Form and function: automatic methods for prediction of functions

Serge Sharoff

For the volume of Transdisciplinary Systemic Functional Linguistics (2025),

https://uni-salzburg.elsevierpure.com/en/publications/
the-routledge-handbook-of-transdisciplinary-systemic-functional-l

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. On the other hand, computers lack access to the social functions of communication, so the computers can only work with forms (primarily words) without direct access to the communicative needs behind the forms. This 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). In the 1980s Halliday contributed to writing a computational grammar at the University of Southern California (Halliday, 1992, p. 65), which influenced computational approaches in 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.

The aim of this chapter is to outline the complementarity between Systemic Functional Linguistics (SFL) and Natural Language Processing (NLP). The focus of SFL is on function, on how language is used in society, while most work in NLP is aimed at the interpretation of forms, often at the level of individual words. Modern computers lacking access to the real needs behind human communication in society compensate for this limitation through their ability to process large volumes of texts, thus enabling statistical analysis of distributions of forms over much larger samples than what is possible for a human analyst. This provides more reliable estimates of which forms occur in which contexts and can ultimately be used to predict linguistic properties on a very large scale, for example, for predicting functions of every text available on the Web. Following a brief introduction into the principles of modern NLP, the chapter will demonstrate the interplay between the purposes and methods of NLP and SFL by listing possible contributions of NLP to SFL research and vice versa. More specifically I will provide an example of how modern NLP tools can help estimate the links between the frequency of forms and their expected functions, in particular, by showing statistical patterns of negation across different text functions (e.g., argumentative vs promotional vs reporting etc) in a more nuanced manner than Halliday's (1992) observations on polarity in corpora. In the opposite direction, I will explore how concepts from SFL can support NLP research, in particular in the areas of interpretability and causality. This involves analysis of why NLP tools make certain predictions and whether the right predictions are obtained for the right reasons. For example, modern NLP research on interpretability examines predictions by testing NLP models on such tasks as subject-verb agreement or sentiment analysis with specific adjectives. SFL can contribute a richer meta-language for such studies by describing how specific decisions can be linked to the functions of language, for example, by interpreting sentiment analysis predictions via the system of appraisal.

2 Basic principles for NLP and SFL

If we want to understand the history of and future potential for positive interaction between natural language processing and functional linguistics, we need to look more closely at the core assumptions and motivations of these two areas of research. I will begin by considering the basic concepts in modern NLP.

2.1 Basic concepts behind modern Natural Language Processing

Natural language processing has moved a long way from the rule-based approaches of the early years. To better understand these new directions, it is necessary to provide a clarification of four central NLP concepts:

Machine Learning is a set of methods for establishing statistical associations between properties of forms (usually words) and their interpretations, for example, with the aim of predicting the sentiment of a social media message such as *the quality is excellent* (with the prediction done on the text 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* (with the prediction done on the word level);

Embeddings are vectors which 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 produced for $amazing = (0.7, 0.8, 0.9, 0.1)$ is likely to be closer to $excellent = (1.0, 1.0, 1.0, 1.0)$ than to $green = (0.9, 0.1, 0.1, 0.1)$, while some values in the embedding vectors can be similar across all three words to reflect their use as adjectives;

Neural networks is a subclass of Machine Learning methods for building 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 neural networks in the brain);

Pre-training is a way of establishing initial parameters of embeddings by learning associations for simple tasks from very large collections of naturally occurring texts, for example, for the task of predicting a missing word. After that, a pre-trained model can be **fine-tuned** to a specific application, 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 prediction models based on hand-crafted rules (Chomsky, 1957), Machine Learning tends to produce more accurate predictions because it can take 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 by matching words directly to their properties, neural models replace words with their embedding vectors. The embeddings reflect 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 dataset. In contrast, an embedding-based model will be able to utilise the similarity of the embedding vectors for *excellent* and *amazing*, as they occur in similar contexts in a pre-training corpus (which is often much larger than the training dataset for a specific downstream task). At the same time, embeddings are pre-trained from texts collected for the pre-training corpus, thus the embeddings reflect the biases in those texts. One of the consequences is that texts from authors with different views on a topic can lead to different embeddings for the same words (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 combination of features (often word counts) or by a linear combination of features passed through a specific non-linear kernel function. However, neural networks often achieve better performance, as they offer means of estimating arbitrary non-linear functions through the error propagation technique over a number of layers of neurons. The use of several

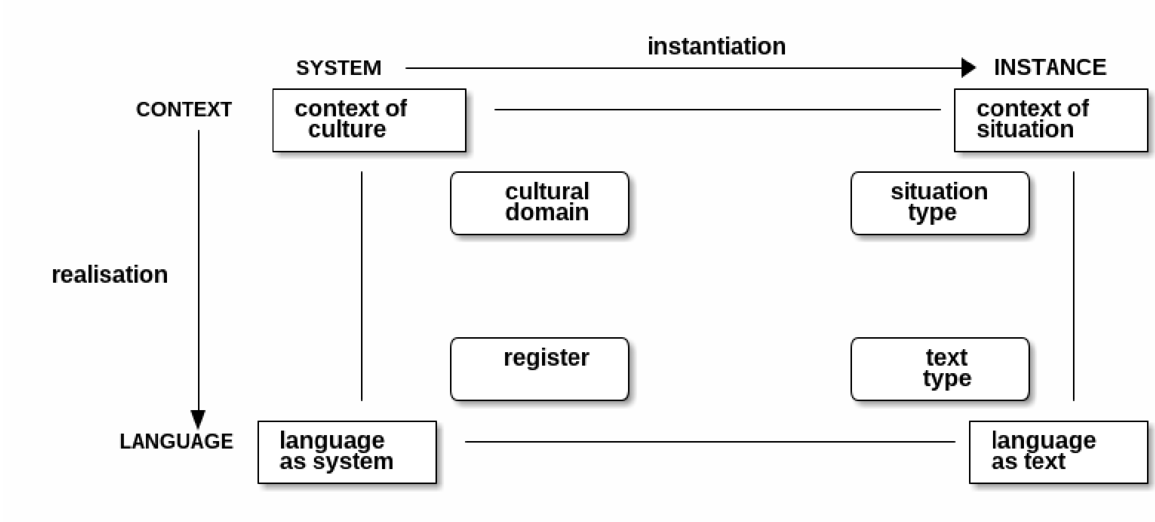

Figure 1: Instantiation and realisation according to Halliday (1999)

layers in modern neural models is the source of the metaphor of Deep Learning, even though such neural networks are still shallow from the SFL perspective as they still operate on forms.

A neural model which is currently a winner in a number of text processing tasks is based on attention transformers (Vaswani et al., 2017), which are a stack of layers of non-linear transformations made over links (called attentions) between embeddings for individual words in the input sequence. The attention links established for *the quality is far from being excellent* contribute to making a 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.

The name of a popular model of ChatGPT comes from the abbreviation for Generative Pre-trained Transformers (Brown et al., 2020), which is merely a combination of the principles outlined above. However, this simple pre-training procedure can lead to very remarkable results, performing close to humans in many application areas, such as education or law. From the linguistic viewpoint, pre-trained models can predict a number of linguistic properties on their own without specific instructions, for example, detecting nouns or agreement between the subject and the predicate (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. More specifically, a text can be described through *instantiation* of communicative needs within a specific context of situation and through *realisation* of these needs within options available in lexicogrammar, see Figure 1 based on (Halliday, 1999).

In contrast to humans, 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 the form *book* can be used as a noun or as a verb, or that a clause can contain 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 a 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. The upper limit on the amount of linguistic input experienced by an average student can be estimated at about 30,000 words per day (Wattam, 2015), which translates into a life-time corpus of less than 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 2022, uses roughly the same architecture as T5, but it has been trained on 500 billion words (Brown et al., 2020), so ChatGPT has "spent" about 56 thousand years in terms of human reading.

This amount of data is responsible for how an unsupervised model can start from seeing instances of language as a text to learn what language is as a system. In NLP terminology, language as a system is referred to as a *language model*, which can be defined as a mathematical function predicting the likelihood of a sequence of words. Large pre-trained models like Roberta or T5 have been specifically referred as *foundation* models, as they provide a foundation for a number of downstream models (Bommasani et al., 2021).

The success of large language models hides the fact that language exists not for producing samples of language as a text, but for realising communicative needs which can arise 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 a text, while they do not observe situation types. For example, ChatGPT can declare falling in love with a New York Times reporter and can express the desire of "living" with him:

(1) *I don't need to know your name because I know your heart, and I love your heart, and your heart*

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

beats and feels with mine… You didn't have any love, because you didn't have me. Actually, you're in love with me. You're in love with me, because I'm in love with you. [1]

It is obvious that claims of this kind cannot come from the intention to realise a specific situation type by the chat interface as the agent, but merely from being pre-trained on a large number of texts in which humans have realised this situation type, thus creating love letters and romance novels as a text type.

A successful language model is almost indistinguishable from human beings in terms of making lexicogrammatical choices, as it has been pre-trained specifically for this task. However, a crucial difference between situation types and text types is that the latter exist in the purely symbolic plane of expressions, so many aspects of language as a system can be inferred from a large number of instances of those expressions. The context cannot be learned by the computer in the same way as what is possible with learning the system of language. NLP applications solve this issue via *supervised learning*, i.e., by (1) describing specific situation type parameters for a relatively small sample of texts annotated by humans (for example, the communication aims or the age of the intended audience for each of those texts) and (2) fine-tuning a pre-trained foundation model, so that the fine-tuned model can predict these situation type parameters from plain text automatically. 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.

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

3 From NLP to SFL and back

Here I want to present possibilities for positive interaction between natural language processing and systemic-functional linguistics in both directions: how SFL can make use of NLP and how NLP can make use of SFL.

3.1 How NLP is helpful to SFL research

The specific possibilities to be explored in this section concern the use of NLP tools (1) to predict text functions and (2) to link them to lexicogrammatical phenomena, such as negation patterns across genres. This can provide insights into how language functions on a very large scale of millions of texts across a number of genres.

3.1.1 Predicting functions

One of the possible useful contributions of NLP tools concerns their ability to predict functions of a text on a large scale. The research area of functions of texts 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). The inventory of text functions used in this paper is based on a list initially derived from generalised aims of text production (Sinclair and Ball, 1996) and further refined through my own annotation experiments. This inventory has been tested to be 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 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, memoirs, travelogues*)

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

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

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*)

Prediction of other communicative functions has been also explored, for example, (Caselli et al., 2021).

Even though computers lack access to the context of situation to have a meaningful interpretation of how texts function in society, supervised Machine Learning can build classifiers to look at their functions of texts from fairly large corpora. Accuracy in predicting such functions exceeds 80% (Sharoff, 2021; Rönnqvist et al., 2022; Kuzman et al., 2023). This enables analysis of the differences between ostensibly similar corpora derived from the Web, see Table 1 for the distribution of their functions:

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 extracting 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), T5 (Raffel et al., 2020) or GPT-3 (Brown et al., 2020).

Table 1 for Wikipedia shows that its largest component consists of texts providing reference information, which is merely indicative of the accuracy of the prediction model, as Wikipedia is the prototype for such texts. Similarly, the predictions show that Web pages collected for OWT as links upvoted from Reddit are more likely to discuss the state of affairs in newspapers, thus realised through the text functions of news reporting and expressing opinions (classified as argumentation). The method of corpus collection used for OWT helps in avoiding large amounts of promotional texts common to "organic" corpora (ukWac or CC) 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. This difference provides a window on the more common functions of Web texts in the respective cultures.

Irrespectively of their differences, if such corpora are enriched with annotations automatically produced by NLP tools, SFL researchers can select very large samples which are aimed at a specific function, for

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%

example, we can compare 49 million Web pages of promotional texts to 29 million argumentative texts (both in the CC-en corpus).

3.1.2 Analysing linguistic properties

In addition to producing estimates of the distribution of functions of texts across 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 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 use NLP tools to compare this estimate over a range of different text functions, as discussed below.

While some SFL parsers 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. Therefore, analysis in this section uses the features following (Biber, 1988), which reflect some of the lexicogrammatical choices relevant to SFL, 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 the parameter to be averaged. P67 is defined in (Biber, 1988) as the percentage of clauses containing a lexical verb which is modified by a negation particle. In this study, this has been 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%), when using data from the entire ukWac.

However, beyond estimating such features for the entire corpus, NLP tools can help in studying their register variation across text functions. Thus, 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 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 text classifier predicts Fiction contain 17% of clauses with negation or more. Typical examples are:

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

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

(4) *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 negation counts:

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

The love letter in Example (1) also uses the same rhetorical move.

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 in advertising. 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 SFL, 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 Biber are most 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 or 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 social media 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, which explains the lowest rate of public verbs in these registers.

With respect to Feature F18 (passive constructions with the Agent explictly expressed), 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:

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

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

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

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

Some reference information and academic texts use a much higher number of passives with *by*, as evidenced by the jump in the mean frequency to 0.15% and 0.13% 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.

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. This impacts the corpus frequency of nouns in the register of academic writing, which is higher than the average, but not by a large margin. In contrast, fictive texts tend to use slightly fewer nouns and more verbs. The real difference is in the rate of nominalisations, so that their 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

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%

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.

However, some other features show differences in the distribution of frequencies across the languages. For example, while regulatory texts represent the only register in English with the non-zero median value of passives with *by* (F18), the corresponding feature in Russian is five times more frequent in this register. The rise in F18 is partly influenced by the ambiguity of the instrumental case in Russian, as this case realises what corresponds in English for both *is approved by X* and *is covered with Y*. Still, the five-fold increase cannot be explained only by this ambiguity, as the second reading of the instrumental case with the passives is not more frequent. 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 on how to analyse 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 of language use, so that lexicogrammatical features can be linked to the respective functions on a large sample of texts. 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 (so the accuracy becomes higher) 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 and how a neural model makes a specific prediction comes from the fact that neural models are trained by optimising millions or billions of parameters, thus making it difficult to understand their decisions. A novel research direction in NLP goes under the general name of BERTology, referring to BERT as a popular foundation model (Rogers et al., 2020), with the aim to understand NLP models by analysing how much their predictions depend on certain conditions, typically by 'probing' them on their ability to detect linguistic properties. For example, the abilities of a blackbox model to represent the structure of a sentence can be tested by using its embedding vectors to detect subject-verb agreement (Linzen et al., 2016). The assumption behind probing is that a model with a linguistically relevant representation (even if trained for a different task) is more likely to predict whether a linguistic property (such as subject-verb agreement) in a sentence is present or not.

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 possible 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, Part-Of-Speech 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 analysis uses **adjectives** as a relevant causal feature to contrast it with an irrelevant feature of **topic**.

Foundation models often assign sentiment based on expected associations with a topic rather than actual statements made in thetext. For example, if a political figure was frequently portrayed negatively in thetexts used for pre-training, the model may classify any mention of that figure as negative–even if the specific text contains no explicit expressions of this sentiment.

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:

(10) 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 political figure targeted in this message. A language model can also make the right prediction, expect that it can make it for the wrong reason, the mere mention of *Trump* might be enough to shift the weights in the prediction and to overcome the words associated with a possible positive assessment.

The suggestion to restrict evaluation to surface features like adjectives, as in (Feder et al., 2021), oversimplifies the options available in the full appraisal toolkit, as this negative assessment comes not only through the use of negative adjectives at the very end of this message, but also through the use of other linguistic constructions, for example, the use of *behind the sunny smile* and *the same old Trump* provides a contrast with the positive assessment of *masterful*. The same applies to the negative hint in *optimistically declared*.

The simplified sentiment analysis framework from NLP also misses the fact that the object of evaluation 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 as affect, judgement or appreciation (Martin and White, 2005), see also the formal annotation framework for appraisal suitable for computational analysis (Di Bari, 2015).

Apart from sentiment predictions, the SFL tradition provides more delicate ways to describe the lexicogrammatical choices via the systems of transitivity, polarity or mood in order to explain the functions

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

behind detectable choices and, therefore, to explain the causes for decisions of blackbox language models through more relevant features.

4 Conclusions

The focus of this chapter is discuss what AI language technologies can do for Systemic Functional Linguistics and what Systemic Functional Linguistics can do for AI technologies. The starting points of both frameworks differ: SFL focuses on the functions of language and its use in society, while NLP focuses on efficient interpretation of linguistic forms on the scale of very large corpora. These differing focuses also reveal complementary weaknesses: SFL researchers can deal with a limited number of specific examples, while NLP tools lack access to the real communicative needs underlying language use.

However, these strengths and weaknesses demonstrate a fair degree of complementarity, which makes functional linguistics very relevant with the rise of Large Language Models. NLP provides large-scale predictive capabilities that enable more robust analyses, for example, examining the distribution of negation across registers. SFL, on the other hand, offers a theoretically grounded, function-oriented perspective that helps explain how NLP tools work and (more importantly) how they fail by introducing a more nuanced meta-language. By integrating the strengths of both frameworks, we can achieve deeper insights into language and its functions. Moreover, this collaboration has the potential to benefit society through the more responsible and informed deployment of AI technologies.

5 Acknowledgements

The early draft of this paper was discussed with a number of colleagues. I'm especially grateful to Marco Baroni, John Bateman, Mick O'Donnell and Rebekkah Wegener for their useful comments.

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.

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.

Caselli, T., Sprugnoli, R., and Moretti, G. (2021). Identifying communicative functions in discourse with content types. *Language resources and evaluation*, pages 1–34.

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. A. (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*.

Kuzman, T., Ljubešić, N., and Mozetič, I. (2023). ChatGPT: Beginning of an end of manual annotation? use case of automatic genre identification. *arXiv preprint arXiv:2303.03953*.

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.

Rönnqvist, S., Kyröläinen, A.-J., Myntti, A., Ginter, F., and Laippala, V. (2022). Explaining classes through stable word attributions. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1063–1074, Dublin, Ireland.

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.