The Calculus of Language

Explicit Representation of Emergent Linguistic Structure Through Type-Theoretical Paradigms

Juan Luis Gastaldi Luc Pellissier

Abstract

The recent success of deep neural network techniques in natural language processing rely heavily on the so-called distributional hypothesis. We suggest that the latter can be understood as a simplified version of the classic *structuralist hypothesis*, at the core of a program aiming at reconstructing grammatical structures from first principles and analysis of corpora. Then, we propose to reinterpret the structuralist program with insights from proof theory, especially associating paradigmatic relations and units with formal types defined through an appropriate notion of interaction. In this way, we intend to build original conceptual bridges between computational logic and classic structuralism, which can contribute to understanding the recent advances in NLP.

Keywords Natural Language Processing · Structuralism · Distributional Hypothesis · Structuralist Hypothesis · Paradigm Derivation · Computational Logic

1 The Triumph of Distributionalism

The past decade has witnessed the success of deep neural networks (DNNs) in the most diverse domains of our cultures. However, if that success is usually acknowledged from a technical and societal perspective, its scientific and epistemological import is more difficult to assess. The latter is no less real notwithstanding, both in the fields with which artificial intelligence (AI) is directly concerned (such as computer science, data analysis, mathematics or engineering) and in those to which it can be directly or indirectly applied.

The debate in this regard has been mostly governed by the revived alternative between connectionist and symbolic approaches. Significantly, both perspectives covet the same battlefield of the "human mind" as the object and the source of epistemological enquiry, thus centering the discussion around the

Originally published in *Interdisciplinary Science Reviews* 46.4 (2021), pp. 569–590. DOI: 10.1080/03080188.2021.1890484.. Please cite the original version.
This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 839730.
File version: 1.0.2023.02.16.

validity of DNNs as models of intelligence or human cognition[1]. However, if we take a closer look at the recent developments of AI in the specific yet central field of natural language processing (NLP), it appears that the success of DNN techniques is not the consequence of computers becoming more "intelligent" in any sense other than metaphorical. More precisely, the remarkable results exhibited by AI in NLP *do not find their source in any successful attempt to explicitly model human faculties, competences or behaviors.* Instead, those results are to be attributed to the capacity of a family of algorithms implementing different DNN models to solve a series of tasks associated with the properties of natural language—such as machine translation, question answering, sentiment analysis or summarization—by processing ever-increasing amounts of linguistic data.

Interestingly, the performance increase for the treatment of a given task has been commonly brought about by the substitution of one network architecture by another. Yet, those models differ by significant features—customarily named, still owing to a metaphorical perspective, after cognitive faculties such as "perception" (eg. MLP), "memory" (eg. LSTM) or "attention" (eg. Transformer). This variety of architectures prevents us from attributing to any one of them a decisive epistemic capacity with respect to general linguistic phenomena. However, devoid of the specifics by which each algorithm organizes the internal representation of the input data, DNN models can only be characterized through a high level strategy consisting in approximating a function through successive layers of distributed representations of a given input, which can compute the expected output for a given task. Unsurprisingly, another cognitive metaphor accompanies this characteristic mechanism of DNNs, that of "learning", which remains as insufficient as the others to explain the efficacy of such models in the treatment of natural language.

Now, if we take our eyes off their strictly technical aspects and the metaphors that usually surround their epistemic claims, it is possible to see that all those models, insofar as they take natural language as their object, share a unique *theoretical* perspective, known as the *distributional hypothesis*. Simply put, this principle maintains that the meaning of a word is determined by, or at least strongly correlated with, the multiple (linguistic) contexts in which that word occurs (its "distribution")[2].

As such, a distributional approach is at odds with the generative perspective that dominated linguistic research during the second half of the 20$^{\text{th}}$ century. Indeed, the latter intends to account for linguistic phenomena by modeling linguistic competence of cognitive agents, the source of which is thought to reside in an innate grammatical structure. In such a framework, the analysis of distributional properties in linguistic corpora can only play a marginal role, if any, for the study of language[3]. By referring the properties of linguistic units to intralinguistic relations, as manifested by the record of collective linguistic performance in a corpus, the distributional hypothesis imparts a radically different

[1] For a representative instance of this widespread positioning of the AI problem, one can refer to the recent "AI Debate" between Yoshua Bengio and Gary Markus (Montreal, 2019). Cf. https://montrealartificialintelligence.com/aidebate/.

[2] Cf. Sahlgren (2008); Lenci (2008, 2018); Gastaldi (2020) for an in-depth presentation and discussion of the distributional hypothesis.

[3] Chomsky's rejection of probabilistic methods is well-known, as is his frequently quoted statement that "the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term" (Chomsky, 1969). For an early exposition of this viewpoint, see (Chomsky, 1957, § 2.4).

direction to linguistic research, where the knowledge produced is not so much about cognitive agents than about the organization of language. It follows that, understood as a hypothesis, *distributionalism constitutes a statement about the nature of language itself*, rather than about the capacities of linguistic agents. Hence, if the success of DNN models is to be endowed with epistemological significance, it is as the triumph of this conception of language that it should be primarily understood.

Linguistic distributionalism is far from new. As often recalled in the recent NLP literature, the distributional hypothesis finds its roots in the decades preceding the emergence of generative grammar, in the works of authors such as J. R. Firth (1957) or, more significantly before him, Z. Harris (1960; 1970a). It can be argued that this classical work in linguistics was chiefly theoretical for, although classical distributional methods provided some formal models (by the standards of that time) and even some computational tests on specific aspects of linguistic structure (cf. Harris (1970b,c)), they were not generally applied on real-life corpora at a significant scale.

And yet, DNN models are not the first to have achieved such large-scale formal implementation either: their use of the distributional hypothesis was long preceded by a family of matrix models whose origins go back to the early 1970s[4]. The main idea of such models—collectively referred to as *vector space models* (VSMs)—consists in representing linguistic units as *vectors*, whose dimensions are given by the possible linguistic contexts in which those units occur in a given corpus[5]. Each word in the vocabulary is then represented as a row in a matrix, whose cells collect information about the distribution of those words with respect to the linguistic contexts, represented by the columns. Finally, a dimensionality reduction can be performed upon that high-dimensional matrix through classic factorization methods (such as SVD), yielding low-dimensional dense vector representations for the words of the vocabulary, endowed with better generalization capabilities than sparse high-dimensional explicit vectors. Computing the distance between any pair of such vectors amounts to computing their distributional similarity (the more similar the distribution of two units, the smaller the distance between their vector representations) which turns to be directly correlated with different forms of linguistic relatedness[6].

DNN models for NLP can be seen as a way of producing and manipulating low-dimensional dense vector representations by other means than those of matrix models. Indeed, in the wake of the first DNN architectures introduced for specific linguistic tasks[7], researchers progressively realized that the network's initial (or "projection") layer could be considered as producing generic vector representations for the corresponding input words, and could be independently trained accordingly, to be used as the standard input form for DNNs oriented towards different tasks. In their most elementary form, such neural models for computing word vector representations, or *word embeddings*, associate a random vector of arbitrary fixed length to each word in a vocabulary, and train those vectors as a hidden layer in a dedicated neural network whose task is to predict

[4]See Turney and Pantel (2010) for an overview.

[5]Within the scope of these matrix models, different configurations of linguistic contexts have been studied. Cf. Sahlgren (2006) for a historical overview.

[6]See, for instance, Landauer et al. (2007) for a comprehensive presentation of *Latent Semantic Analysis* (LSA), one of the most popular models among VSMs.

[7]See Bengio (2008) for an overview of early DNN NLP models.

words out of the words surrounding them in a given corpus.

Although produced very differently than dense vector representations of previous matrix models, neural word embeddings rely on the same distributional phenomenon. Indeed, it has been shown that such word embeddings encode a great amount of information about word co-occurrence (Schnabel et al., 2015). More significantly, in a series of papers following the introduction of neural word embedding models, Levy and Goldberg (2014a) showed that one of the most successful of them, the Skip-gram model (Mikolov et al., 2013), was performing an implicit factorization of a (shifted) pointwise mutual information word-context matrix. What is more, the authors were capable of exhibiting performances comparable to that of neural models by transferring some of the latter's design choices and hyperparameter optimizations to traditional matrix distributional models (Levy and Goldberg, 2014b; Levy et al., 2015).

The pioneering neural embedding models succeeded in establishing distributed vector representations as the fundamental basis for the vast majority of DNN NLP models. Since then, increasingly sophisticated embedding models have been proposed, which take into account, among others, sub-lexical units (Bojanowski et al., 2016; Sennrich et al., 2016) or contextualized supra-lexical ones (Peters et al., 2018; Devlin et al., 2018; Radford, 2018; Brown et al., 2020). Their architecture and computational strategies differ in multiple ways[8], and their greater complexity, compared to initial neural word embeddings, would require that the formal link to more interpretable frameworks, like the one established by Levy and Goldberg, be reassessed. However, at their most elementary level, all those models share the same simple yet not trivial theoretical grounding, namely that of the distributional hypothesis, and even akin basic means of setting it up to determine the properties of linguistic units out of the statistics of their contexts in a given corpus.

2 Under Distributions, the Structure!

If we look back to its origins, it is possible to see that the distributional hypothesis constitutes, in fact, a corollary, or rather a simplified and usually semantically oriented version of a classic and more comprehensive approach to linguistic phenomena, known as *structuralism*. Structuralist linguistics precedes, and at least in part includes Harris's work, finding its most prominent American exponent in Harris's mentor, L. Bloomfield (cf. 1935), while its European roots go back to the seminal work of F. de Saussure, at the beginning of the 20$^{\text{th}}$ century, further developed by authors such as R. Jakobson and L. Hjelmslev (cf. de Saussure (1959); Jakobson (2001); Hjelmslev (1953, 1975)).

As distributionalism, structuralism is above all a theory about the nature of language rather than linguistic agents, based on a series of interconnected conceptual and methodological principles aiming at (and to a great extent required by) the complete description of linguistic phenomena of any sort. All those principles are organized around the central idea that linguistic units are not immediately given in experience, but are, instead, the formal result of a system of oppositional or differential relations that can be established, through linguistic analysis, at the level of the multiple supports in which language is manifested.

[8] For a good presentation of the variety of word embedding models, one may refer to Pilehvar and Camacho-Collados (2020).

A thorough assessment of the whole set of structuralist principles falls out of the scope of the present paper[9]. However, it is worth focusing on one of those principles which represents a key component of what structuralism takes to be the basic mechanism of language, namely the idea that those oppositional relations constituting linguistic units are of two irreducible yet interrelated kinds: *syntagmatic* and *paradigmatic*.

2.1 Syntagmas and Paradigms

In their most elementary form, *syntagmatic* relations are those constituting linguistic units (eg. words) as part of an observable sequence of terms (eg. phrases or sentences). For instance, the units `finds`, `me` and `in` are syntagmatically related in the sequence of terms: `The useless dawn finds me in a deserted streetcorner`[10]. Such units are thus recognized as coexisting in the same linguistic context, bearing different degrees of solidarity. It is this syntagmatic solidarity that contains the essence of the distributional hypothesis, as evidenced by Saussure's words:

> What is most striking in the organization of language are *syntagmatic solidarities*; almost all units of language depend on what surrounds them in the spoken chain or on their successive parts. (de Saussure, 1959, p. 127)[11]

Yet, structuralism considers another kind of relations that a linguistic unit can contract, namely associative or *paradigmatic* relations with all the other units which could be substituted to it at that particular position. Such units are not—and could not be as such—present in the explicit linguistic contexts of the term being considered. In the context given by our example, `me` bears a paradigmatic relation to units such as `you`, `her` or `someone`, which are not present in that context. While syntagmatic relations establish *coexisting* linguistic units, paradigmatic relations hold between *alternative* ones, thus implying an exclusive disjunction. Sets of units related syntagmatically are said to form *syntagmas* (or *chains*), while sets of paradigmatically related units constitute *paradigms*.

Figure 1 shows an illustration of syntagmatic and paradigmatic relations between lexical units (i.e. words) in the context of a linguistic expression, with the horizontal axis representing possible syntagmatic relations and the vertical paradigmatic ones. But such relations are not restricted to lexical units, and can be shown to hold between linguistic units at different levels, both supra and sub-lexical. For instance, following another example from Hjelmslev (1953, p. 36), from the combinations allowed by the successive paradigms {`p`, `m`}, {`e`, `a`} and {`t`, `n`} we can obtain the words `pet`, `pen`, `pat`, `pan`, `met`, `men`, `mat` and `man`, as syntagmas or chains of a higher level than that of the initial units (characters in this case).

[9] One may consult Ducrot (1973) for synthetic yet precise and faithful presentation of linguistic structuralism, as well as Maniglier (2006) for an in-depth analysis of its conceptual and philosophical stakes. We have addressed the connection between the structuralist approach and current trends in NLP in Gastaldi (2020).

[10] All our subsequent examples will be taken from the English language. We introduce the convention of writing linguistic expressions under analysis in a `monospace` font.

[11] Notice, however, that Saussure's notion of syntagmatic relations takes into account not only the relations between units but also those between sub-units within a unit.

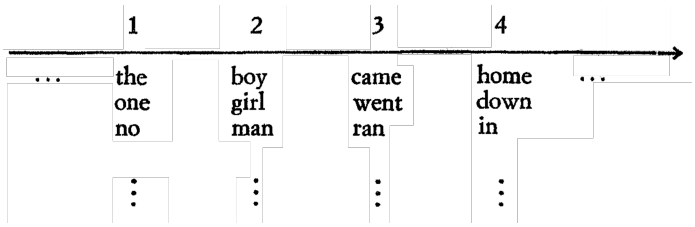


Figure 1: Hjelmslev's illustration of syntagmatic and paradigmatic relations respectively represented by the horizontal and the vertical axes (Hjelmslev, 1971, p. 210).

2.2 Varieties of formal content

It follows that, from a structuralist point of view, the properties of linguistic units are determined at the crossroads of syntagmatic and paradigmatic relations. Now, even without entering into all the subtleties associated with this dual determination of units[12], considering paradigmatic relations in addition to syntagmatic ones broadens the perspectives and goals of the linguistic analysis stemming from a popularized version of distributional semantics. For the paradigmatic organization revealed by the old structuralist lens can provide a much more precise account of the mechanisms involved in the relation among terms and contexts and of their effect on linguistic meaning. Indeed, if at any point of a linguistic sequence we can establish the multiple paradigmatic relations at play by providing the specific list of possible units from which the corresponding unit is chosen, a manifold of both syntactic and semantic structural features can be represented. If we agree to speak about "content" instead of "meaning" to circumvent the exclusive semantic import usually attributed to the latter, then we can say that paradigms helps us to identify at least three different dimensions of the content of linguistic units.

2.2.1 Syntactic Content

As for the first of those dimensions, returning to the example in Figure 1, we can see that a word like `boy` can be substituted by `girl` or `man`, or even `sky` or `sadness`. But it could not be substituted by `the`, `have` or `from` without making the corresponding sentences ungrammatical, i.e. without making the sequence of words not be a sentence at all. Such limitation of the domain of possibilities operated by paradigmatic relations at each position of a syntagmatic chain ensures the successful interaction between the terms in that chain, i.e. their capacity to combine in such a way that they constitute a unit of a higher level. The corresponding restrictions respond to multiple dependencies within the syntagmatic chain, recalling that language cannot be reduced

[12]Actually, for structuralism, linguistic or semiological units are determined at the intersection of not one but two sets of such series of syntagmatic and paradigmatic relations: the signifier and the signified, or the expression and the content planes. The need of a second set of syntagmatic-paradigmatic relations can, in principle, be explained by the insufficiency of just one series to determine all the relevant properties of units (the two sets borrowing determinations from one another). For the sake of simplicity, in this paper we restrict ourselves to syntagmatic and paradigmatic relations of expressions only, which is also closer to the way in which this problem is treated in NLP.

to a mere "bag of words". In this sense, they concern above all syntactic phenomena, like the one evidenced by Chomsky's famous example (1957) opposing `Colorless green ideas sleep furiously` to `Furiously sleep ideas green colorless`, where the interaction of the terms in the first case succeeds in establishing a sentence as a linguistic unit of higher level than those terms, even if the semantic content is suspended, while the interaction in the second case fails, making it significantly challenging, if not impossible, to attribute a content whatsoever to this expression other than that of being an apparently random sequence of terms. We can then call *syntactic content* this particular aspect of the meaning of linguistic units which is the effect of multiple dependencies with respect to other units in the context. Yet, such dependencies or restrictions do not hold directly between terms but between classes of terms, and are thus difficult to capture explicitly for an analytic procedure focused exclusively in syntagmatic relations. In contrast, the classes established by a paradigmatic viewpoint can contribute to restore those structural syntactic properties explicitly.

2.2.2 Characteristic Content

The second dimension of a unit's content revealed by paradigms is what we can term *characteristic content*. By being included in a class of substitutable terms, a term receives from the latter a positive characterization, given by the properties shared by all the terms in the class. In our example, all the words susceptible of occupying the place of `boy` will most certainly not only be nouns (syntactically) but also agents who can come, go and run. Incidentally, if those terms happen to have other common properties that the ones explicitly expressed in the context (eg. being human), the latter will also tend to be considered part of the characteristic content. Even in the case of unusual substitutions, such as `sky` or `sadness`, the common characteristics of regular terms constituting the paradigm induced at that position by the syntagmatic chain will invest those unusual terms with specific content attributes. If instead of `boy` we had found the term `gavagai` (Quine, 2013) at that place, then the corresponding paradigm $\{\texttt{boy}, \texttt{girl}, \texttt{man}, \ldots\}$ would contribute to reducing that term's complete semantic indeterminacy by projecting upon it all the characteristics shared by the terms in the paradigm. Characteristic content thus concerns the positive characterization of the meaning of terms from their distribution alone, and it is clear how challenging it would be to do so explicitly without some representation of paradigmatic units.

2.2.3 Informational Content

If the characteristic content was the only understanding of the semantics of linguistic units provided by paradigmatic relations, then the content of one term would be indistinguishable from that of any of the members of its paradigm. At best, its meaning could only be singularized at the intersection of all the paradigms to which it belongs. Yet, the mere existence of more than one member in a paradigm is an indication of the fact that the content of those members is not identical, as subtle as the difference may be. From this perspective, the choice of a particular term within the syntagmatic chain is done at the expense of all the others in the corresponding paradigm. Not only is such a choice related to the content of the term, but it can also be understood as constitutive

of it. Indeed, following the classic views of Shannon (1948), in line with those of structuralism on this point[13], the content conveyed by a term is completely determined by its choice among a class of other possible terms. We can thus call *informational content* this third dimension of content which singularizes each term by contrast with all the others belonging to the same paradigm.

Here again, we can see how the explicit derivation of paradigms can provide new perspectives compared to usual probabilistic models. For instead of computing the information conveyed by one term with respect to the entire vocabulary, paradigms restrict at each point of the syntagmatic chain the domain of terms whose distribution is relevant for such computation. In this way, the probability of a term acquires a semantic value. Take, for instance, the probability of `boy` in the English language, which is ~ 0.00010[14]. If this probability is compared to, say, that of `no` ($\sim$0.00145) or `down` ($\sim$0.00067) or any other random word in the entire English vocabulary, it is hard to see how this comparison can contribute to extending our knowledge of the meaning of `boy`. However, if we compare it to that of `girl` ($\sim$0.00013) and of `man` ($\sim$0.00063), to which it is paradigmatically related, then the resulting distribution becomes semantically relevant. Certainly, if conditional probabilities are considered, terms like `no` or `down`, conditioned on the same context as that of `boy`, will have a probability close to 0. Yet, by bringing paradigms into focus, we can understand that conditional probabilities result from the combination of at least two main components, distinguishable in principle: the probability of a paradigm or class as a whole, and the probability of terms within that class, both of which concern important aspects of that term's content. Computing probability distributions for terms only within specific paradigms can help to reveal the effect of structural features on the informational content of terms (eg. `one` being uninformative as a pronoun but informative as a determiner). More generally, the probability distribution of terms within paradigms can contribute to characterizing those paradigms at a more abstract level, for example, by noticing that some classes (like those containing nouns, for instance, or roots in the case subword units) will tend to have many members with low absolute frequency, while others (like classes containing prepositions, determiners or inflections at the subword level) will be rather composed of a small, and sometimes even definite, number of terms with high absolute frequency.

It appears that, by focusing on paradigmatic relations and units, the vague notion of meaning referred to in the distributional hypothesis can be specified into syntactic, characteristic and informational content, each of which manifests distinctive effects of contexts on words. This is not to say that meaning is entirely reducible to these three kinds of contents. Indeed, one can easily imagine others, such as referential, pragmatic or psychological contents. Yet, the three varieties revealed by the action of paradigms can very well be considered as the main dimensions of the *formal content* of linguistic units, that is the content resulting from formal relations, if we understand by "formal", following the young Chomsky, "nothing more than that it holds between linguistic expressions" (Chomsky, 1955, p. 39). This notion of formal also recalls Saussure's famous tenet that

[13] For a historical connection between the structuralist and the information-theoretical approaches to language, see Apostel et al. (1957); Jakobson (1967).

[14] According to Google Books Ngram Viewer (Michel et al., 2010, https://books.google.com/ngrams, `en_2019corpus`).

language is a form, not a substance (de Saussure, 1959, p. 113). And indeed, in the three varieties presented here, the content of terms is the result of oppositional relations determining differential entities, as in Saussure's account of linguistic mechanisms. Yet, in each case those relations are of a specific kind: in the case of syntactic content, the relation between classes of terms constitute paradigmatic units of different types, while the characteristic content results from differentiating paradigms irrespective of their type and the informational content emerges from differentiating singular terms within a paradigm.

It is worth insisting on the fact that, through the derivation of paradigmatic relations, the structuralist approach can capture both syntactic and semantic properties of language as the result of one and the same procedure. In this way, it recovers one of the most remarkable aspects of current distributional models, and of word embeddings in particular, which also exhibit this joint treatment of syntax and semantics (Mikolov et al., 2013; Avraham and Goldberg, 2017; Gastaldi, 2020). But unlike the latter, the structuralist representation of those properties is not limited to elementary probability distributions, similarity and relatedness measures or even clustering methods in the global embedding space. Relying on the derivation of paradigms, the structuralist approach promises to provide a representation of language as a complex system of classes and dependencies at different levels.

2.3 The Structuralist Hypothesis

The strengthening of the distributional hypothesis through structuralist methods, and especially through the derivation of explicit paradigms, entails some important conceptual consequences. Starting with the fact that, owing to the specification of the mechanisms by which linguistic context conditions the content of terms, a structuralist approach can dispense with the rather elusive notion of *use* supposed to be somehow reflected in the organization of language. Significantly, while resorting to such a notion of use would imply opening the linguistic model to the study of extralinguistic pragmatic or psychological aspects, the remarkable results of current distributional models do not benefit from any substantial contribution from them, other than those recorded in the corpus under analysis. Certainly, corpora are not disembodied devices unrelated to extralinguistic dimensions. Despite a general tendency to treat corpora as neutral and unbiased datasets, it is in the nature of a corpus to be an expression of concrete practices, as well as of a partial way of recording, selecting, normalizing and organizing them. However, within the limits of a corpus, those practices can only take a linguistic form. Corpus analysis is therefore entirely formal, in the double sense advanced above, that is relying on relations between expressions only, without considering any kind of substance[15]. This is not to say that psychological or pragmatical studies are not interesting *per se*, or that the results of current models should not be complemented with such studies, but only that, as a matter of fact, those results do not depend on such investigations. The resort to a notion of use in most of the literature around current distributional models thus remains mostly speculative and ineffective. In line with this situation, a structuralist viewpoint suggests that the source of linguistic content (both syntactic and semantic) is to be sought, neither in pragmatic

[15]Notice that, under such definition, formal analysis and properties are not supposed to be neutral or unbiased.

or psychological dimensions beyond language nor in any substantial organization of the world, but primarily in the fairly strict (although not closed) system of interdependent paradigms derivable, in principle, from the explicit utterances that system is implicitly governing. As Harris puts it:

> The perennial man in the street believes that when he speaks he freely puts together whatever elements have the meanings he intends; but he does so only by choosing members of those classes that regularly occur together, and in the order in which these classes occur. [...] the restricted distribution of classes persists for all their occurrences; the restrictions are not disregarded arbitrarily, e.g. for semantic needs. (Harris, 1970a, pp. 775-776)

It follows that the analysis of a linguistic corpus, inasmuch as it succeeds in deriving the system of classes and dependencies that can formally account for the regularities in that corpus, is a sufficient explanation of everything that is there to be *linguistically* explained. This idea constitutes a key component of what can henceforth be called the *structuralist hypothesis*, namely that *linguistic content is the effect of a virtual structure of classes and dependencies at multiple levels underlying (and derivable from) the mass of things said or written in a given language*. Accordingly, the task of linguistic analysis is not just that of identifying loose similarities between words out of distributional properties of a corpus, but rather this other one—before which the latter appears as a rough approximation—of explicitly drawing from that corpus the system of fairly strict dependencies between implicit linguistic categories. If we agree to adopt Hjelmslev's terminology and call *process* a complex of syntagmatic dependencies and *system* a complex of paradigmatic ones (Hjelmslev, 1975, p. 5), then the following passage from Hjelmslev's *Prolegomena* can be reasonably taken to express the essence of the structuralist hypothesis:

> *A priori* it would seem to be a generally valid thesis that for every *process* there is a corresponding *system*, by which the process can be analyzed and described by means of a limited number of premisses. It must be assumed that any process, can be analyzed into a limited number of elements recurring in various combinations. Then, on the basis of this analysis, it should be possible to order these elements into classes according to their possibilities of combination. And it should be further possible to set up a general and exhaustive calculus of the possible combinations. (Hjelmslev, 1953, p. 9)

3 The Challengens of an Emergent Calculus

Notice that, in Hjelmslev's view, the ultimate goal of linguistic analysis goes beyond the pure description of the data, and pursues the derivation of an exhaustive calculus. This goal is at least partially fulfilled by current distributional models, which are intended to be applied to data outside the training corpus or to be used as generative models. But if a calculus is necessarily at work in those models once they are trained, its principles remain entirely implicit. Here too, we can see how the structuralist derivation of a (paradigmatic) system out of (syntagmatic) processes can contribute to providing an *explicit representation* of

such a calculus, based on the particular way in which generalization is achieved trough paradigms. The example in Figure 1 can offer an elementary intuition of this mechanism. If, from a hypothetical corpus made of the three expressions corresponding to the three horizontal lines of the table, we are able to derive the four paradigms A, B, C, D, corresponding to the latter's columns, and then establish some of their combinatorial properties, for instance, the capacity of composing in the order $A \times B \times C \times D$,[16] then the explicit calculus that starts to be drawn in this way appears as the correlate of the generalization achieved by considering all possible combinations of the members of the paradigms at their corresponding positions (such as `the girl ran home`), the vast majority of which was not present as such in the original data upon which the system was built.

Incidentally, under this interpretation, the structuralist program challenges the classic distinction between connectionist and symbolic methods and its philosophical consequences (cf., for instance, Minsky (1991)). While beginning with combinatorial properties of linguistic units as raw data whose structure is only presupposed, the structuralist hypothesis aims at reconstructing an explicit and interpretable representation of the structure underlying such data, taking the form of a symbolic system at different levels (from the phonological all the way up to the grammatical or even stylistic level). From this perspective, symbolic systems implementing different aspects of algebraic structures are the direct result of the interaction of terms (including sub- or pre-symbolic ones) reflected in the statistics of given corpora. Conversely, when those symbolic systems are put into practice—in the performance of linguistic agents, for instance—the corresponding symbolic processes cannot but reproduce to a significant extent the statistical properties of the terms upon which that system was derived. Hence, from a structuralist perspective, connectionist and symbolic properties appear as two sides of the same phenomenon.

With the rather frail means of the epoch, the classic structuralist approach was able to prove its fecundity in the description of mainly phonological and morphological structures of multiple languages. However, empirical studies of more complex levels of language, and of grammar in particular, received mostly circumscribed and limited treatment. More generally, despite some valuable early efforts (Hjelmslev, 1975; Harris, 1960) structuralist linguistics encountered difficulties in providing effective formalized methods to describe syntactic structures in their full generality. The rise of Chomsky's generativist program in the late 1950s pushed the structuralist approach into obsolescence, until some of the latter's intuitions were recovered in the form of distributional methods by the resurgence of empiricist approaches in the wake of the emergence of new computational techniques in the 1980s (cf. McEnery and Wilson (2001); MacWhinney (1999); Chater et al. (2015)). As it turns out, the resurgence of distributional methods was mostly driven by semantic concerns, to the extent that, in the current state of the art, distributionalism has become indistinguishable from distributional semantics. Therefore, the question of a distributional calculus of a broader scope, including structural aspects of syntax no less than semantics, remains largely open.

[16] Of course, this example is for illustrative purposes only. The analysis of real corpora renders this task far more difficult, in particular due to distributional sparsity, which far from being a simple technical problem, has deep theoretical implications. See Yang (2016).

3.1 Obstacles to Paradigm Derivation

As we have seen, structuralism's main strategy to tackle the problem of an explicit calculus focuses on the derivation of paradigmatic units. However, establishing paradigms is a highly challenging task outside strongly controlled and circumscribed conditions. For if, at first sight, paradigms appear as simple classes of terms, such classes have the particularity of being at the same time of an extreme *precision*—since the inclusion of one incorrect term would be enough to jeopardize the successful interaction of linguistic terms—and perfectly *general*—since paradigms may contain an indefinite number of terms, either unseen in the data upon which they were derived or even not yet existent in the language under analysis, thus virtually allowing for an indefinite number of syntagmas or linguistic processes.

Those two conditions are somewhat in tension: generality excludes any purely extensional definition of paradigms, while precision makes intensional or logical definitions particularly complex, especially considering that they are to be drawn exclusively from distributional properties. Indeed, such precision is the result of the simultaneous action of multiple restricting principles, which are realized by terms interacting within a definite context.

Take, for instance, the expression `one girl has gone`. The paradigm of `has`, which can contain the terms `has`, `had`, `is` and `was`, is here delineated, from a purely distributional viewpoint, simultaneously by `gone`, selecting possible auxiliary verbs, and by `one` and `girl` selecting verbs or verb phrases that are singular. As simple as this example may be, it already allows us to indicate three major obstacles to the derivation of paradigms.

3.1.1 Circularity

The first and most important of those obstacles concerns the nature of the dependencies upon which a paradigm is supposed to be established. As it appears clearly in the example, the restricting principles establishing a paradigm correspond to several dependencies within the syntagmatic chain. But we also saw that such dependencies do not hold directly between terms but between classes of terms. From the point of view of paradigmatic derivation, this means that, while we can only rely on terms within the syntagmatic chain—the only ones accessible to experience—, paradigms do not contract dependencies with terms but with their respective characteristic contents (for instance, with the noun and singular characters of the term `girl`). Significantly, the characteristic content of a term can only be established through its paradigmatic relations, which rely, in turn, upon the same kind of dependencies with paradigms in the context, including the original paradigm we were intending to establish. Indeed, the singular character of `girl` is nowhere to be found other than in the fact that, within this context, terms like `boy` or `man`, but not `boys`, `men` or `girls`, belong to the paradigm of `girl`, a circumstance that depends, among other things, on the fact that the context of the paradigm including `has`, `had`, `is` and `was`, but not terms like `have` or `were` is present in the context.

The circularity of the task is manifest: paradigms are needed to establish paradigms. Yet, this circularity is not to be attributed to the method itself, but to the very nature of its object. Indeed, from a purely internal or empirical viewpoint, what are, for instance, adjectives, other than a particular class of

terms associated to nouns? And what are nouns if not something that can be modified by adjectives? These mutual dependencies are by no means restricted to syntactic classes, but pervade all levels of language: phonological (eg. vowels and consonants), morphological (eg. verb stems and inflections), semantic (eg. agent and actions), or even stylistic (eg. formal and familiar).

3.1.2 Hierarchical Compositionality

A second difficulty defying paradigmatic derivation concerns the composite organization of the restrictions delineating a paradigm. In the example above, for instance, the context of the term `has` most certainly allows drawing a paradigm including `has`, `had`, `is` and `was` and excluding `have`, `are` and `were`. But the interaction of this paradigm with the context is not homogenous and unidimensional. We mentioned at least two different features of that interaction: one with (certain aspects of) the characteristic content of `gone` and the other with that of `girl`. However, if the context is considered as a unanalyzed whole, as it usually is in current distributional methods, those dimensions remain indistinguishable. Considering the different paradigms composing that context could certainly help (if the circularity stated above was somewhat overcome), but it would not be enough. For it would still be required to know how those paradigms interact with each other establishing a complex system of dependencies. The singular character determining `has`, `had`, `is`, `was`, for instance, may find its source in the paradigm of `girl`, but also in that of `one`. Nevertheless, this would no longer be the case for a sentence like `one day the girl was gone`, where `one` and `girl` interact in a different way than in the original example. If we recall the idea of successful interaction of the previous section, we can say that in the first case both units successfully interact, composing a unit of a higher level, which in turn contributes, as a new unit, to the definition of the paradigm of `has`, while in the second they only interact indirectly (after interacting with other units) in such a way that the paradigm of `one` does not affect the definition of that of `has`. Into this kind of difficulties fall also classic examples such as `the boy and the girl have`, where the paradigm of `have` should contain plural terms and exclude singular ones, even if none of the paradigms for the words in the context can be expected to explicitly exhibit that characteristic (Chomsky, 1957, § 5.2-3).

To deal with this difficulty, some sort of compositionality principle should be found. But the mere composition of paradigms is not enough either. A more subtle mechanism is needed to assess the multiple ways in which different compositional principles are capable of interacting to derive a hierarchical structure.

3.1.3 Internal Paradigmatic Structure

Finally, when establishing a paradigm, it can happen that what constitutes its unity might not be immediately evident from the list of terms it contains. In our example, the words `courage` or `money` could also occupy the place of `gone`, and it is not clear what the unified content of the paradigm delineated by these terms could possibly be. It follows that, even if its members are not drawn in a purely random way, the internal coherence of a paradigm may not be completely guaranteed by the context, requiring further specification. And indeed, a quick

inspection of those members in our example suggests that the paradigmatic class could be analyzed into different subclasses, such as past participle verbs and nouns.

While the previous difficulty can be understood as concerning syntagmatic relations between paradigmatic units defining the structure of linguistic contexts, in this case we are confronted with the problem of the paradigmatic relations between (sub-)paradigms defining the structure of a paradigm containing them. The difficulty of this task resides in that, in principle, the context upon which the paradigm was derived in the first place has no explicit means to perform further discriminations within that paradigm, and it is not obvious what could be the source of those discriminations.

These difficulties have not remained unnoticed, even in the old days of structuralist research (see, for instance, Chomsky (1953)). They are also not the only ones that derivation of paradigms can encounter[17], most of all considering we have only presented them though extremely simple illustrations. Real-life analysis can only make this situation worse. In particular, trying to derive paradigms exclusively through corpus analysis can raise new difficulties unforeseen to a pre-computational structuralist perspective, for which the automatic processing of corpus of significant size remained after all a promising but peripheral possibility. Indeed, most of the structuralist original theoretical and methodological constructions are consciously or unconsciously conceived on the basis of linguistic data that can be produced by elicitation from an informant, if not through simple introspection. Problems like adequacy of probability measures, scarcity of data or impossibility statements (establishing, for instance, that two terms cannot stand in a given relation) barely appear among its original theoretical concerns.

And yet, in view of the resurgence of distributional methods and the growing necessity of making explicit the mechanisms directly or indirectly responsible for that success, it seems worth readdressing those main difficulties concerning paradigmatic inference, in the perspective of the renewal of those methods in this new setting. For the structural features current models have been shown to grasp, if only implicitly, are an indication that *such difficulties can be overcome.*

4 Towards a Type-Theoretical Emergent Calculus of Language

Despite their mainly semantic orientation, DNN models for NLP seem to capture a significant amount of structural features of language out of distributional properties, making them available for their precise application to a vast range of downstream tasks. This tends to confirm the claims of the structuralist

[17] In particular, we have disregarded here a fundamental problem which is nevertheless central from a structuralist standpoint, namely that syntagmatic relations between terms upon which our construction of paradigms relies as a given, do in fact require to be established in a way that also depends on the paradigmatic relations they are supposed to help constructing. Hence, in order to be entirely faithful to the structuralist perspective, a segmentation procedure should make part of the derivation of a linguistic system, not just as a preliminary step (such as "tokenization") but on a par with paradigmatic derivation. We leave the treatment of segmentation within this framework for an upcoming work.

hypothesis that those features can be derived from the analysis of linguistic performance alone, contributing to a better grasp of linguistic content. Indeed, several recent studies have provided evidence of the fact that those structures are indeed encoded in the corresponding models (Linzen et al., 2016; Enguehard et al., 2017; Dinu et al., 2018; Blevins et al., 2018; Goldberg, 2019; Clark et al., 2019; Hewitt and Manning, 2019; Manning et al., 2020; Bradley, 2020). However, the question of a method—be it neural or not—that could provide an explicit representation of such structure remains open.

4.1 Paradigms as Types

If we accept that paradigm derivation is a promising strategy to provide a representation of implicit structural features, then we should adopt an analytic framework where we can tackle the obstacles presented in the previous section. Now, all of those obstacles revolve around the idea that *paradigmatic units do not pre-exist the dependency relations they contract with other units of the same kind*. Therefore, the intended framework should address the dynamic establishment of dependencies as constitutive of its elementary classificatory objects. For this reason, we propose to represent paradigms as computational *types*.

The idea of representing linguistic phenomena through types is not new (Lambek, 1958; McGee Wood, 1993; Moot and Retoré, 2012; Fouqueré et al., 2018). Moreover, the recent success of DNN models has motivated several attempts to use types to provide explicit representations for both semantic and syntactic structures of language, based on current techniques, and of embeddings and vector representations in particular. However, in the vast majority of the cases, a separation persists between the construction of atomic (mostly semantic) types (Choi et al., 2018; Chen et al., 2020; Lin and Ji, 2019; Raiman and Raiman, 2018; Krishnamurthy et al., 2017; Abzianidze, 2016) and the establishment of (mostly syntactic) dependencies among them (Coecke et al., 2010; Clark et al., 2016; Coecke, 2019; Wijnholds and Sadrzadeh, 2019).

Within the type-theoretical tradition stemming from the study of the correspondence between logical proofs and computational processes (i.e. the Curry-Howard correspondence; cf. Groote (1995)), a singular research program originating in French proof-theory brings to the fore a notion of *interaction* upon which the types of a system are built from an intricate web of dependencies (Girard, 1989, 2001; Krivine, 2010; Miquel, 2020). This original perspective offers a powerful framework which could be mobilized to address the difficulties associated with the derivation of paradigms as more than simple classes, since not only derived types, but also atomic ones can be conceived as resulting from one and the same procedure, and are endowed with an internal structure that contain traces of the principles of their mutual relationships[18].

In this approach, types are conceived as sets that are closed under operations resulting from a notion of successful interaction defined at the level of their elements. More precisely, given a set A of elements of some set $\mathcal{L}$, we can consider its *orthogonal*, the set $A^{\perp}$ containing all the elements of $\mathcal{L}$ that successfully interact with the elements of A, for a given notion of successful interaction.

[18]Indeed, in Girard's *Ludics* (Girard, 2001), a tentative reconstruction of the whole of logic from an interactive point of view, there are no atomic types *per se*: atomic types are just types that are not yet decomposed.

Types are then defined as exactly the sets that are orthogonal to a set, and for any given set A, it is possible to construct a type that includes it by considering its *bi-orthogonal* (i.e. $A^{\perp\perp}$), which is fully defined by A. As a consequence, all the elements of a type constructed in this way are characterized by a common interactive behavior with respect to all the others elements in $\mathcal{L}$, whose behavior is, in turn, represented by their respective types[19].

To the best of our knowledge, this approach to types through interaction has not yet been applied to the treatment of natural language in a way that can contribute to the intelligibility and development of current NLP methods. However, the capabilities exhibited by this framework permit to suggest that a proper interpretation of interaction within natural language can help developing current distributional methods in the direction established by the structuralist hypothesis, by addressing the challenges to which the latter is confronted. In the rest of this article, we indicate how the obstacles presented in the previous section could find a suitable treatment from this perspective.

4.2 Circularity as (Bi-)Orthogonality

By understanding types as sets which are the orthogonal of other sets, types are conceived as the sets that are stable by the operations resulting from correct interaction. Hence, the circularity intrinsically involved in the extraction of paradigms is here embedded in the formal definition of types through orthogonality. Types are exactly the fixed points of this circularity, whose construction is inseparable from their dependencies with other types. That A is a type means nothing more (and nothing less) than that there is a certain dependency (captured by a notion of interaction) between the terms in A and other classes of terms, which can, in turn, be constructed as other types, thanks, among others, to the action of A.

All that is needed to put this framework into practice is an adequate notion of successful interaction defined over terms and classes of terms. In a pure computational setting, where terms are computational processes (i.e. programs), termination of interacting programs is often used (Riba, 2007). Although there is no unique natural way of defining such interaction in the case of natural language, we can intuitively associate it with distributional properties, i.e. two linguistic terms successfully interact if they co-occur with statistical significance within relevant contexts across a given corpus[20].

To give a general intuition of orthogonal types in natural language, consider, for instance, the expression `she must know`. Following the classical distributional method, we can determine, in a given corpus, the most frequent words that appear at each position of the context defined by this expression. We can then expect to obtain something similar to the following three classes of terms[21]:

[19] For a fairly accessible presentation of the technical framework, see Girard (1989).

[20] Certainly, the problem of statistical significance (and relevant contexts!) in this framework is extremely challenging, and falls outside the scope of the present paper. The work of Yang (2016) offers a promising perspective in this sense.

[21] For this toy example, we compute the paradigmatic classes rather naively, using Google Books Ngram Viewer (Michel et al., 2010, https://books.google.com/ngrams), with the following parameters: `en_2019` corpus, from 1900 to 2019, with a smoothing of 3. The use of wildcards permits to recover the most frequent (up to 10) words at a given place. Since this toy example has only an illustrative character, we disregard the difference in frequency of words within each class, and we order them alphabetically.

A	B	C
`he`	`could`	`also`
`i`	`did`	`be`
`one`	`may`	`do`
`she`	`might`	`get`
`they`	`must`	`go`
`we`	`should`	`have`
`you`	`would`	`know`
	`'d`	`make`
	`will`	`not`
	`'ll`	`take`

If we call A, B, C those three classes, we can now consider the set $A \times B \times C$ containing all the possible combinations of the terms of those classes, in that order (eg. `he did get`, `they could be`, `we should have`, etc.). We thus obtain a generalization of the linguistic data, since some of the expressions in $A \times B \times C$ (and in fact most, in the general case) will not exist in the corpus, although they constitute correct expressions of the language under study. As a result, the analysis can be carried beyond the original available data[22].

We can then say that a set A is orthogonal to a set B if all the of terms of A can co-occur with the terms of B. In our simple example, we can restrict the notion of successful interaction to simple concatenation of terms. Since the interaction between terms is not commutative in this case, a given set A will have two orthogonals: its left-orthogonal ${}^{\perp}A$, which contains all the left terms with which A interacts correctly, and its right orthogonal $A^{\perp}$, which is the same on the right. For example, if we take any subclass b of the class B defined before, say $b = \{\texttt{may}, \texttt{might}, \texttt{must}\}$, we have that its left orthogonal wihin the set $A \times B \times C$ coincides with the class A (i.e. ${}^{\perp}b = A$) and its right orthogonal is equal to C ($b^{\perp} = C$), both of which become types following our definitions. Moreover, we can consider the (right or left) bi-orthogonal of b, which is equal to the entire class B, and thus also a type, i.e. ${}^{\perp}(b^{\perp}) = ({}^{\perp}b)^{\perp} = B$.

In this way, we have constructed three types which are nothing more than the expression of the mutual dependencies that hold between them. Such types can behave like idealized paradigms which could be further refined based on the statistical properties of the initial corpus. More significantly, their formal construction permits to mobilize the entire type-theoretical apparatus in such a way that the remaining obstacles concerning paradigm derivation can be addressed in a new perspective.

4.3 Compositionality Through Connectives

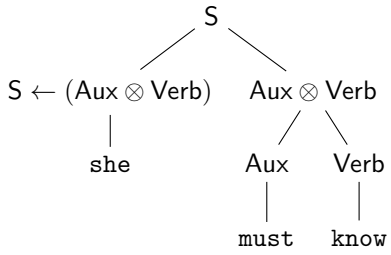
As we said, types constructed in this way, are endowed with an internal structure: not every set is a type, but only those that behave uniformly with respect to the selected notion of interaction. As it turns out, this internal structure can be used to govern the hierarchical compositionality of types. For, from a logical viewpoint, only the composition of types yielding another type (in the defined sense) are legitimate. Since there are, in principle, multiple ways in which we can compose types to build other types, owing to different properties

[22] See Yang (2016) for an insightful way of distinguishing between plausible and implausible generalizations.

of the interaction between those type's terms, and between those terms and their (respective or shared) contexts, different modes of compositionality can be expressed, which take the form of emergent logical *connectives* (Girard, 2001).

Turning back to our example, we can see that the type B contains principally modal verbs. It is not unreasonable to think that the consideration of other expressions in the corpus would reveal the significant presence of similar types in such a way that a type Aux of auxiliary verbs can be established as a formal unit referring to all those similar types. Moreover, suppose that we are also able to identify in a similar way a type Verb of verbs, and that we denote by $\otimes$ the compositional relation of succession. We can then establish that the compound term `must know` has the derived type $\mathsf{Aux} \otimes \mathsf{Verb}$. This means not only that its first element is in Aux and the second in Verb, but that, as a whole, they behave following the common behavior of the elements in $\mathsf{Aux} \otimes \mathsf{Verb}$. This common behavior can in turn be expressed in the usual way, namely by specifying its orthogonal: the set of terms that interact well with all its elements.

Suppose now that `she` is in ${}^{\perp}(\mathsf{Aux} \otimes \mathsf{Verb})$. If we also know that `she must know` is in S (the type of sentences, for instance), then, opening the door to an iterative definition, we can establish that `she` has type $\mathsf{S} \leftarrow (\mathsf{Aux} \otimes \mathsf{Verb})$ meaning that juxtaposing a $\mathsf{Aux} \otimes \mathsf{Verb}$ left to it yields a sentence. From this we can draw a syntax tree which expresses both the type of each component term, but also the way the different types are related through connectives:



As each different connective generates a different relation between the component types and the one resulting from their composition, different connectives are necessary. Indeed, here `she` might also be typed with a type Pron of pronouns (constructed through orthogonality, like the other types), and the whole chain then typed as $\mathsf{Pron} \otimes (\mathsf{Aux} \otimes \mathsf{Verb})$. Yet, if we focus on the entire chain being of type S, a different connective is required. Other connectives than the one presented here allow to express a rich variety of compositional relationships between types and their interaction with their respective contexts, following the perspectives of linear logic (Girard, 1987).

4.4 Analyzing Paradigms With Subtyping

Finally, if we take a closer look to our example, we can see that it relies not only on the orthogonals of the three classes we presented, but also on our ability, among the third type

$$C = \{\texttt{also}, \texttt{be}, \texttt{do}, \texttt{get}, \texttt{go}, \texttt{have}, \texttt{know}, \texttt{make}, \texttt{not}, \texttt{take}\},$$

to discriminate a specific subset $C' = \{\texttt{be}, \texttt{do}, \texttt{get}, \texttt{go}, \texttt{have}, \texttt{know}, \texttt{make}, \texttt{take}\}$ representing verbs, from $C'' = \{\texttt{also}, \texttt{not}\}$ containing adverbs. This problem

corresponds precisely to the third of the obstacles presented in the previous section.

The difficulty resides in the fact that, within the context considered, there is no means to operate the necessary distinction. Of course, if the type Verb mentioned before was already constructed, then the difficult would be easily circumvented (since $C \cap \mathsf{Verb} = C'$). But constructing the type Verb might, of course, require that we are able to distinguish C' from C'' in the first place. However, C' can result from more direct distributional properties if other interactions are taken into account. For instance, if we consider, in the way described, the most frequent words coming after `she must also` and `she must not`, we get the following classes, respectively:

$$\{\texttt{she must also}\}^{\perp} = \{\texttt{be}, \texttt{have}, \texttt{keep}, \texttt{know}, \texttt{learn}, \texttt{make}, \texttt{show}, \texttt{sign}, \texttt{take}, \texttt{understand}\}$$

$$\{\texttt{she must not}\}^{\perp} = \{\texttt{be}, \texttt{do}, \texttt{expect}, \texttt{forget}, \texttt{go}, \texttt{have}, \texttt{know}, \texttt{let}, \texttt{only}, \texttt{think}\}$$

Using the information provided by these new types, it should be possible to produce subtypes in C so that the distinction between C' and C'' is conveniently approached.

Conclusion

In this paper we have attempted to extend the scope of the distributional hypothesis underlying current DNN NLP models with the perspectives of classic structural linguistics. In particular, we suggested that, by focusing on paradigmatic units, we can both specify the mechanisms by which distributional properties produce linguistic content and provide explicit representations for the implicit structural features involved in that process. Moreover, we have identified three major challenges for the derivation of paradigms from distributional or syntagmatic properties alone and we have proposed a type-theoretical framework where they could be assessed and overcome.

The ideas presented in these pages should be seen as an exercise of conceptual bridging between disciplines which are not usually considered together (NLP with AI, philosophy of language, structuralist linguistics, computational logic). As such, they have a primarily speculative value. Their validity can only be established by empirical results, which are the object of our current work.

References

Lasha Abzianidze. Natural solution to FraCaS entailment problems. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 64–74, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-2007.

L. Apostel, B. Mandelbrot, and A. Morf. *Logique, Langage et Théorie de l'Information*. Presses Universitaires de France, 1957.

Oded Avraham and Yoav Goldberg. The interplay of semantics and morphology in word embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short*

Papers, pages 422–426, Valencia, Spain, April 2017. Association for Computational Linguistics.

Yoshua Bengio. Neural net language models. *Scholarpedia*, 3(1):3881, 2008. doi: 10.4249/scholarpedia.3881.

Terra Blevins, Omer Levy, and Luke Zettlemoyer. Deep rnns encode soft hierarchical syntax, 2018.

Leonard Bloomfield. *Language*. G. Allen & Unwin, Ltd, London, 1935. ISBN 0044000162.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomáš Mikolov. Enriching word vectors with subword information. *CoRR*, abs/1607.04606, 2016. URL http://arxiv.org/abs/1607.04606.

Tai-Danae Bradley. At the interface of algebra and statistics. *ArXiv*, abs/2004.05631, 2020.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

Nick Chater, Alexander Clark, John A. Goldsmith, and Amy Perfors. *Empiricism and language learnability*. Oxford University Press, Oxford, United Kingdom, first edition edition, 2015. ISBN 978-0-19-873426-0. OCLC: ocn907131354.

Tongfei Chen, Yunmo Chen, and Benjamin Van Durme. Hierarchical entity typing via multi-level learning to rank. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8465–8475, Online, July 2020. Association for Computational Linguistics.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. Ultra-fine entity typing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 87–96, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1009.

Noam Chomsky. Systems of syntactic analysis. *The Journal of Symbolic Logic*, 18(3):242–256, 1953. ISSN 00224812.

Noam Chomsky. Logical syntax and semantics: Their linguistic relevance. *Language*, 31(1):36–45, 1955. ISSN 00978507, 15350665.

Noam Chomsky. *Syntactic Structures*. Mouton and Co., The Hague, 1957.

Noam Chomsky. *Quine's Empirical Assumptions*, pages 53–68. Springer Netherlands, Dordrecht, 1969.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does bert look at? an analysis of bert's attention, 2019.

Stephen Hedley Clark, Laura Rimell, Tamara Polajnar, and Jean Maillard. The categorial framework for compositional distributional semantics. 2016.

Bob Coecke. The mathematics of text structure, 2019.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. Mathematical foundations for a compositional distributional model of meaning, 2010.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Georgiana Dinu, Miguel Ballesteros, Avirup Sil, Sam Bowman, Wael Hamza, Anders Sogaard, Tahira Naseem, and Yoav Goldberg, editors. *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Oswald Ducrot. *Le structuralisme en linguistique*. Éditions du Seuil, Paris, 1973. ISBN 9782020006194.

Émile Enguehard, Yoav Goldberg, and Tal Linzen. Exploring the syntactic abilities of rnns with multi-task learning. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 3–14, 2017.

John Rupert Firth. A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis*, pages 1–32. Blackwell, Oxford, 1957.

Christophe Fouqueré, Alain Lecomte, Myriam Quatrini, Pierre Livet, and Samuel Tronçon. *Mathématique du dialogue: sens et interaction*. Hermann, 2018. ISBN 9782705697211.

Juan Luis Gastaldi. Why can computers understand natural language? *Philosophy & Technology*, May 2020.

Jean-Yves Girard. Linear logic. *Theoretical Computer Science*, 50(1):1 – 101, 1987. ISSN 0304-3975. doi: https://doi.org/10.1016/0304-3975(87)90045-4.

Jean-Yves Girard. *Towards a Geometry of interaction*, volume 92 of *Contermporary Mathematics*, pages 69–108. AMS, 1989.

Jean-Yves Girard. Locus solum: From the rules of logic to the logic of rules. *Mathematical Structures in Computer Science*, 11(3):301–506, 2001. doi: 10.1017/S096012950100336X.

Yoav Goldberg. Assessing bert's syntactic abilities, 2019.

Philippe De Groote. *The Curry-Howard Isomorphism*. Academia, 1995.

Zellig Harris. *Structural linguistics*. University of Chicago Press, Chicago, 1960. ISBN 0226317714 0226217714.

Zellig Harris. Distributional structure. In *Papers in Structural and Transformational Linguistics*, pages 775–794. Springer, Dordrecht, 1970a.

Zellig S. Harris. *Computable Syntactic Analysis: The 1959 Computer Sentence-Analyzer*, pages 253–277. Springer Netherlands, Dordrecht, 1970b.

Zellig S. Harris. *Morpheme Boundaries within Words: Report on a Computer Test*, pages 68–77. Springer Netherlands, Dordrecht, 1970c.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419.

Louis Hjelmslev. *Prolegomena to a Theory of Language*. Wawerly Press, Baltimore, 1953.

Louis Hjelmslev. *La structure fondamentale du langage*, pages 177–231. Éditions de Minuit, Paris, 1971. ISBN 9782707301345.

Louis Hjelmslev. *Résumé of a Theory of Language*. Number 16 in Travaux du Cercle linguistique de Copenhague. Nordisk Sprog-og Kulturforlag, Copenhagen, 1975. ISBN 0-299-07040-9.

Roman Jakobson. *Preliminaries to speech analysis: the distinctive features and their correlates*. M.I.T. Press, Cambridge, Mass, 1967. ISBN 9780262600019.

Roman Jakobson. *Roman Jakobson: Selected Writings*. Mouton De Gruyter, Berlin Berlin, 2001. ISBN 9783110173611.

Jayant Krishnamurthy, Pradeep Dasigi, and Matt Gardner. Neural semantic parsing with type constraints for semi-structured tables. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1516–1526, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1160.

Jean-Louis Krivine. Realizability algebras: A program to well order r. *Logical Methods in Computer Science*, 7(3), 05 2010. doi: 10.2168/LMCS-7(3:2)2011.

Joachim Lambek. The mathematics of sentence structure. *The American Mathematical Monthly*, 65(3):154–170, 1958. doi: 10.1080/00029890.1958.11989160.

Thomas K. Landauer, Danielle S. McNamara, Simon Dennis, and Walter Kintsch, editors. *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, Mahwah, New Jersey, USA, 2007.

Alessandro Lenci. Distributional semantics in linguistic and cognitive research. *From context to meaning: distributional models of the lexicon in linguistics and cognitive science, Italian Journal of Linguistics*, 1(20):1–31, 2008.

Alessandro Lenci. Distributional models of word meaning. *Annual Review of Linguistics*, 4(1):151–171, 2018. doi: 10.1146/annurev-linguistics-030514-125254.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2177–2185, Cambridge, MA, USA, 2014a. MIT Press.

Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning, CoNLL 2014, Baltimore, Maryland, USA, June 26-27, 2014*, pages 171–180, 2014b.

Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *TACL*, 3:211–225, 2015.

Ying Lin and Heng Ji. An attentive fine-grained entity typing model with latent type representation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6197–6202, Hong Kong, China, November 2019. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of lstms to learn syntax-sensitive dependencies, 2016.

Brian MacWhinney, editor. *The emergence of language*. Carnegie Mellon symposia on cognition. Lawrence Erlbaum Associates, Mahwah, NJ, 1999. ISBN 978-0-8058-3010-1 978-0-8058-3011-8.

Patrice Maniglier. *La vie énigmatique des signes*. Léo Scheer, Paris, 2006.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 2020. doi: 10.1073/pnas.1907367117.

Anthony M. McEnery and Anita Wilson. *Corpus Linguistics: An Introduction*. Edinburgh University Press, Edinburgh, 2001.

Mary McGee Wood. *Categorial grammars*. Routledge, London, 1993.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Holberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 2010.

Tomáš Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

Marvin L. Minsky. Logical versus analogical or symbolic versus connectionist or neat versus scruffy. *AI Magazine*, 12(2):34, Jun. 1991. doi: 10.1609/aimag.v12i2.894.

Alexandre Miquel. Implicative algebras: a new foundation for realizability and forcing. *Mathematical Structures in Computer Science*, 30(5):458–510, May 2020.

Richard Moot and Christian Retoré. *The Logic of Categorial Grammars: A Deductive Account of Natural Language Syntax and Semantics*. Lecture Notes in Computer Science 6850. Springer-Verlag Berlin Heidelberg, 1 edition, 2012. ISBN 978-3-642-31554-1,978-3-642-31555-8.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *CoRR*, abs/1802.05365, 2018. URL http://arxiv.org/abs/1802.05365.

Mohammad Taher Pilehvar and Jose Camacho-Collados. Embeddings in natural language processing. theory and advances in vector representation of meaning. Draft, 2020. URL http://josecamachocollados.com/book_embNLP_draft.pdf.

W. V. Quine. *Word and object*. MIT Press, Cambridge, Mass, 2013. ISBN 978-0-262-51831-4.

Alec Radford. Improving language understanding by generative pre-training. 2018.

Jonathan Raiman and Olivier Raiman. Deeptype: Multilingual entity linking by neural type system evolution, 2018.

Colin Riba. Strong normalization as safe interaction. In *LiCS '2007*, pages 13–22, 2007.

Magnus Sahlgren. *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*. PhD thesis, Stockholm University, Stockholm, Sweden, 2006.

Magnus Sahlgren. The distributional hypothesis. *Special issue of the Italian Journal of Linguistics*, 1(20):33–53, 2008.

Ferdinand de Saussure. *Course in General Linguistics*. McGraw-Hill, New York, 1959. Translated by Wade Baskin.

Tobias Schnabel, Igor Labutov, David M. Mimno, and Thorsten Joachims. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 298–307, 2015.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. ACL. doi: 10.18653/v1/P16-1162.

Claude E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948.

Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *CoRR*, abs/1003.1141, 2010.

Gijs Wijnholds and Mehrnoosh Sadrzadeh. A type-driven vector semantics for ellipsis with anaphora using lambek calculus with limited contraction. *J. of Logic, Lang. and Inf.*, 28(2):331–358, June 2019.

Charles Yang. *The price of linguistic productivity : how children learn to break the rules of language*. The MIT Press, Cambridge, Massachusetts, 2016.