why we trained two sets of models, with window sizes 3 and 10. The evaluation process is described in Section 4.

The resulting models indeed demonstrate semantic specificity of different registers. It can be observed through comparing the lists of nearest associates for a given word in different models. The Table 1 gives one example of such specificity.

One can clearly see the difference in the meaning of the word '*bank*' when it is used in the fiction register. While the dominant meaning in general English (and in academic and news registers) is related to financial institutions, fiction texts use the word in an absolutely different sense related to river shores. It is possible to quantify these differences using any of set comparing methods. At the moment we implemented Jaccard distance (Jaccard, 1901) which estimates the number of intersecting elements in

[1] `http://www.natcorp.ox.ac.uk/docs/URG/codes.html#classcodes`

Table 1: First 5 associates for '*bank*'

Model	**Whole BNC**	**Academic**	**News**	**Fiction**
1	banker	mortgage	banker	**spate**
2	banking	wales	banking	**slope**
3	loan	overdraft	deposit	**gully**
4	deposit	money	lender	**shore**
5	overdraft	loan	branch	**hill**
Jaccard distance to the whole BNC	0	0.75	0.57	**1**
Kendall's τ distance to the whole BNC	0	0.66	0	**0.85**

Table 2: First 5 associates for '*star*'

Model	**Whole BNC**	**Academic**	**News**	**Fiction**
1	hollywood	sun	superstar	moon
2	superstar	earth	singer	galaxy
3	movie	jupiter	legend	light
4	galaxy	galaxy	heart-throb	cloud
5	entertainer	stripe	guitarist	sky
Jaccard distance to the whole BNC	0	0.9	0.9	0.9
Kendall's τ distance to the whole BNC	0	0.56	0.26	0.4

two sets and normalized Kendall's τ (Knight, 1966) which calculates the differences between rankings. However, we also plan to test other metrics, for example, as described in (Kutuzov and Kuzmenko, 2016). Larger distance between the 'general' model and a particular register means that the key word is semantically shifted in this register.

The word '*bank*' is an example of homonymy, where different senses are totally unrelated. However, our approach also captures subtler cases of polysemy, in which senses are still connected to each other, like with the noun '*star*' described in the Table 2.

In this case, all the register-specific models seem to be on approximately the same distance from the general model (which comprises several meanings at once). However, it is easy to see that the academic and fiction registers are closer to each other, featuring the astronomical sense of '*star*'. At the same time, they share no associates with news register, which primarily employs the word in the sense of '*celebrity*'. The Table 3 shows the matrix of mutual distances between different registers for the word '*star*'.

RegisterExplorer, our demo web service at `http://ltr.uio.no/embeddings/registers`, based on *Gensim* framework (Řehůřek and Sojka, 2010), provides easy access to such comparative tables for the models trained on the BNC register subcorpora. It also features visualizations of the interaction between nearest associates within one register and between registers as well. The models are available to download, and the source code for the service is released as free software, making it easy to adapt the system to any set of models, depending on a researcher' aims. Among other applications, one can use

Table 3: Mutual Kendall's τ distances between registers for the word '*star*'

Register	**Spoken**	**Academic**	**News**	**Fiction**	**Non-fiction**
Spoken	0	**0.69**	0.50	**0.69**	**0.69**
Academic	**0.69**	0	**0.69**	0.40	0.49
News	0.50	**0.69**	0	**0.69**	**0.69**
Fiction	**0.69**	0.40	**0.69**	0	0.51
Non-fiction	**0.69**	0.49	**0.69**	0.51	0

the system to explore particular *genres*, not only large-scale registers, as in this work. For this, one needs only a large genre-annotated corpus to train genre-specific models (the BNC size is not enough for that).

The distributional approach allows to reveal register-specific senses, which cannot be discovered by traditional frequency-based analysis. Frequency distributions across registers are meaningful only for words which possess one dominant meaning. For instance, let us consider a word '*room*' (noun). Based on the 'Words and Phrases' web resource[2], compiled using the Corpus of Contemporary American English (COCA), we can only conclude that this word is much more frequent in the fiction register (frequency of 87 748, while the respective counts for the spoken, magazine, newspaper and academic registers are 19 948, 36 225, 31 908, 10 113). However, no conclusions about meaning differences across registers can be made.

The distributional approach (implemented in *RegisterExplorer*) supports the observation that in the fiction register the word 'room' is more frequent. However, it additionally allows us to see that the spoken register – unlike other registers and general English – is strongly associated with the sense 'the amount of space' (apart from much more common sense 'a part of a building with a floor, walls, and a ceiling').

Senses and registers are not always in one-to-one association. A register may feature two or more specific senses, while two or more registers may be similar in terms of the senses they share. Considering the verb 'mean', we can clearly see that the senses 'to be of a specified degree of importance to (someone)' and 'to intend something, often bad or wrong' are associated with the fiction register, while the academic and news registers share the sense 'to have as a consequence or result'.

4 Detecting registers with word embeddings

Any word embedding model can be turned into a text classifier via Bayes rule. The idea (dubbed 'deep inverse regression') was first introduced in (Taddy, 2015) and essentially allows to calculate the likelihood of an arbitrary sequence of words in the context of any trained distributional model.

Recall that we have a set of models trained on register-specific texts. It means that we can find out the extent to which a particular sentence or a text is prototypical for a given register or registers.

For, example intuitively it is obvious that the sequence '*star divorced yesterday*' is quite likely in the news texts, very unlikely in the academic texts, and in the spoken texts its likelihood should be somewhere in between. If we apply deep inverse regression to our models with window size 10 and the aforementioned sequence (lemmatized and PoS-tagged), it produces the following likelihood values (more negative values mean less likelihood):

- News: -27.53
- Spoken: -42.39
- Academic: -43.25
- Fiction: -48.42

The news register indeed turns out the most likely to produce such a sentence, with spoken register a bit less likely, and academic and fiction most unlikely.

RegisterExplorer allows users to analyse an arbitrary text and receive lists of registers ranked by their likelihood to produce such a text. This can be used to quickly explore 'linguistic profiles' of texts and to find out how different registers are manifested in language, helping to study their interactions.

To evaluate the applicability of this approach to unseen data we randomly sampled 1 000 sentences from each of our register-specific subcorpora (for the purposes of evaluation, these sentences were not used in the subsequent model training). For these sentences, we calculated the most likely register according to the deep inverse regression method and evaluated it against the real sentence affiliations. The results are summarized in the Table 4.

[2] `http://www.wordandphrase.info/`

Table 4: Models' performance scores

	Text classification, F1		Simlex999, Spearman correlation	
Register	Window size 3	Window size 10	Window size 3	Window size 10
Academic	0.52	0.50	0.05	0.05
Fiction	0.48	0.46	0.17	0.15
News	0.49	0.50	0.03	0.04
NonAcademic	0.39	0.39	0.03	0.11
Spoken	0.48	0.48	0.03	0.04
Average	**0.47**	0.46	0.06	**0.08**

The models with the smaller window size performed a bit better in text prediction, which supposedly implies that they are better in capturing linguistic specificity of language registers in the BNC. We hypothesize that the reason for this is that the smaller windows work as filters against too much influence of topics and content of the sentences. As a result, models are more 'concentrated' on lexical semantics *per se*. However we acknowledge that the difference is negligible and more experiments are needed to find out the best models' hyperparameters for our task.

For reference, we additionally evaluated the resulting subcorpora models against the well-known *SimLex999* semantic similarity dataset (Hill et al., 2016), as a measure of general 'quality' of the models. Relatively low values did not come as a surprise, considering rather small size of the subcorpora (as compared to billion-word corpora used in contemporary state-of-the-art models). There is no consistent advantage of one window size over another (unlike with text classification), the results vary depending on a register.

One can notice unusual behaviour of the *nonAcademic* model: with the increase of the window size, it becomes much better in *Simlex999* task, unlike models trained on other subcorpora. In our opinion, the reason for this is comparatively high average sentence length in the texts belonging this register (about 12 words), favouring settings with large window sizes. Another register with long sentences is *academic* (12.4 words), but it does not demonstrate such a behaviour, supposedly because of denser information load (less 'irrelevant' words), which allows to grab the meaning even with a narrow window. Of course, another reason for this can be a simple fluctuation caused by stochastic nature of word embedding models and small corpus size. It is also interesting that the *fiction* register model seems to represent semantic structure of language as a whole much better than others. However, we leave studying these phenomena for the future research.

Unfortunately, there is no clear way to directly evaluate how good the models are in capturing particular words' semantic differences across registers (as described in the previous section). That's why we use the values in the Table 4 as proxy measure of this, and subsequently, *RegisterExplorer* features the models with window size 3.

5 Conclusion

Thus, we presented an approach to detect differences in word semantics dependent on the language registers. It is believed that register variation can be found at different language levels. Our contribution is related to exploring this variation on the semantic tier of language. Our approach is very straightforward in implementation and based on distributional semantic models yielding state-of-the-art results in many NLP tasks.

We gave examples of semantic shifts detected by our approach and showed how it can be easily extended to detect the most likely register of an arbitrary text. As a proof of concept, we presented *RegisterExplorer* (`http://ltr.uio.no/embeddings/registers`), a web service to study difference in lexical semantics across several language registers extracted from the BNC corpus. The service comes with freely available source code[3] to enable researchers to quickly set up their own experimental design,

[3] `https://github.com/ElizavetaKuzmenko/dsm_genres`

with their own models and text types.

Among others, this framework can be used to study differences in word usage between different fiction genres or even between particular writers, or one and the same writer at certain time periods. It is also possible to analyse how dominant word meanings change depending on types of communicative situations, etc. Once the necessary corpora are available, distributional models can be trained using many available off-the-shelf tools and then easily loaded in to our framework.

In the future we plan to extend the web service with additional models trained on larger corpora (to allow exploration of finer genres), as well as improve its visualization abilities. We are also interested in defining what measures of similarity between lists of nearest associates in different models yield better performance and how to evaluate it. Furthermore, we hope that this service will enable us to investigate cross-linguistic register variation by examining comparable registers in different languages, for example, English and Russian.

References

M. Fatih Amasyal and Banu Diri, 2006. *Automatic Turkish Text Categorization in Terms of Author, Genre and Gender*, chapter Natural Language Processing and Information Systems, pages 221–226. Springer Berlin Heidelberg.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 238–247, Baltimore, USA.

Yoshua Bengio, Rejean Ducharme, and Pascal Vincent. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

Douglas Biber and Jesse Egbert. 2015. Using grammatical features for automatic register identification in an unrestricted corpus of documents from the open web. *Research Design and Statistics in Linguistics and Communication Science*, 2(1):3–36.

Douglas Biber. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

Daniel Chandler. 1997. An introduction to genre theory. *The Media and Communications Studies Site*.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2013. Linguistic profiling of texts across textual genres and readability levels. An exploratory study on Italian fictional prose. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 189–197. INCOMA Ltd. Shoumen, BULGARIA.

John Firth. 1957. *A synopsis of linguistic theory, 1930-1955*. Blackwell.

Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*.

Paul Jaccard. 1901. *Distribution de la Flore Alpine: dans le Bassin des dranses et dans quelques régions voisines*. Rouge.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schutze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL/EACL97)*, pages 32–38.

Adam Kilgarriff. 1997. "I don't believe in word senses". *Computers and the Humanities*, 31(2):91–113.

William R Knight. 1966. A computer method for calculating Kendall's tau with ungrouped data. *Journal of the American Statistical Association*, 61(314):436–439.

Andrey Kutuzov and Elizaveta Kuzmenko. 2016. Cross-lingual trends detection for named entities in news texts with dynamic neural embedding models. In *CEUR Workshop Proceedings*, volume 1568, pages 27–32.

Yong-Bae Lee and Sung Hyon Myaeng. 2002. Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 145–150.

David Lee. 2001. Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, 5(3):37–72.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. 2000. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471–495.

Matt Taddy. 2015. Document classification by inversion of distributed language representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 45–49. Association for Computational Linguistics.

Peter Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

Sven Meyer zu Eissen and Benno Stein. 2004. Genre classification of web pages. *Advances in Artificial Intelligence*, pages 256–269.