# Exploring natural language principles with respect to algorithms of deep neural networks

Thesis proposal

**Tomáš Musil**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague, Czech Republic
musil@ufal.mff.cuni.cz

## Abstract

*Neural networks are the state-of-the-art method of machine learning for many problems in natural language processing (NLP). Their success in machine translation and other NLP tasks is phenomenal, but their interpretability is challenging. Inspired by the semantic properties of the word2vec vector representations of words, we want to find out how neural networks represent meaning. In order to do this, we propose to examine neural networks in NLP, research methods for their interpretation, concept of meaning in the philosophy of language and find a methodology that would enable us to connect these areas in a principled manner.*

## 1 Artificial Intelligence and Language

Language was one of the central topics of artificial intelligence (AI) research ever since Turing [1950] considered the question "Can machines think?" and proposed to replace it with the "imitation game", based purely on textual communication.

There has been a tremendous development in recent years in NLP, even though language is still one of the hardest problems in AI. Machine translation systems achieve super-human performance (at least in a competition setting) [Barrault et al., 2019]. Voice assistants are getting better and better. Some text generation models are so powerful that their authors consider them to pose a danger to society [Radford et al., 2019a].

Artificial neural networks are behind a lot of these achievements. Being machine learning (ML) models with up to billions of parameters and very little structure related to the task that they are learning, neural networks are often regarded as black boxes, and interpretation of the trained models presents a major scientific challenge [Belinkov et al., 2019].

Some questions are relatively easy to answer. These include inquiry into to what extent neural networks represent linguistic information, for which there are annotated datasets. Other problems are harder than that. How do neural machine translation (NMT) systems achieve the level of translation quality comparable to humans? How do neural networks represent meaning? Not only is there a shortage of annotated data, but the problem is also much more complicated on the theoretical side. The nature of meaning is itself a subject of debate in the philosophy of language. Together with asking the questions mentioned above, we, therefore, need to pick a suitable theory of meaning to specify them.

The philosophical aspects of these questions and the artificial essence of the objects of our research pose two methodological problems. We need to specify what are we searching for along with the search. And the objects of our research are not independent of the tools that we are using for our experiments and of the community that develops both the NLP systems and the tools of research. It is essential to keep this in mind and reflect on the research from the position of the philosophy of science as well.

This proposal is organised as follows: in Section 2, we introduce the word2vec model and its semantic properties as a motivational example for our research. In Section 3, we describe representations of language in neural networks and the various tasks for which they are used. In Section 4, we review the methods of examining neural networks. We present results obtained with these methods in Section 5. We conclude that an intuitive concept of meaning is not sufficient to interpret the results. We discuss possible theoretical approaches to meaning in Section 6. In Section 7, we sketch a possible methodology for combining the research in the theory of meaning and interpretation of neural networks in NLP in a unified framework that will enable us to learn more about language itself. In Section 8, we summarize related work, our prior work and plans for further research. Full-page figures are given as supplementary material after the bibliography.

## 2 Semantic Spaces in Neural Networks

Vector word representations sometimes have interesting semantic properties. The most well-known example was found by Mikolov et al. [2013a], who created the word2vec model. Representations from this model obey the vector arithmetic of meanings illustrated by Figure 1 and the following equation:

$$v_{king} - v_{man} + v_{woman} \approx v_{queen},$$

meaning that if we start with the word “king”, by subtracting the vector for the word “man” and adding the vector for the word “woman” we arrive at a vector that is nearest in the vector space to the one that corresponds to the word “queen”. This means that *queen* is to *woman* as *king* is to *man*.

Mikolov et al. [2013b] also trained the word2vec model with phrases, resulting in even simpler and more elegant equations, such as

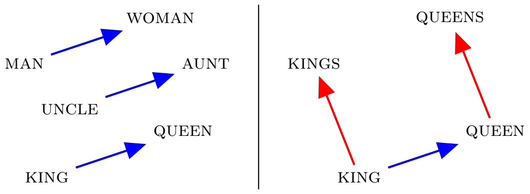


*Figure 1: Examples of semantic vector arithmetic according to Mikolov et al. [2013a].*

$$v_{Germany} + v_{capital} \approx v_{Berlin}.$$

Why does the model learn these analogies? The authors also find this question interesting:

> The model itself has no knowledge of syntax or morphology or semantics. Remarkably, training such a purely lexical model to maximize likelihood will induce word representations with striking syntactic and semantic properties. [Mikolov et al., 2013c]

Unfortunately, neither they nor anybody else (as far as we know) has published an answer. Goldberg and Levy [2014] ask:

> Why does this produce good word representations?
>
> Good question. We don’t really know.
>
> The distributional hypothesis states that words in similar contexts have similar meanings. The objective [of the model] clearly tries to increase the [dot product of the representations of the context and the word] for good word-context pairs, and decrease it for bad ones. Intuitively, this means that words that share many contexts will be similar to each other (note also that contexts sharing many words will also be similar to each other). This is, however, very hand-wavy. Can we make this intuition more precise? We’d really like to see something more formal.

Why does the word2vec model produce word representations with remarkable semantic properties? To answer the question, we need to understand both *how the model works* and *what meaning is.* The first part is the task of NLP and ML in general; the second is the task of the philosophy of language.

# 3 Representations of Language

Before we present our account of what the answers to the questions about neural networks and semantics should look like and how we propose to get closer to it, we need to introduce the representations of language that we are going to study and the tasks for which they are used.

In neural networks, language is represented by vectors of weights or activations. The vector representations of language units (documents, sentences, words, parts of words, characters), usually called *embeddings*, are mappings from a discrete and sparse space of individual units to a continuous and dense vector space.

Word embeddings are used in various applications. In this section, we describe tasks that are important for our research. Other tasks that we are interested in include text generation, question answering, sentiment analysis, summarization, and image captioning.

## 3.1 Language Models

One of the first applications that used word embeddings was a neural language model [Bengio et al., 2003]. It was a simple feed-forward neural network that predicted the next word given the $k$ preceding words.

Today, the state-of-the-art language models (e.g. GPT-2, Radford et al. [2019b]) use the Transformer architecture [Vaswani et al., 2017] with billions of parameters.

## 3.2 Pretrained Representations

Some unsupervised models are used specifically for obtaining vector representations of words. These representations are then used in other tasks where data scarcity prevents training the embeddings from scratch. The embeddings may be adapted to the task together with the rest of the network (this process is called *fine-tuning*).

One such model is the word2vec mentioned above; others include Glove [Pennington et al., 2014] and fastText Bojanowski et al. [2017]. Also interesting are the language models with the so-called context embeddings, e.g. eLMo [Peters et al., 2018] and BERT [Devlin et al., 2019]. These models can also be trained on a corpus consisting of text in many languages, creating one multilingual model that is able to produce representations independent on the input language.

## 3.3 Neural Machine Translation

Translation can be taken as a direct application of semantics if we assume there is a procedure for comparing the meaning of expressions in different languages. There are also other possibilities, such as defining meaning through a process of translation of symbols [Peirce, 1935], or describing meaning as that what is invariant in the empirical process of translation [Toury, 1980]. In any case, translation and meaning are closely related.

NMT started as a general sequence to sequence learning algorithm with a simple recurrent neural network (RNN) architecture [Sutskever et al., 2014]. Later, the attention mechanism [Bahdanau et al., 2014] was added to help the translation model with alignment. In recent years, the attention mechanism is being used without recurrent network cells in the Transformer model [Vaswani et al., 2017].

Unsupervised NMT [Lample et al., 2018] is a recent technique that makes it possible to train translation models without parallel corpora. It starts with mapping the word embedding spaces of the languages from the translation pair on each other. Then it creates a simple word-for-word translation model for each translation direction and creates a training corpus by translating monolingual data with these models. It iteratively makes the models and the corpus better, by training one of the systems on the data produced as translations by the other system. Each system is learning to translate from the synthetic data (translations) to the natural data (original monolingual corpus). As it is getting better, it produces better translations and therefore better training data for the other model, which translates in the opposite direction.

The first step of the unsupervised NMT, the mapping of the embedding spaces can be done in an unsupervised manner, based on the fact that embedding spaces have similar shape even for different languages. This leads us to believe that the embedding space is structured in a meaningful way.

Machine translation is also interesting for translation theory and philosophy of language. The previous machine translation (MT) paradigm, statistical machine translation, was explicitly based upon the idea of the noisy channel model [Weaver, 1955]. Neural machine translation is still implicitly based on the probabilistic interpretation of neural networks. It is the task for translation theory and philosophy of language to determine to what extent is the probabilistic paradigm applicable to translation or find another one if need be.

# 4 Examining Representations of Language

In this section, we present two groups of methods for investigating language representations in neural networks: probing and unsupervised methods.

Bakarov [2018] presents an overview of the methods of examining vector representations in NLP. Further information can also be found in the overview of methods for analysing deep learning models for NLP [Belinkov and Glass, 2019].

## 4.1 Probing

According to Belinkov and Glass [2019], the most common approach for examining linguistic properties in neural network components is using a classifier to predict such properties from activations of the neural network. We refer to this approach as "probing".

Probing is a supervised method, so using it requires data annotated for the studied property. This means that with the help of probing, we can only reveal in representations the kind of information that we have previously decided to look for in them and we have an annotated dataset for them.

This introduces a systemic bias into the research. It is easier to probe for properties that are already described in formal frameworks with large annotated datasets. The results that find these properties in the representations then retroactively affirm the correctness of the formal frameworks.

## 4.2 Unsupervised Methods

The goal of unsupervised methods is first to find what plays an important role and only then label it. One example would be clustering, where we first find the clusters in the space and then try to assign labels to them. Unsupervised methods avoid the bias inherent in the choice of the information to classify in probing. The cost is that the results are usually harder to quantify.

To illustrate this with an example: it is possible to show that we can partition the Czech embedding space from NMT encoder into parts that correspond to part of speech (POS) [Musil, 2019]. Does this mean that information about POS is important for the NMT system? Maybe it merely shows that we can divide the high dimensional space by criteria that are mostly arbitrary? We can make stronger claims with the unsupervised approach, by showing that in the first few dimensions of principal component analysis, the embeddings naturally form clusters that correspond to POS.

paragraphClustering is an unsupervised method that serves to separate the data into classes. Although hierarchical clustering can give us more information than just the classes themselves, it may be too coarse – words have multiple features and assigning them to just one class or one place in the hierarchy is often too much of a simplification.

**Principal component analysis (PCA)** is a process commonly used to decorrelate data or transform data into a lower-dimension space, for example, to visualize it. In the context of word representations, PCA was used, for example, to identify a subspace containing gender information and then modify representations to remove gender bias (debiasing, Bolukbasi et al. [2016]).

**Independent component analysis (ICA)** [Jutten and Herault, 1991, Comon, 1994, Hyvärinen and Oja, 2000] is a similar method that looks for components that have as little Gaussian data distribution as possible. It was used, for example, to extract features from distribution representations of the words [Honkela et al., 2010].

# 5 Properties of Word Embeddings

Because we are interested in representations of meaning, we will restrict this section to embeddings of words. The space of sentences is too large and sparse for effective computations. Many applications use smaller units than words, but a substantial portion of them only has meaning in composition with others.

The rest of this Section is arranged as follows: Section 5.1 is a survey of related work; Sections 5.2–5.5 presents our contributions.

## 5.1 Related Work

Most of the research of interpretation of word embeddings consists of probing for morphology and syntax:

Chen et al. [2013] have designed several classifier tasks to help us better understand the information encoded in the word embeddings, one of which is sentiment polarity.

Köhn [2015] shows for several languages that from the representation of a word, we can predict its syntactic properties, such as POS, gender, case, or number.

Qian et al. [2016] examined the properties (including POS and sentiment) of embeddings from 3 different architectures of language models for more than 20 languages, including Czech. They use a multi-layer perceptron classifier trained on parts of the dictionary and evaluated on the rest of the dictionary.

Belinkov et al. [2017a,b] have found out what NMT models learn about morphology and semantics by training POS, morphological and semantic taggers on representations from various models.

Saphra and Lopez [2018] showed that language models learn POS first. The role of the language model is to predict the next word, so the representation of the POS is a side effect, but it becomes apparent as the learning progresses, even before the model can successfully estimate the probabilities of the next word. The authors use the SVCCA (Singular Vector Canonical Correlation Analysis) method to demonstrate that different aspects of the language structure are learned at different rates, with information on POS acquired at the beginning of learning.

Hewitt and Manning [2019] probed that representations in machine translation encode syntactical properties.

Vylomova et al. [2017] have shown that if representations from the neural machine translation encoder are similar to each other, then the words they represent are semantically and morphologically similar. What distinguishes this procedure from classical probing is that it does not train the classifier, which is replaced by an external annotation of semantic (or morphological) similarity.

There is less work about representation of meaning in word embeddings:

Gupta et al. [2015] show that representations obtained by the Mikolov et al. [2013a] method contain information about the reference properties of words, for example, it is possible to predict the country in which a city is located from the embedding of the city's name.

Hollis and Westbury [2016] have shown that principal components of word2vec embeddings correlate with various psycholinguistic properties.

Word embeddings are clustered according to meaning [Liu et al., 2018] in t-SNE [Maaten and Hinton, 2008].

## 5.2 Examining Word Embeddings with PCA

Our research Musil [2019] shows that with the help of PCA, we can show that a neural translation model divides Czech words into POS classes. It also distinguishes between proper names and general nouns. The structure of representation varies between the encoder and the decoder of the NMT system.

The structure of the representation of the same data in the word2vec model is different, for example, in that it distinguishes infinitive forms of verbs or modal verbs. A completely different structure is found in the space of representations of words in the neural model for sentiment analysis. All of these facts can be shown without annotated data and thus without deciding beforehand what we will look for in the space of representations. For this reason, we find our results more convincing than if they had been obtained through probing.

Inspired by Hollis and Westbury [2016], we compare the structure of Czech word embeddings for English-Czech NMT, word2vec and sentiment analysis. We show [Musil, 2019] that although it is possible to successfully predict POS tags from word embeddings of word2vec and various translation models, not all of the embedding spaces show the same structure. The information about POS is present in word2vec embeddings, but the high degree of organization by POS in the NMT decoder suggests that this information is more important for machine translation and therefore the NMT model represents it more directly. Our method is based on correlation of PCA dimensions with categorical linguistic data. Figure S.1 (on page 17) shows correlations of POS information with PCA components for NMT encoder, NMT decoder and word2vec embeddings. We also show that further examining histograms of classes along the principal component is important to understand the structure of representation of information in embeddings (see Figure S.3).

### 5.3 Sentiment Analysis

We have found [Musil, 2019] that the shape of the space of word embeddings for a model trained for sentiment analysis is triangular. In Figure S.2 (on page 18), we see a sample of the words plotted along the first two principal components. The first component represents the polarity of the words (good/bad); the second component represents intensity (strong/neutral). The triangular shape may be explained by the fact that words that are far from the centre on the polarity axis are never of low intensity.

### 5.4 Clustering Word Derivations

Derivation is a type of a word-formation process which creates new words from existing ones by adding, changing or deleting affixes.

To examine derivational relations in word embeddings, we used DeriNet [Kyjánek, 2018], a Czech lexical network, which organizes almost one million Czech lemmata into derivational trees. For each such pair, we compute the difference of the embeddings of the two words and perform unsupervised clustering of the resulting vectors. Our results [Musil et al., 2019] show that these clusters mostly match manually annotated semantic categories of the derivational relations (e.g. the relation 'bake–baker' belongs to the category 'actor', and a correct clustering puts it into the same cluster as 'govern–governor'). See Figure S.4.

### 5.5 Preliminary ICA Results

We have used NMT and word2vec embeddings from models trained on *fiction* part of the Czech side of the Czeng corpus Bojar et al. [2016].

We look at the words that are strongest in each ICA component. We have found that the components represent various types of categories. We list a few of them with examples here:

**Semantic category:** words with similar semantic content (e.g. law and justice) from various syntactic categories: *zákona Unie členských zákon stanoví Komise zákony soud zákonů zákonem Evropské práva práv ustanovení nařízení porušení soudu tj souladu podmínek*

**Semantic and syntactic category:** words that are defined both semantically and syntactically, in this case, verbs associated with *going somewhere* in the past tense and masculine gender: *šel zašel zajít jít spěchal šla zavedl vešel dopravit nešel vrátil poslal vydal šli poslat přišel odjel přijel jel dorazil*

**Syntactic subcategory:** words with specific syntactic features, but semanticaly diverse (in this case, adjectives in feminine singular form): *Velká moudrá občanská dlouhá slabá čestná železná překrásná hladká určitá marná tmavá hrubá příjemná bezpečná měkká svatá nutná volná zajímavá*

**Feature across POS categories:** e.g. feminine plural form for adjectives, pronouns and verbs: tyto tyhle neměly byly mohly začaly vynořily zmizely měly objevily všechny vypadaly nebyly zdály změnily staly takové podobné jiné tytéž

**Stylistic:** in this case non-standard forms: *máš bys tý nemáš seš ses víš Hele kterej sis jseš bejt vo svýho celej děláš chceš teda každej velkej*

We are finding that many ICA components represent various features of words. It seems to classify not only morphology and syntax, but also semantics, so it is a promising research direction for inquiries about representations of meaning.

## 6 NLP and Philosophy of Language

Towards the end of Sections 5.1 and 5.5, we talked about meaning. What is meant by *meaning* is part of the problem that we aim to solve. There is no agreed-upon general definition of 'meaning' (or 'sense', 'semantics', ...), as, for example, Stokhof [2013] explains:

> Usually, [the theoretical and conceptual diversity in formal semantics] is not regarded as particularly problematic, and is often explained by pointing out that they are merely different ways of addressing the phenomena that semanticists are interested in. Be that as it may, what does seem puzzling to us is that there is no firm consensus on what constitutes a proper conceptualisation of the core phenomena. Thus we find meaning described in terms of truth-conditions (intensionally or extensionally conceived), as constituted by assertability conditions, characterised in terms of update conditions or context-change potentials, analysed in terms of inference potential, and so on. And then there is the added dimension of speaker's meaning and conversational implicature, and the concomitant discussions about the dependence between such notions and literal meaning (if such is acknowledged as a bona fide entity to begin with).

To be able to talk about representations of meaning, we will have to review different conceptualizations of meaning and find one that is useful for describing the phenomena we encounter when we examine how neural networks work in NLP.

There is almost no related work that would connect NLP with the philosophy of language. Honkela [2007] links neural language models, self-organizing maps and Quine's semantic holism.

## 6.1 The Distributional Hypothesis

If neural language models or pretrained embeddings represent meaning at all, it must be derived from the training corpus. The language model does not have access to any information besides the corpus, which is only seen through a sliding window of tokens. This may be the reason behind the popularity of the distributional hypothesis in neural language model literature. The famous saying by Firth Firth [1957], "You shall know a word by the company it keeps!", is quoted in almost every paper concerned with vector space models of language.

The general distributional hypothesis states that the meaning of a word is given by the contexts in which it occurs. It is, however, worth noticing that in Firth's theory, collocation is just one among multiple levels of meaning, and his text does not support the idea of meaning based on context alone.

## 6.2 The *Use Theory* of Meaning

The *use theory* of meaning can be summed up as "the meaning of a word is its use in the language" [Wittgenstein, 1953, § 43]. It is associated with late Wittgenstein's concept of language game. In *Philosophical Investigations* [1953, §§ 499–500], he writes:

> To say "This combination of words makes no sense" excludes it from the sphere of language and thereby bounds the domain of language. [...] When a sentence is called senseless, it is not as it were its sense that is senseless. But a combination of words is being excluded from the language, withdrawn from circulation.

This "bounding of the domain of language" is precisely what language model does; therefore, the use theory may be one way to connect language modelling and semantics.

That "knowledge of language emerges from language use" is also one of the main hypotheses of cognitive linguistics [Croft and Cruse, 2004].

## 6.3 Structuralism

In structuralism, the meaning of a word is given by its relation to the other words of the language [de Saussure, 1916]. This holds for word representations in artificial neural networks as well. The vectors representing the words do not have any other meaning than their position among the rest of the vectors, and a single vector does not have any significance outside the model. This is also demonstrated by the vectors being different every time the model is trained because of random initialization.

## 6.4 Semantic Holism and Atomism

Word representations obtained from the word2vec model exhibit interesting semantic properties. This is usually explained by referring to the general distributional hypothesis. We propose [Musil, 2020] a more specific approach based on Frege's holistic and functional approach to meaning.

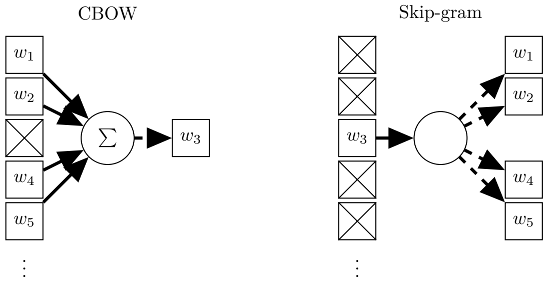


*Figure 2: CBOW and Skip-gram language models according to Mikolov et al. [2013a].*

There are two variants of the word2vec model [Mikolov et al., 2013a]. The CBOW variant predicts a missing word based on the context; the Skip-gram variant predicts context words based on a single word (see Figure 2). The Skip-gram variant performs better in analogy tasks [Mikolov et al., 2013c]. We show that the training process the Skip-gram variant of word2vec is analogous to a holistic definition of meaning.

*Semantic holism* (or *meaning holism*) is "the thesis that what a linguistic expression means depends on its relations to many or all other expressions within the same totality. [...] The totality in question may be the language to which the expressions belong or a theory formulation in that language." [Fodor and Lepore, 1992] The opposing view is called semantic atomism, and it claims that there are expressions (typically words), whose meaning does not depend on the meaning of other expressions. The meaning of these expressions is given by something outside language (e.g. their relation to physical or mental objects).

Taking Tugendhat's formal reinterpretation of Frege's work [Tugendhat, 1970] as a starting point, we demonstrate that it is analogical to the process of training the Skip-gram model and it offers a possible explanation of its semantic properties. Tugendhat's definition of meaning as truth-value potential is:

> [T]wo expressions $\varphi$ and $\psi$ have the same truth-value potential if and only if, whenever each is completed by the same expression to form a sentence, the two sentences have the same truth-value.

This definition has one crucial aspect in common with the Skip-gram version of the word2vec model: while we examine the meaning of an expression, the expression is fixed, and the context is changing for comparison. Therefore, it presupposes the context as the source of meaning, in the same way, that Skip-gram learns the representation of a word from the representation of the context. The fact that the holistic Skip-gram version of word2vec works better in analogy tasks than the complementary atomistic CBOW version supports the holistic approach to meaning [Musil, 2020].

## 6.5 Objectivism, Subjectivism and Experientialism

Study of metaphor and its connection to experience led Lakoff and Johnson [1980] to criticise both the objectivist and subjectivist approaches to language. Melby [1994] applies this critique to MT and says that "most work in machine translation is explicitly or implicitly based on [the objectivist framework]." He lists the following beliefs as characteristic for objectivism:

1. Words and expressions are mapped to senses.
2. Each sense exists independently and has the properties of mathematical sets.
3. The meaning of a sentence can be obtained by combining the word senses from the bottom up.

Although this may seem similar to semantic atomism, the "myth of objectivism" is a broader view. In the original formulation Lakoff and Johnson [1980], it talks about epistemology and ontology in a way that is not necessarily implied by semantic atomism. To overcome the myth, we need to accept experience as the source of knowledge.

Melby [1994] claimed that contemporary techniques of machine translation will never be extended to handle general language texts and that entirely new techniques that avoid the assumptions of objectivism will be needed; the systems need to understand dynamic metaphor and exhibit flexibility in handling new situations. If Lakoff and Johnson's theory of metaphor holds, this is a trivial consequence: since understanding metaphor is based on experience and contemporary translation systems do not experience anything, they cannot understand and translate metaphors.

More than 25 years later, NMT is based on principles that can hardly be construed as an extension of the old techniques. They are much more flexible and in general, produce significantly better translations. Do neural networks somehow evade the pitfalls of objectivism? Maybe going repeatedly through the enormous quantity of textual data constitutes a kind of experience; perhaps it is possible to extract the experience of others from the data? May that be one of the reasons for their sudden success in MT and other NLP applications?

# 7 AI and Philosophy of Science

In the previous sections, we examined NLP applications and came to the conclusion that we should interpret the principles of their operation from the point of view of the philosophy of language. What kind of scientific methodology should we apply in this case?

When we investigate NLP (and artificial intelligence in general), we are not dealing with objects that are independent of the community of researchers as does biology or physics. Furthermore, we are using machine learning methods to run experiments on the results of other machine learning methods.

The question of how to incorporate results of machine learning into the scientific workflow is starting to come up in other sciences as well, e.g. biology [Smyčka, 2020].

In this section, we sketch one possible approach to this problem. Because we are talking about *artificial intelligence* (in a broad sense), we will start by distinguishing between *natural* and *artificial.* This distinction is going to be useful in Section 7.2, where we discuss current NLP practice from the point of view of the philosophy of science.

## 7.1 Natural and Artificial

According to Romportl [2015], *natural* is "that which defies being captured by language". It is associated with organic growth and the Greek concept of φύσις (*physis*). *Artificial* is "that whose essence is fully determined by language". It is associated with human reason, rationality, (interpretable) structure and the Greek concept of λόγος (*logos*).

Most things have both of these aspects. For example:

> Let's imagine an old rustic wooden table. What is artificial about it? That which we can grasp with words: shape and size of its geometrical idealisation, its weight, colour tone, purpose, or perhaps a description of the way it was made by a carpenter with an axe, a saw and a jack plane. However, we cannot describe how exactly it looks, how it feels when being touched, the exact look of its texture and wood structure, its smell. [Romportl, 2015]

The undescribable remainder is the *natural.*

In this sense, emergent strong AI would be partly natural because we would not control every particular aspect of its creation. There are two strong arguments for the *naturalness* of NLP applications as well: solving tasks like *natural language understanding* may be equivalent to creating general strong AI. Also, language itself lies between the natural and the artificial:

> Language in general is a long bridge between *physis* and *logos*, with deixis and protolanguages close to the bank of *physis*, formal languages, mathematics, geometry etc. close to the bank of *logos*, and natural language somewhere in between, where human minds operate. [Romportl, 2015]

Can we say that contemporary artificial neural network models applied to NLP are already *natural* in this sense? We think that we can. The fact that there are workshops such as BlackBox NLP dedicated to interpreting neural networks in NLP proves that there are aspects of their behaviour that we do not yet understand. The general conclusion was also stated by Romportl [2015]: "[W]e should seriously start to think how to live with the natural machine intelligence that has already started to emerge on top of our technological artefacts."

## 7.2 Science and NLP

We can use the distinction between *natural* and *artificial* to categorise various scientific methods. Some scientific disciplines study objects that are clearly artificial (e.g. literary studies), others are dedicated to clearly natural fields (physics as a classic example of a science, biology, chemistry, ...). A different situation arises in mathematics because mathematicians are not working with something given by nature or created by other people. They are creating and developing a rational (and therefore *artificial*) structure of objects that are purely conceptual (whatever that means—even the most stubborn supporter of mathematical platonism would concede that numbers exist in a different way than rocks, plants and animals). This self-referential process makes its *artificialness* even more apparent.

NLP (at least in its current, machine-learning-driven form) is based on mathematics. However, as we have demonstrated in Section 7.1, it is also becoming *natural*.

The *natural* is examined by natural (or empirical) sciences. We can illustrate the process of research in natural sciences by the diagram in Figure S.5 on page 21. The goal of science is to develop a theory about reality. We observe reality through the phenomena that we experience. Based on our current theories, we design instruments and experiments to measure the observed phenomena. The results of these measurements and experiments inform the development of our theories, completing the circle. One concrete example of this scientific approach would be cosmology, illustrated by the diagram in Figure S.5.

The situation in (contemporary) NLP is different because the instruments that we are developing are not just a mean to understanding an independent reality; they are the goal itself. The theory of NLP is about the instruments of NLP. We illustrate this by the diagram in Figure S.6.

The architectures of neural networks are often developed independently of the tasks for which they will eventually be used. Examples of this include the same neural network architecture being used for all text-related tasks, including MT, and even applications unrelated to their original purpose such as game playing [Upadhyay et al., 2019]. This universality of the instruments is one of the factors enabling the theory to focus on the instruments themselves and ignore what was in the previous figure labelled as "reality".

To examine the *natural* quality that emerges in using deep neural networks in NLP, we need to take a step back and look at the NLP research as a whole, concentrating on the relation between language itself and the neural networks (Figure S.7).

What should the theory be in this new picture? We believe that linguistics, with its descriptivist mode (and its objectivist syntax/semantics/pragmatics divide Melby [1994]) is not the best candidate. It should be the philosophy of language, asking questions such as "What is *meaning*?". Asking this kind of questions is vital to understand the relationship between language and current technology.

We propose to attack the problem from two sides:

- from the perspective of the *natural*, we will examine the structure of data, representations and the algorithms that produce them,
- from the perspective of the *artificial*, we will find a theory of language that would lead to the same structures.

Examples of the first part of our research are given in Section 5, examples of the second type are given in Section 6.

# 8 Conclusion and Future Work

Interpretability is an important challenge for neural networks in NLP. There is a limited amount of findings about linguistic phenomena that we are able to predict from embeddings. Much less is known about the general properties of the embedding space and about its semantic properties. As NLP technologies—such as MT—are becoming competitive with humans, we should be able to learn something about language itself by studying the way these technologies work. Finding a plausible explanation of state-of-the-art technology from the point of view of the philosophy of language would contribute to the theory of meaning.

We have contributed to the knowledge in this field by examining and comparing embeddings from sentiment analysis, NMT and word2vec [Musil, 2019]. We have demonstrated that word embeddings capture information about semantic classes of word derivations [Musil et al., 2019]. We examined the relationship between the word2vec model and semantic holism [Musil, 2020].

Future work includes more examining of the embedding space with component analysis, notably ICA, extending the research of word derivations to more languages and completing the methodological reflections sketched in Section 7.

# 9 Bibliography

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Amir Bakarov. A Survey of Word Embeddings Evaluation Methods. *arXiv:1801.09536 [cs]*, January 2018. URL `http://arxiv.org/abs/1801.09536`. arXiv: 1801.09536.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, August 2019. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W19-5301`.

Yonatan Belinkov and James Glass. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, March 2019. doi: 10.1162/tacl_a_00254. URL `https://www.aclweb.org/anthology/Q19-1004`.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. What do Neural Machine Translation Models Learn about Morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada, July 2017a. Association for Computational Linguistics. doi: 10.18653/v1/P17-1080. URL `https://www.aclweb.org/anthology/P17-1080`.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Taipei, Taiwan, November 2017b. Asian Federation of Natural Language Processing. URL `https://www.aclweb.org/anthology/I17-1001`.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. On the Linguistic Representational Power of Neural Machine Translation Models. *arXiv:1911.00317 [cs]*, November 2019. URL `http://arxiv.org/abs/1911.00317`. arXiv: 1911.00317.

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

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. URL `http://arxiv.org/abs/1607.04606`. arXiv: 1607.04606.

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered. In Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala, editors, *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, Lecture Notes in Artificial Intelligence, pages 231–238, Cham / Heidelberg / New York / Dordrecht / London, 2016. Springer International Publishing. ISBN 978-3-319-45509-9.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv:1607.06520 [cs, stat]*, July 2016. URL `http://arxiv.org/abs/1607.06520`. arXiv: 1607.06520.

Yanqing Chen, Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. The Expressive Power of Word Embeddings. *arXiv:1301.3226 [cs, stat]*, May 2013. URL `http://arxiv.org/abs/1301.3226`. arXiv: 1301.3226.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

William Croft and D. Alan Cruse. *Cognitive Linguistics*. Cambridge University Press, 1 edition, January 2004. ISBN 978-0-521-66114-0 978-0-521-66770-8 978-0-511-80386-4. doi: 10.1017/CBO9780511803864. URL `https://www.cambridge.org/core/product/identifier/9780511803864/type/book`.

Ferdinand de Saussure. *Course in General Linguistics*. Duckworth, London, 1916.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

John R Firth. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*, 1957.

Jerry A Fodor and Ernest Lepore. *Holism: A shopper's guide*. Blackwell, 1992.

Yoav Goldberg and Omer Levy. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. Distributional vectors encode referential attributes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 12–21, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1002. URL `https://www.aclweb.org/anthology/D15-1002`.

John Hewitt and Christopher D Manning. A Structural Probe for Finding Syntax in Word Representations. page 10, 2019.

Geoff Hollis and Chris Westbury. The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic bulletin & review*, 23(6):1744–1756, 2016.

Timo Honkela. Philosophical aspects of neural, probabilistic and fuzzy modeling of language use and translation. In *Neural Networks, 2007. IJCNN 2007. International Joint Conference on*, pages \mbox2881–2886. IEEE, 2007.

Timo Honkela, Aapo Hyvärinen, and Jaakko J. Väyrynen. WordICA—emergence of linguistic representations for words by independent component analysis. *Natural Language Engineering*, 16(3): 277–308, July 2010. ISSN 1351-3249, 1469-8110. doi: 10.1017/S1351324910000057. URL `https://www.cambridge.org/core/product/identifier/S1351324910000057/type/journal_article`.

A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, June 2000. ISSN 08936080. doi: 10.1016/S0893-6080(00)00026-5. URL `https://linkinghub.elsevier.com/retrieve/pii/S0893608000000265`.

Christian Jutten and Jeanny Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal processing*, 24(1):1–10, 1991.

Lukáš Kyjánek. Morphological Resources of Derivational Word-Formation Relations. Technical Report TR-2018-61, Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, Czech Republic, 2018.

Arne Köhn. What's in an Embedding? Analyzing Word Embeddings through Multilingual Evaluation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1246. URL `https://www.aclweb.org/anthology/D15-1246`.

George Lakoff and Mark Johnson. *Metaphors we live by.* University of Chicago press, 1980.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised Machine Translation Using Monolingual Corpora Only. page 14, 2018.

Shusen Liu, Peer-Timo Bremer, Jayaraman J. Thiagarajan, Vivek Srikumar, Bei Wang, Yarden Livnat, and Valerio Pascucci. Visual Exploration of Semantic Relationships in Neural Word Embeddings. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):553–562, January 2018. ISSN 2160-9306. doi: 10.1109/TVCG.2017.2745141.

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

Alan Melby. Machine translation and philosophy of language. 1994.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781, 2013a. URL `http://arxiv.org/abs/1301.3781`.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality, 2013b.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June 2013c. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/N13-1090`.

Tomáš Musil. Examining Structure of Word Embeddings with PCA. In Kamil Ekštein, editor, *Text, Speech, and Dialogue*, pages 211–223, Cham, 2019. Springer International Publishing. ISBN 978-3-030-27947-9. doi: 10.1007/978-3-030-27947-9_18.

Tomáš Musil. Semantic Holism and Word Representations in Artificial Neural Networks. In *Submitted to AISB 2020.*, 2020.

Tomáš Musil, Jonáš Vidra, and David Mareček. Derivational Morphological Relations in Word Embeddings. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 173–180, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4818.

Charles Sanders Peirce. *Collected papers of Charles Sanders Peirce.* Harvard University Press, 1935.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL `https://www.aclweb.org/anthology/D14-1162`.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, 2018.

Peng Qian, Xipeng Qiu, and Xuanjing Huang. Investigating Language Universal and Specific Properties in Word Embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1478–1488, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1140. URL `https://www.aclweb.org/anthology/P16-1140`.

Alec Radford, Jeffrey Wu, Dario Amodei, Jack Clark, Amanda Askell, Miles Brundage, and Ilya Sutskever. Better Language Models and Their Implications, February 2019a. URL `https://openai.com/blog/better-language-models/`.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):9, 2019b.

Jan Romportl. *Naturalness of Artificial Intelligence*, pages 211–216. Springer International Publishing, Cham, 2015. ISBN 978-3-319-09668-1. doi: 10.1007/978-3-319-09668-1_16. URL `https://doi.org/10.1007/978-3-319-09668-1_16`.

Naomi Saphra and Adam Lopez. Language Models Learn POS First. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 328–330, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5438. URL `https://www.aclweb.org/anthology/W18-5438`.

Jan Smyčka. Diverzifikace evropských horských rostlin – o pravidlech, výjimkách a předpovědích do budoucna, 2020. URL `http://www.cts.cuni.cz/index.php?m=43&akce=1535&lang=cs`.

Martin Stokhof. Formal semantics and Wittgenstein: An alternative? *The Monist*, 96(2):205–231, 2013. doi: 10.5840/monist20139629. URL `http://dx.doi.org/10.5840/monist20139629`.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to Sequence Learning with Neural Networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014. URL `http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf`.

Gideon Toury. In search of a translation theory. *Tel Aviv: The Porter Institute for Poetics and Semiotics*, 1980.

Ernst Tugendhat. The meaning of 'Bedeutung' in Frege. *Analysis*, 30(6):177–189, 1970.

Alan M. Turing. Computing Machinery and Intelligence. *Mind*, LIX(236):433–460, October 1950. ISSN 1460-2113, 0026-4423. doi: 10.1093/mind/LIX.236.433. URL `https://academic.oup.com/mind/article/LIX/236/433/986238`.

Uddeshya Upadhyay, Nikunj Shah, Sucheta Ravikanti, and Mayanka Medhe. Transformer Based Reinforcement Learning For Games. *ArXiv*, abs/1912.03918, 2019.

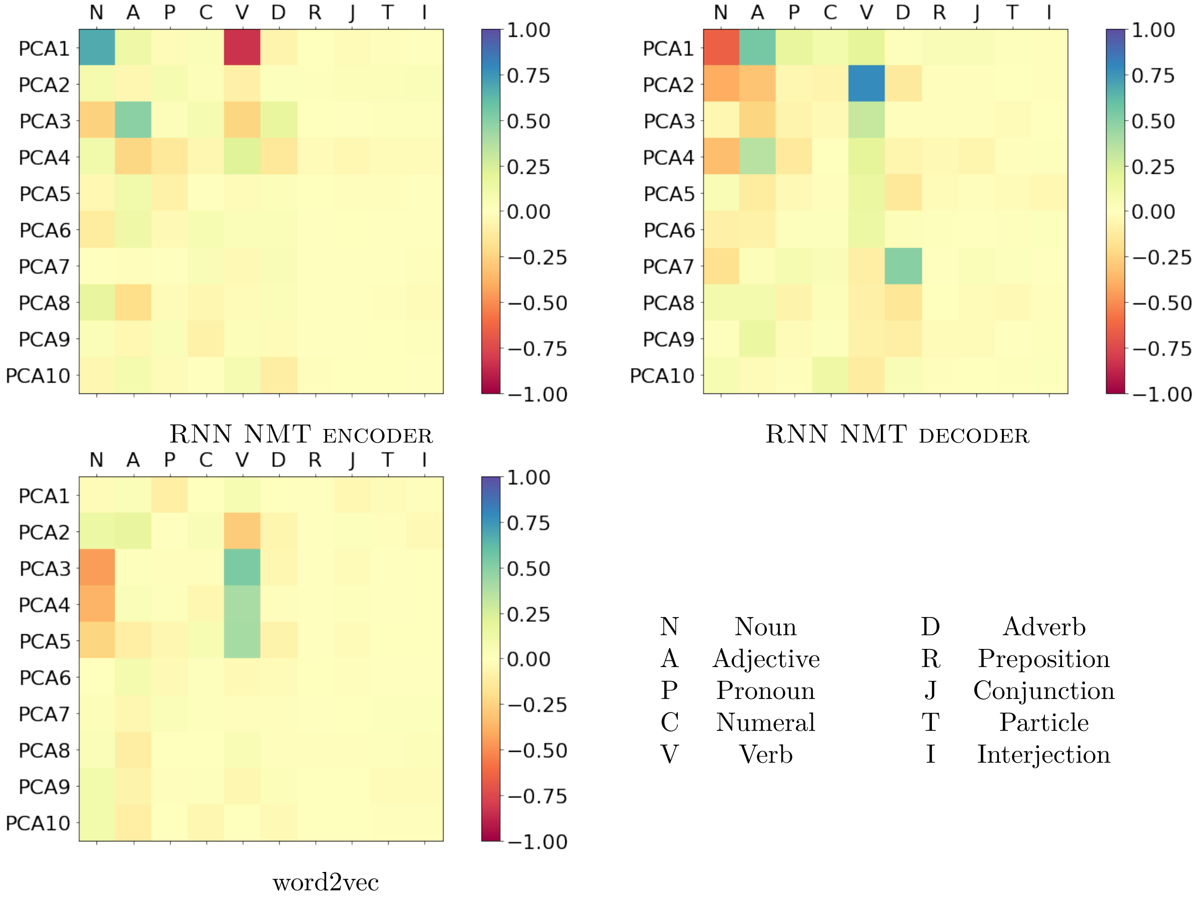
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, \Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Ekaterina Vylomova, Trevor Cohn, Xuanli He, and Gholamreza Haffari. Word Representation Models for Morphologically Rich Languages in Neural Machine Translation. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 103–108, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4115. URL `https://www.aclweb.org/anthology/W17-4115`.

Warren Weaver. Translation. *Machine translation of languages*, 14:15–23, 1955.

Ludwig Wittgenstein. *Philosophical Investigations*. Wiley-Blackwell, 1953.

# 10 Supplementary Material



RNN NMT encoder

RNN NMT decoder

word2vec

| | | | |
|---|---|---|---|
| N | Noun | D | Adverb |
| A | Adjective | R | Preposition |
| P | Pronoun | J | Conjunction |
| C | Numeral | T | Particle |
| V | Verb | I | Interjection |

*Figure S.1: Correlations of POS and PCA dimensions from the encoder of the Czech-English RNN NMT model (top left), the decoder of the English-Czech RNN NMT model (top right) and the word2vec model (bottom). The direction of the PCA dimensions is arbitrary, so the sign of the correlation is not important in itself, only if there are values with opposite signs in the same row we know that they are negatively correlated.*

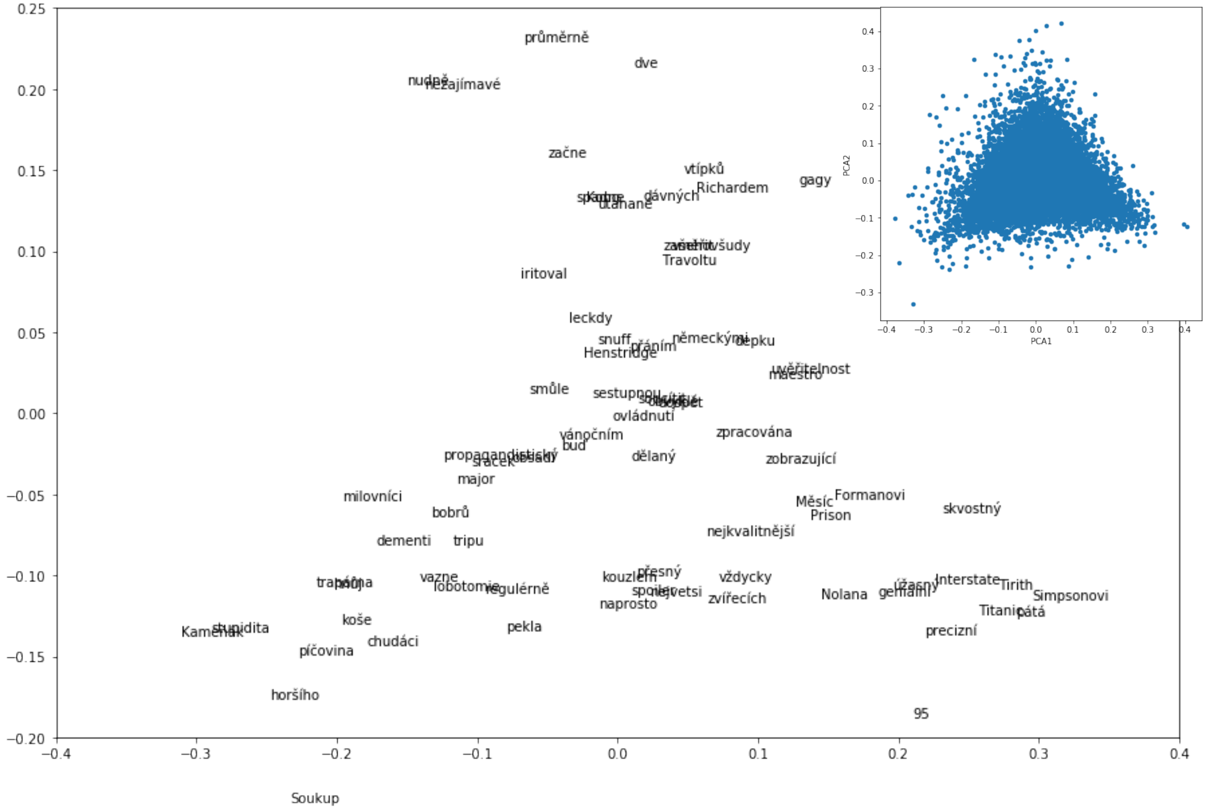


*Figure S.2: A random sample of words from the distribution of the embeddings from the sentiment analysis CNN model along the first (horizontal) and second (vertical) PCA dimension. The top right subplot shows the complete distribution.*

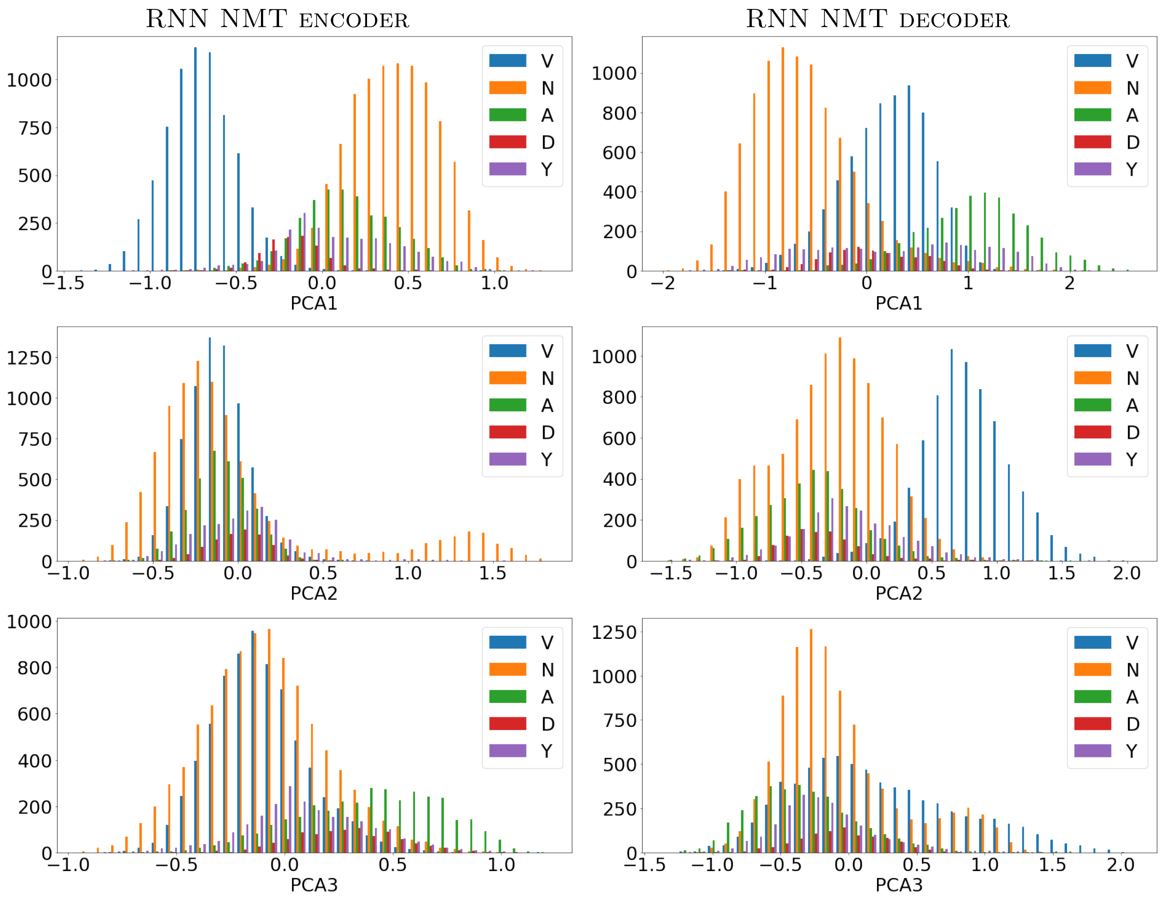
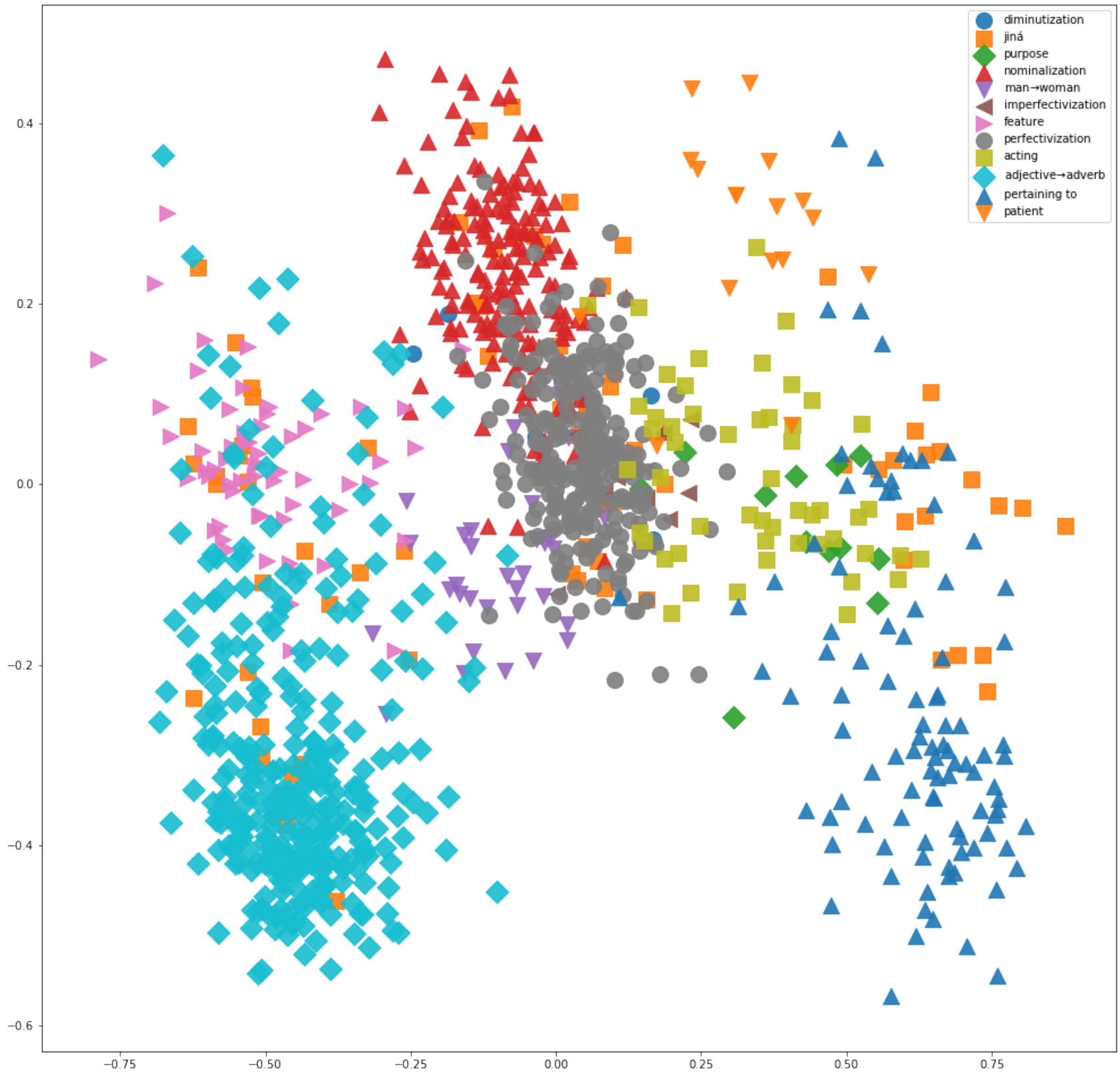


*Figure S.3: Histograms of the four largest POS classes along the first three PCA dimensions of the embeddings from the NMT* RNN MODEL*. The Czech-English* RNN NMT ENCODER *is on the left and the English-Czech* RNN NMT DECODER *on the right. V = verbs, N = nouns, A = adjectives, D = adverbs, Y = other.*

e.g. kompenzovat – kompenzace (compensate – compensation)



e.g. luxus – luxusní (luxury – luxurious)

e.g. filosofie – filosofický (philosophy – philosophical)

*Figure S.4: Clusters of the derivation types.*

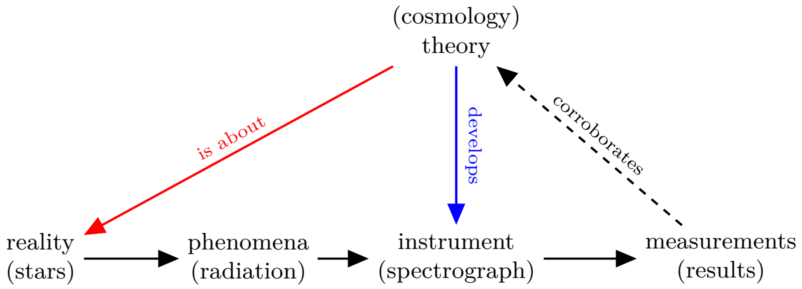


*Figure S.5: Methodology: Cosmology*

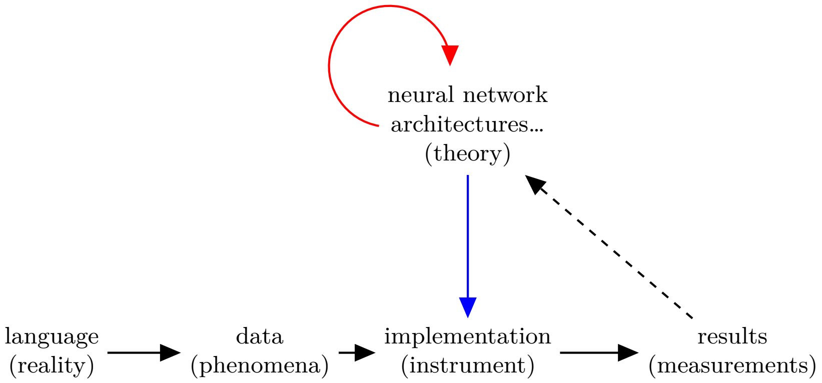


*Figure S.6: Methodology: NLP*

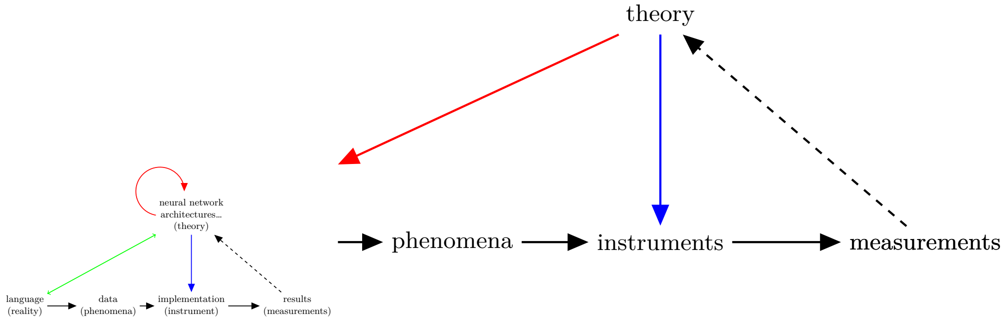


*Figure S.7: Methodology: Solution*