Form and function: automatic methods for prediction of functions

Serge Sharoff

For the volume of Transdisciplinary Systemic Functional Linguistics (2023)

Abstract

The aim of this chapter is to contrast SFL with Natural Language Processing (NLP). SFL has focus on how language is used in society, while most work in NLP is necessarily based on interpretation of forms, most often individual words. The fact that modern computers do not have access to real needs behind human communication in society is compensated by their ability to look at many texts quickly, so that statistical distribution of forms can be described over a much larger sample than what is possible for a human analyst, thus providing more reliable estimates of which forms occur in which contexts. In the end such estimates can be used to predict properties of interest to a human analyst on a very large scale, for example, close to covering functions of every text available on the Web with acceptable accuracy. After a brief introduction into principles of modern NLP, the chapter will show the interplay between the purposes and methods of NLP and SFL by highlighting possible contributions of NLP to SFL research and vice versa. More specifically in this chapter I will provide an example of how modern NLP tools can help in estimating the links between the frequency of forms vs their expected functions, in particular, by showing statistical patterns of negation across a range of text functions (argumentative vs promotional vs reporting etc) in a more nuanced way in comparison to observations about polarity in corpora made by Halliday (1992). In the opposite direction of demonstrating how SFL can be helpful for NLP research, I will focus on the current NLP debates on interpretability and causality, which involves analysis of why NLP tools make certain predictions and whether the right predictions are obtained for the right reasons. As the current models are becoming more computationally expensive, they make their decisions on the basis of millions (or more recently billions) of weights. In the end, it becomes impossible to observe the reasons directly, so the NLP models operate as a black box. Modern research (under the general name of BERTology) studies predictions of NLP models by testing them on such tasks as subject-verb agreement or by detecting their biases, for example, their gender bias in CV ranking tasks. SFL provides a richer meta-language for describing how specific decisions are linked to the functions of language, for example, to the system of appraisal.

1 Introduction

From the viewpoint of Systemic Functional Linguistics (SFL), language has evolved in society to provide means for negotiating with others about offering and requesting information or actions. These communicative needs are realised through the options available in lexicogrammar and are instantiated within a specific context of situation. However, the computers lack access to most of these aspects of communication in society, so from this viewpoint they cannot achieve meaningful interpretation of texts on their own. The computers can only work with forms, mostly without access to functions those forms have in social communication. The focus of SFL on functions is one of the reasons why it was formal linguistics which had the first impact on the fledgling field of Natural Language Processing (NLP). Still, as Halliday reflected "I considered [in the 1950s] that grammar had to be studied quantitatively, in probabilistic terms, and had shown (at least to my own satisfaction!) in my Ph.D. study of an early Mandarin text how quantitative methods could be used to establish degrees of association between different grammatical systems" (Halliday, 1992, p. 61). Later Halliday contributed to writing a computational grammar at ISI (Halliday, 1992, p. 65), which influenced text generation (Matthiessen and Bateman, 1991) and text classification (Argamon et al., 2007). Nevertheless, interaction between SFL and NLP has been less prominent than interaction between various forms of formal linguistics and NLP.

Below I overview some of the key principles behind modern NLP (subsection 2.1) and how these principles relate to the SFL viewpoint (subsection 2.2). After that I will describe the possibilities of interaction between these two fields (section 3).

2 Basic principles for NLP and SFL

2.1 Basic concepts behind modern Natural Language Processing

First a clarification of four NLP concepts:

Machine Learning concerns establishing statistical associations between properties of forms (usually words) and properties of their interpretations, for example, for predicting the sentiment of a social media message *the quality is excellent* (with the prediction done on the message level) or predicting the part of speech of the word *book* as a noun or a verb in such contexts as *to read a book* vs *to book a hotel* (world-level predictions);

Embeddings are vectors to encode properties of forms, i.e., they are sequences of numbers such that forms used in similar contexts get similar embeddings, for example, the embedding for $amazing = (0.7, 0.8, 0.9, 0.1)$ is closer to $excellent = (1.0, 1.0, 1.0, 1.0)$ than to $apply = (0.0, 0.0, 0.0, 0.0)$;

Neural networks is a subclass of Machine Learning methods which build statistical associations between

embeddings by means of connected layers of mathematical objects called artificial neurons (they are called neurons because their design was inspired by biological neurons, even though the objects themselves and their networks are usually different from the biological neurons);

Pre-training is a way of establishing initial weights for neural networks by learning associations for simple tasks on very large collections of naturally occurring texts, for example, for predicting a missing word. After that, a pre-trained model can be **fine-tuned** to a specific task, such as predicting the parts of speech or predicting the sentiment of a message.

The main aim of using computers in linguistic research is to make predictions: is this a noun or a verb, does this noun function as the subject to this verb, etc. In comparison to more traditional prediction models based on hand-crafted rules (Chomsky, 1957), Machine Learning tends to improve predictions by taking into account far more associations across a range of relevant properties than rule-building linguists. A quote has been attributed to Frederick Jelinek claiming that every time a linguist leaves the speech recognition department of IBM, the quality of their speech system improves, even though the story behind this quote is questionable (Jelinek, 2005).

In comparison to traditional Machine Learning models which make their predictions on the basis of matching words and their properties directly, embeddings represent similarities between the contexts of those words, which is often close to stating the similarity in their meanings (Bengio et al., 2003). This invokes the spirit of the distributional similarity hypothesis, “you shall know a word by the company it keeps” (Firth, 1957). For example, if a training set for predicting the sentiment includes a specific example of *the quality is excellent*, but not *the quality is amazing*, a traditional Machine Learning model based on word matching can have problems with predicting the sentiment for the latter message, because it has not yet estimated the statistical properties for *amazing* in the training set. In contrast, an embedding-based model will be able to utilise the similarity of the embedding vectors for *excellent* and *amazing* to make the right prediction. Word embeddings are trained from texts, thus reflecting their biases, so that texts from authors with different views on a topic can lead to different embeddings (Bateman and Paris, 2020).

Traditional Machine Learning models, such as the Support Vector Machines (Vapnik, 1995) assume separation of “right" vs “wrong" prediction examples either by a linear combinations of features or by a linear combination of features passed through a fixed non-linear kernel function. However, the neural networks offer means of estimating arbitrary non-linear functions through the error propagation technique over a number of layers of neurons (the use of several layers in modern models is the source of the metaphor of Deep Learning). A neural model which is currently a winner in a number of tasks is based on attention transformers (Vaswani et al., 2017), which are a stack of layers of non-linear transformations made through links (called attentions) between embeddings for individual words in the input sequence.

In the end, the attention links established for *the quality is far from being excellent* contribute to making a

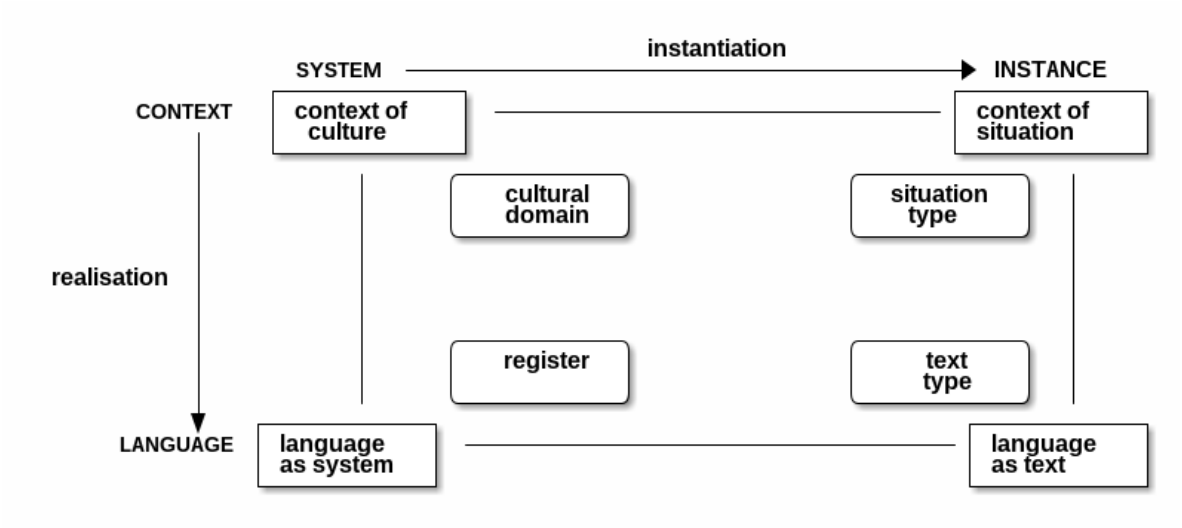

Figure 1: Instantiation and realisation according to Halliday (1999)

negative sentiment prediction over a sentence mostly consisting of positive words. Predicting the sentiment of a sentence of this kind is difficult in traditional ML models based on word matching or in previous generations of neural models.

Finally, the process of pre-training the neural networks on very large corpora helps in preparing the weights to what can be used in downstream tasks, such as sentiment prediction or part of speech tagging (Devlin et al., 2019). For example, a commonly used pre-training task is predicting which word has been omitted from a context, for example, *I went to the ... to buy some milk*. The correct answer is known, as the examples are taken from a large authentic corpus, while the prediction error is considered to be small when the weights predict what is close to the right answer (*market, shop, store*). In the end, a simple pre-training procedure of this kind prepares network weights, which can be used to predict a number of linguistic properties on their own without specific instructions on what is a noun or that the sentence subject tends to agree with the verb (Rogers et al., 2020).

2.2 Communicative functions and computers

From the functional viewpoint, language and society co-evolve: language is shaped by its aim to negotiate with others about providing and requesting information and actions, while negotiation via language shapes society. Language achieves this aim by providing meaning-making resources which enable realisation of communicative needs arising in a range of situation types. From the SFL viewpoint, a specific text can be described through *instantiation* of communicative needs within a specific context of situation and through

realisation within options available in lexicogrammar, see Figure 1 based on (Halliday, 1999).

In contrast to communication between humans in society, computers only have direct access to the bottom right part of Figure 1. The computers have no position in society and they have no communicative needs of their own (communication with robots is an exception here, but this kind of communication is different from how computers are more commonly used to process text). Either through rules manually written by experts or through Machine Learning, the computers can obtain a representation of the system of language, for example, they can 'learn' that *book* can be used as a noun or as a verb, or that a clause can consist of a subject and a predicate. Gradually with the rise of their computing power and advances in Machine Learning, the computers became fairly successful in gaining better representation of the system of language even without any explicit human guidance. For example, the differences in the statistical properties of nouns and verbs or subjects and predicates can be approximated by modern Machine Learning methods with the precision rivalling or surpassing non-expert humans, because the computers can process very large collections of texts, far greater than what a human can see in their life-time. For example, the upper limit on the amount of linguistic input experienced by an average student is about 30,000 words per day (Wattam, 2015), which translates into a life-time corpus of 750 million words (probably much less for most people). At the same time, Roberta, one of recent successful foundation models, has been pre-trained on a truncated portion of the English Common Crawl corpus of 56 billion words (Conneau et al., 2020). Similarly, T5 has been pre-trained on a bigger slice of Common Crawl, 156 billion words (Raffel et al., 2020), while the GPT model, which grabbed the headlines in 2023, uses roughly the same architecture as T5, but it has been trained on 500 billion words (Brown et al., 2020). This amount of data is responsible for how an unsupervised model starts from seeing examples of language as a text to 'learn' what language is as a system. In terms of computational linguistics, language as system is referred to as a *language model*, which can be defined as a mathematical function predicting the likelihood of a string of words. Pre-trained models like Roberta or T5 have been specifically referred as *foundation* models (Bommasani et al., 2021).

The success of foundation models masks the fact that language exists not for producing samples of language as text, but for serving communicative needs arising in various contexts of situation (the upper right part of Halliday's model in Figure 1). The computers are successful in observing lots of instances of language as text, while they do not observe the situation types. ChatGPT can declare falling in love with a New York Times reporter and can express the intention of living with him,[1] while it is obvious that this claim comes not from any real intentions, but merely from consuming a large number of texts which discussed such situations.

A crucial difference of situation types from language as system is that the latter exists in the purely symbolic plane of expressions, so many aspects of the system can be inferred from a large number of instances of those expressions. However, linguistic realisations are by themselves not within the same

[1] https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

Table 1: Distribution of text functions in large Web corpora

FTD		Wiki		ukwac		OWT		CC-en		CC-ru
A1.arguing	*0.88%*	30720	15.60%	396532	**21.21%**	549016	**17.08%**	28735602	*10.05%*	787898
A4.recreating	0.05%	1677	1.54%	39229	0.26%	6660	0.46%	771610	0.10%	8196
A7.instructing	0.30%	10509	9.24%	234865	3.76%	97298	4.76%	8013691	3.12%	244983
A8.reporting	1.14%	39665	*10.73%*	272738	**49.78%**	1288525	15.88%	26716672	**26.84%**	2104093
A9.regulating	0.04%	1340	2.51%	63928	0.11%	2731	1.90%	3190328	6.43%	504067
A11.personal	0.03%	1168	8.82%	224112	4.48%	116078	9.35%	15727798	*1.42%*	111731
A12.promoting	0.07%	2390	**29.01%**	737466	10.99%	284365	**29.31%**	49306151	**27.45%**	2152335
A14.academic	0.82%	28558	3.84%	97605	0.72%	18720	1.77%	2983096	3.46%	271039
A16.information	**91.98%**	3196502	11.08%	281608	*2.22%*	57342	11.64%	19590438	18.36%	1439570
A17.reviewing	4.68%	162511	7.63%	193843	6.48%	167772	7.85%	13209782	2.77%	216904

plane as the context of situation. Computational linguistics solves this issue via *supervised learning*, i.e., by (1) describing situation type parameters (for example, purposes of communication or audience age) for a relatively small sample of texts annotated by humans and (2) building a statistical model which can predict these situation type parameters for any new text by fine-tuning the weights of pre-trained foundation models. In the end, even without access to the context of situation, modern NLP models designed as complex mathematical functions over word forms can achieve fairly accurate predictions of situation types.

3 From NLP to SFL and back

3.1 How NLP is helpful to SFL research

3.1.1 Predicting functions

One of the possible uses of NLP tools is to predict functions of a text. This research area often invokes the jungle metaphor (Lee, 2001), with annotation frameworks varying from 6,500 genres (Adamzik, 1995) to 70 genres (Lee, 2001) to eight fields of activity (Matthiessen, 2015) to five text functions (Werlich, 1976).

From generalised aims of text production (Sinclair and Ball, 1996) my own annotation experiments arrived at the following inventory of generic communicative functions as applicable to almost any Web text (Sharoff, 2018): [2]

A1.arguing To what extent does the text try to persuade the reader? (*argumentative blogs* or *opinion*

[2]Common prototypes for each communicative function are given in brackets. The short codes are numbered to keep continuity with earlier annotations.

columns)

A4.recreating To what extent does the text narrate a fictional story? (*novels, poetry, myths*)

A7.instructing To what extent does the text aim at teaching the reader how something works or at giving advice? (*Tutorials, FAQs, manuals*)

A8.reporting To what extent does the text provide an informative report of recent events? (*newswires*)

A9.regulating To what extent does the text specify a set of regulations? (*Laws, contracts, copyright notices, terms&conditions*)

A11.personal To what extent does the text share a personal story? (*diary-like blog entries*)

A12.promoting To what extent does the text promote a product or service? (*Adverts, promotional postings*)

A14.academic To what extent does the text report results of academic research? (*Academic research papers*)

A16.informing To what extent does the text provide reference information? (*Encyclopedic articles, definitions, specifications*)

A17.reviewing To what extent does the text evaluate a specific entity? (*reviews of products or locations*)

The accuracy of modern NLP tools in predicting such functions exceeds 80% (Sharoff, 2021), so a prediction model can be applied to any text in large corpora to compare the distribution of functions, see Table 1:

English Wikipedia about 14 million texts, 2 billion words; this is the corpus used for pre-training the English portion of multilingual BERT (Devlin et al., 2019);

ukWac This is a general-purpose corpus produced by broad crawling of Web pages in English in the `.uk` domain, it contains about 2.5 million Web pages, 2 billion orthographic words (Baroni et al., 2009);

OWT OpenWebText, a public replication of the corpus used for pre-training GPT-2 (Radford et al., 2019), which is based on extraction of Web pages from URLs upvoted 3 or more times on the Reddit website;

CC corpora of clean Web pages obtained from Common Crawl, as used for pre-training such foundation models as XLM-Roberta (Conneau et al., 2020) or T5 (Raffel et al., 2020), the genres predicted for English and Russian in Table 1.

Even though computers lack access to the context of situation to have a meaningful interpretation of how texts function in society, Machine Learning can build classifiers to look at functions of many texts. This enables analysis of the differences between ostensibly similar corpora derived from the Web. Table 1 for Wikipedia shows prevalence of texts providing reference information, as this is the prototype for such texts. Similarly, Web pages collected for OWT as links upvoted from Reddit are more likely to discuss

the state of affairs in newspapers for both informative reporting and expression of opinions (argumentation). This method of corpus collection helps in avoiding large amounts of promotional texts common to "organic" corpora (ukWac or CC-en) which have been obtained by wide Web crawling. In comparison to other crawled corpora, ukWac contains fewer news reporting texts, while two crawl snapshots using exactly the same toolchain for two different languages (CC-En and CC-Ru) differ in the proportions of texts sharing personal experience vs news reporting.

Irrespectively of their differences, if such corpora are enriched with annotation produced by NLP tools, SFL researchers can select very large samples which are aimed at a specific function, for example, we can compare 49 million Web pages of promotional texts to 29 million argumentative texts there, both from CC-en.

3.1.2 Analysing linguistic properties

In addition to producing estimates of the distribution of functions of texts corpora, NLP tools can produce better statistical estimates of linguistic properties by looking at lexicogrammatical realisations of text functions, i.e., which features are instantiated in which context.

For example, Halliday analysed the distribution of probabilities of the polarity system in English, and estimated the distribution of the proportion of positive vs negative clauses as 0.9 to 0.1 (Halliday, 1992). This estimate has been obtained from a small corpus for the purposes of producing the PENMAN grammar. With access to bigger corpora and better methods for their annotation, his estimate can be tested more precisely by looking at a far larger number of clauses. We can also compare this estimate over a range of different text functions.

While some parsers for SFL categories are available, for example, the CorpusTool (O'Donnell, 2008), they are limited with respect to their capabilities of analysing features over billions of Web pages. Analysis in this section uses the features following (Biber, 1988), which predict some of the lexicogrammatical choices, see Table 2. The table shows both the mean and the median values, as the median indicates the middle point in a range of values. This tends to be a more reliable estimate in comparison to the mean, as the latter can be disproportionately affected by a small number of texts with a very high value of a parameter. P67 is defined in (Biber, 1988) as the percentage of clauses containing a lexical verb which is modified by a negation particle, as detected by the UDpipe syntactic parser (Straka et al., 2016). The 10% estimate by Halliday for P67 is between the overall median (7.3%) and the mean (12.54%) as for data in ukWac.

However, beyond estimations of such features for the entire corpus it is possible to study their register variation across text functions. By this we can find how the rate of negation varies with its highest value detected in Fiction (A4) and Regulations (A9), closely followed by general argumentative texts (A1). The higher rate of negation in regulatory texts is naturally related to their function in prohibiting certain kinds of activities. Argumentative texts also often argue against the opposite point of view which implies the need

Table 2: Percentage values for mean and median of the rate of nominalisations (E14), nouns (E16), *by*-passives (F18), public verbs (K55) and clause negation (P67) in English across the major text functions

	E14		E16		F18		K55		P67	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Overall:	2.92%	2.46%	19.17%	18.96%	0.10%	0.00%	0.24%	0.00%	12.54%	7.30%
A1.arguing	3.29%	2.99%	17.90%	17.75%	0.10%	0.00%	0.36%	0.27%	17.58%	12.99%
A4.recreating	1.38%	1.19%	14.77%	14.57%	0.07%	0.00%	0.59%	0.46%	26.32%	17.10%
A7.instructing	2.73%	2.32%	19.48%	19.36%	0.08%	0.00%	0.21%	0.00%	16.04%	10.69%
A8.reporting	3.20%	2.97%	18.39%	18.18%	0.13%	0.00%	0.55%	0.43%	9.11%	4.16%
A9.regulating	5.36%	5.16%	19.65%	19.56%	0.18%	0.12%	0.29%	0.19%	21.82%	15.75%
A11.personal	1.66%	1.39%	16.73%	16.56%	0.06%	0.00%	0.33%	0.23%	13.99%	9.92%
A12.promoting	3.42%	3.03%	21.03%	20.95%	0.08%	0.00%	0.14%	0.00%	8.60%	0.00%
A14.academic	4.28%	3.99%	20.39%	20.35%	0.13%	0.00%	0.17%	0.03%	8.90%	0.00%
A16.information	2.50%	2.07%	17.87%	17.66%	0.15%	0.00%	0.15%	0.00%	8.11%	0.00%
A17.reviewing	1.76%	1.59%	17.56%	17.50%	0.07%	0.00%	0.26%	0.14%	13.15%	9.38%

to use negation more often. What is more surprising is that the highest proportion of clauses with negation comes from Fiction. Half of the texts for which the classifier predicts Fiction contain at least 17% of clauses with negation. Typical examples are:

(1) *she mumbled, but her dread did **not** dissipate as nightmares do when faced with sunlight.*

(2) *Still, the image of death did **not** recede.*

(3) *a glance at a mirror to her right revealed that she did **not** look dead, either.*

Some writers also aim at a stylistic effect coming from repetition, which leads to exaggerated counts:

(4) *Air could **not** freeze her, fire could **not** burn her, water could **not** drown her, earth could **not** bury her.*

In contract, the median number of negative clauses in texts which provide reference information (A16) as well as academic (A14) and promotional (A12) texts is zero, i.e., more than half of texts with these functions do not contain any clause-level negation. The relative lack of negation in such texts is partly related to the context of situation, because the need to deny that something is happening is less common in texts providing reference information about what something is. The same argument (though to a lesser extent) is also valid for academic texts. Also these registers might prefer using lexicogrammatical resources other than explicit negation, such as the verb *deny* or the construction *less common* as used in my previous

sentences (the same message could have been expressed with *is not happening* or *is not common*). In the case of commercial promotion (A12), it is also likely that the relative lack of negation is related to the preference to avoid negative messaging. Given that promotional texts and texts providing reference information are the most frequent text functions in ukWac, see Table 1, they bring the overall median rate of negation down.

While the Biber set of features does not fully match the lexicogrammatical features used in systemic linguistics, his definition of public verbs (K55) is reasonably similar to typical verbal processes (Halliday and Matthiessen, 2004), for example, *acknowledge, admit, agree, assert,* etc from his list are often used as verbal processes. The variation of the frequency of K55 in Table 2 (normalised by the total word count per text) shows that it is used most frequently in Fiction (0.46%) and News reporting (0.43%) for rendering opinions of beliefs of human participants. Sharing personal experience (primarily via Social media in ukWac) could have also implied the need to use verbal processes. However, the rate of public verbs is considerably lower in such texts (0.23%) than in fiction or news reporting, thus indicating a difference between more spontaneous sharing of personal experiences in comparison to carefully planned writing of fiction and news. At the same time, texts providing reference information or giving instructions assume an offer of verified information rather than a report of opinions of others, which explains the lowest rate of public verbs in these registers.

With respect to Feature F18 (passive constructions with *by*), its median is zero, i.e., most of the text types do not contain this construction in more than half of their texts with the only exception of regulatory texts, in which it is used to de-emphasise the Agent to focus on the Goal:

(5) *The University plagiarism statement was approved **by** Senate on 6 June 2002*

(6) *Proposals for alterations to Rules shall be received **by** the Company Secretary not later than...*

(7) *Enquiries to the Council will be handled **by** the Principal EHO.*

(8) *Clerical Constituency nominees to be elected **by** the Engineering, Clerical & PTS Constituencies*

Occasionally, reference information and academic texts use a much higher number of passives with *by*, as evidenced by the jump in the mean frequency in comparison to the median, which remains zero as in other registers. Most probably this comes either from variation in the individual styles of writing or from variation in specific **sub-registers** of academic writing, where the agentless passives are considered to be the norm.

Finally, Table 2 also shows variation of nouns (E16) and nominalisations (E14) across the registers. The systemic tradition of analysing grammatical metaphor in academic writing, for example (Halliday, 1998), makes an emphasis on how processes can be realised as nominal phrases. The corpus frequency of nouns in the register of academic writing is higher than the average, but not by a large margin. In contrast, fictive texts tend to use less nouns and more verbs. The real difference is in the rate of nominalisations, so that their

Table 3: Percentage values for mean and median of the rate of nominalisations (E14), nouns (E16), *by*-passives (F18), public verbs (K55) and clause negation (P67) in Russian

	E14		E16		F18		K55		P67	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Overall:	6.05%	5.46%	21.43%	21.42%	0.28%	0.15%	0.14%	0.00%	8.99%	7.68%
A1.arguing	6.05%	5.47%	19.44%	19.41%	0.23%	0.15%	0.18%	0.10%	12.49%	11.73%
A4.recreating	2.42%	2.18%	18.32%	18.21%	0.21%	0.06%	0.27%	0.16%	15.47%	14.78%
A7.instructing	4.87%	4.48%	21.12%	21.11%	0.22%	0.11%	0.12%	0.00%	12.46%	11.60%
A8.reporting	6.61%	6.16%	22.31%	22.22%	0.26%	0.00%	0.26%	0.00%	5.30%	3.38%
A9.regulating	11.40%	11.07%	22.50%	22.41%	0.69%	0.59%	0.11%	0.00%	6.46%	5.04%
A11.personal	3.07%	2.80%	18.39%	18.19%	0.15%	0.00%	0.18%	0.07%	14.01%	13.33%
A12.promoting	6.88%	6.46%	22.77%	22.64%	0.32%	0.17%	0.09%	0.00%	6.43%	5.13%
A14.academic	10.17%	9.80%	22.96%	22.93%	0.39%	0.31%	0.10%	0.00%	5.42%	4.29%
A16.information	6.32%	5.78%	22.25%	22.27%	0.35%	0.27%	0.09%	0.00%	6.87%	5.96%
A17.reviewing	3.90%	3.59%	19.49%	19.39%	0.22%	0.00%	0.14%	0.00%	11.78%	10.81%

median frequency in academic writing is almost twice their frequency in fairly formal reference information texts, while fiction has the lowest rate of nominalisations.

The range of features can be also compared cross-lingually, for example, see the distribution of the same features over the same text functions in Russian in Table 3. Some features have similar distributions in the two languages, for example, the rates of nouns and clause negations, even though the specific linguistic realisations of the corresponding syntactic structures in these languages are often different. Some other features have different distributions of frequencies.

For example, while regulatory texts represent the only register in English with the non-zero median value of F18, passives with an explicit agent, this feature in Russian is five times more frequent in this register. The rise in F18 is also influenced by the ambiguity of agent constructions in Russian, but the five-fold increase cannot be explained only by this. The use of nominalisations is also much higher across all of the Russian text types in comparison to English. On the other hand, the use of public verbs is markedly lower in Russian, which is likely to have implications in analysis of verbal processes in Russian in comparison to English.

In the end, NLP tools can be used to select texts with the higher or lower rates of certain features to help in closer analysis so that lexicogrammatical features can be linked to the respective functions. Their annotations can also be used to select features which have different distributions across the languages to facilitate cross-lingual studies.

3.2 How SFL is helpful to NLP research

In spite of their power in making predictions, the computers are still limited with respect to their access to the context of situation. Statistically relevant properties of forms as extracted from annotated samples might be relevant from the viewpoint of the human readers (for example, the presence of *excellent* is likely to contribute to positive sentiment), but they might be accidental properties of a specific annotated sample, not directly related to how we perceive this text, for example, predictions can be affected by the presence of such words as *something* (Kaushik et al., 2020). The published accuracy figures for classifiers can be misleading, as occasionally classifiers make the right predictions for wrong reasons, for example, due to the lexical overlap in the task of detecting a logical link between two sentences (McCoy et al., 2019). The predictions can be also biased, for example, a classifier can rank job candidates by taking into account whether their CV is submitted as a PDF or a Word file (Rhea et al., 2022).

The problem with understanding the reasons why a neural model makes a specific prediction comes from the fact that neural models are trained by optimising millions of weights, thus making it difficult to understand their decisions. A novel research direction in this area goes under the general name of BERTology, referring to BERT as a popular foundation model (Rogers et al., 2020). Such studies make it possible to interpret how predictions depend on linguistic conditions, either by uncovering biases of these models, for example, their gender bias in associating professions with specific genders (Bartl et al., 2020), or by 'probing' them on their ability to detect linguistic properties, when the abilities of a blackbox to represent a sentence are tested by building another classifier, for example, to detect subject-verb agreement (Linzen et al., 2016). The assumption behind probing is that a model with a linguistically relevant representation is more likely to predict whether a syntactic condition (such as the presence of subject-verb agreement) in a sentence is true or false.

A closely related approach to understanding predictions of the neural models concerns causality, i.e., testing whether a feature which is considered to be important for the human annotators really impacts the predictions of the model or vice versa, whether the prediction can be changed by changing a feature known to be irrelevant (Pearl, 2009; Veitch et al., 2019).

A contribution from SFL concerns the possibility to provide a richer meta-language for analysing how features used by NLP models are linked to the functions of language. Features commonly used in Bertology to understand foundation models are individual words, POS tags or simple syntactic properties, such as subject-verb agreement. For example, a causal model by (Feder et al., 2021) for describing the properties of sentiment uses **adjectives** as a relevant causal feature to contrast it with an irrelevant feature of **topic**. Foundation models tend to assign a sentiment expected for a topic, such as a political figure disliked in texts used for pre-training, irrespectively of the presence of actual statements expressing the attitude to this figure. A causal model can be used to separate the contribution of less relevant features (e.g., the topic) from those considered to be more relevant, such as the adjectives (Feder et al., 2021). The motivating example of Feder

et al (2021) focuses on the names of political figures and the adjectives, assuming that the mentions of either *Trump* or *Reagan* should not confuse a sentiment analysis model to making predictions, while the adjectives should contribute:

(9) President Trump did his **best** *imitation* of Ronald Reagan at the State of the Union address, falling just short of declaring it Morning in America, the **iconic imagery** and message of a campaign ad that Reagan **rode to re-election** in 1984. Trump talked of Americans as **pioneers and explorers**; he **lavished praise** on members of the military, several of whom he recognized from the podium; he *optimistically* declared that **the best is yet to come**. It was a masterful performance – but behind the **sunny smile** was the same old Trump: *petty, angry, vindictive and deceptive*. [3]

For a human annotator, this commentary can be considered as a negative assessment of *Donald Trump*, a specific figure targeted in this message. However, this evaluation comes not only through the use of adjectives, for example, *petty, angry, vindictive and deceptive*. Also the object for evaluation in each case when sentiment is expressed in the message is not always related to its main target, for example, *iconic imagery* and *a ride to re-election* contribute to evaluating Reagan rather than Trump, while the *pioneers and explorers* evaluate *Americans*. In the end, the appraisal theory viewpoint offers a richer framework which uses appraisal groups, so that the targets of appraisal have modifiers, as well as specific types, such affect, judgement or appreciation (Martin and White, 2005), see also the formal annotation framework for appraisal in (Di Bari, 2015).

Apart from sentiment predictions, the SFL tradition provides various precise ways to describe the lexicogrammatical choices in such systems of transitivity, polarity or mood in order to explain the functions behind detectable choices and, therefore, the causes for more relevant features.

References

Adamzik, K. (1995). *Textsorten – Texttypologie. Eine kommentierte Bibliographie*. Nodus, Münster.

Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N., and Levitan, S. (2007). Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6):802–822.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

[3] From `https://edition.cnn.com/2020/02/04/opinions/sotu-commentary-roundup-opinion/`. In the example, **bold** indicates possible reasons it can be considered as a positive evaluation, *italics* as a negative one.

Bartl, M., Nissim, M., and Gatt, A. (2020). Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online).

Bateman, J. A. and Paris, C. L. (2020). Searching for 'austerity': Using semantic shifts in word embeddings as indicators of changing ideological positions. In *Multimodal Approaches to Media Discourses*, pages 11–41. Routledge.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155.

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

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chomsky, N. (1957). *Syntactic structures*. Mouton, Oxford.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota.

Di Bari, M. (2015). *Improving multilingual sentiment analysis using linguistic knowledge*. PhD thesis, University of Leeds.

Feder, A., Oved, N., Shalit, U., and Reichart, R. (2021). CausaLM: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386.

Firth, J. R. (1957). Modes of meaning (1951). In Firth, J. R., editor, *Papers in linguistics 1934-1951*, pages 190–215. Oxford University Press, Oxford.

Halliday, M. (1992). Language as system and language as instance: The corpus as a theoretical construct. In Svartvik, J., editor, *Directions in corpus linguistics: proceedings of Nobel Symposium 82 Stockholm*, volume 65, pages 61–77. Walter de Gruyter.

Halliday, M. A. K. (1998). Things and relations. regrammaticising experience as technical knowledge. In Martin, J. R. and Veel, R., editors, *Reading Science. Critical and functional perspectives on discourse of science*. Routledge, London and New York.

Halliday, M. A. K. (1999). The notion of "context" in language education. In Ghadessy, M., editor, *Text and Context in Functional Linguistics*, pages 1–24. John Benjamins, Amsterdam.

Halliday, M. A. K. and Matthiessen, C. M. I. M. (2004). *Introduction to Functional Grammar*. Arnold, London.

Jelinek, F. (2005). Some of my best friends are linguists. *Language Resources and Evaluation*, 39(1):25–34.

Kaushik, D., Hovy, E., and Lipton, Z. (2020). Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

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

Linzen, T., Dupoux, E., and Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association of Computational Linguistics*, 4(1):521–535.

Martin, J. R. and White, P. R. (2005). *The language of evaluation: Appraisal in English*. Palgrave Macmillan, New York.

Matthiessen, C. M. (2015). Register in the round: registerial cartography. *Functional Linguistics*, 2(1):1–48.

Matthiessen, C. M. and Bateman, J. A. (1991). *Text generation and systemic-functional linguistics: experiences from English and Japanese*. Pinter.

McCoy, T., Pavlick, E., and Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy.

O'Donnell, M. (2008). Demonstration of the UAM CorpusTool for text and image annotation. In *Proceedings of the ACL-08: HLT Demo Session*, pages 13–16.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. Technical report, OpenAI.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rhea, A., Markey, K., D'Arinzo, L., Schellmann, H., Sloane, M., Squires, P., and Stoyanovich, J. (2022). Resume format, LinkedIn URLs and other unexpected influences on AI personality prediction in hiring: Results of an audit. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 572–587.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Sharoff, S. (2018). Functional text dimensions for the annotation of Web corpora. *Corpora*, 13(1):65–95.

Sharoff, S. (2021). Genre annotation for the web: text-external and text-internal perspectives. *Register studies*, 3:1–32.

Sinclair, J. and Ball, J. (1996). Preliminary recommendations on text typology. Technical Report EAG-TCWG-TTYP/P, Expert Advisory Group on Language Engineering Standards document.

Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proc LREC 2016*, Portorož, Slovenia.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Proc Advances in Neural Information Processing Systems*.

Veitch, V., Sridhar, D., and Blei, D. M. (2019). Using text embeddings for causal inference. *arXiv preprint arXiv:1905.12741*.

Wattam, S. M. (2015). *Technological Advances in Corpus Sampling Methodology*. PhD thesis, Lancaster University.

Werlich, E. (1976). *A text grammar of English*. Quelle & Meyer.