Generative computational modelling

Paola Merlo
University of Geneva
paola.merlo@unige.ch

Giuseppe Samo
University of Geneva / BLCU
giuseppe.samo@unige.ch

1 Computational Modelling: the third way to science

Computational modelling and simulation-based research are becoming an accepted, and in some cases dominant, way of doing science. Simulations and computational modelling are a third way to science, complementing experiments and theory. These methods have become commonplace in many scientific disciplines, and in some areas of the language sciences.

1.1 What is a model?

So, what is a model? In its most common usage, a model is an abstract representation of an item or a concept: a car, a plane, or a building. Models are created in order to view, manipulate, or test the object they represent without having to build the real thing. For example, an architect draws a model of a building, without having to pick up a hammer or a nail. If the building is complex, the architect builds a small physical model of the building (these days often by 3D-printing) to judge its appearance and aesthetics. For a representation of a problem to be called a model it has to have three properties: (i) *Mapping*: A model is based on an original; (ii) *Reduction*: A model only represents some relevant features of the original; (iii) *Pragmatic*: A model needs to be usable in place of the original with respect to some purpose.[1]

The art of modelling in general is to choose the right features to represent the problem concisely and perspicuously. The art of computational modelling is to do so in a computationally efficient way. A good model is therefore not just a practically manageable representation of a complex problem, but often a good explanation of the problem, if it makes the relations and causalities between the parts perspicuous and precise. A good computational model has the added value of supporting questions about the efficiency of computation.

Traditional computational modelling applied to any research topic has had for a long time the added and often essential value of forcing the researcher to make their theory and assumptions painfully explicit. In the words of Guest and Martin (2021, 689): "The scientific inference process is a function from theory to data (...). It is a path function that steps from theory, specification and implementation before an interpretation can have explanatory force in relation to a theory." Computational modelling helps make this process explicit and reproducible.[2]

The kind of modelling we will mostly talk about will be quantitative and data-driven. Theoretical linguistics can benefit from this way of doing science. For example, a fundamental feature within the Minimalist program (cf. Chomsky 1995 and related works) is that syntactic architectures are built based on computational resources that have a cost, and that natural language syntax is designed to minimize such costs. By using computational modelling, one can quantify and weigh the involved operations, formalize theories and compare their relative complexity.

A computational model needs to be an abstract and reduced version of the conceptual object it is meant to model. One essential contribution of recent computational modelling has been the representation of linguistic problems in forms that lend themselves to automatic manipulation by learning algorithms. A typical technique is to represent the object of study as a collection of relevant attributes and their values. The simplest organisation of attributes

[1] There are some areas of investigations where models in a 1:1 scale can be useful. For example, the Human Brain project (Koslow and Huerta, 2013) aims to reproduce the human brain at a 1:1 scale, and support research that would be invasive and simply impossible for ethical reasons.

[2] It might be that current deep learning models will turn out to be less helpful than older, simpler ones in this respect. Whether current machine learning models, with all their opaque layers of complexity and their random initialisation processes, still perform this conceptual clarification function is the object of active debate and investigation, and we will touch on this issue briefly at the end.

and values is a vector. So we can represent the object of study O as a pair (x, y), where $x = (a_1, a_2, ..., a_n)$ and $y = a_{n+1}$. So O is a vector of attributes where x can be thought of as the independent and y the dependent variable, the variable we want to predict in the context of x. The algorithms to calculate the relation between x and y can vary in complexity and especially can vary in the presence or absence of a latent space of unobserved objects and quantities that need to be postulated.

We will present several types of models. Conceptually, they can be organised by the amount of manual intervention in the construction of the model: some models define both attributes and values by hand; some models predefine the attributes but estimate the values by corpus-driven data collection; some models define both attributes and values by data-driven induction.

1.2 The quantitative aspect: an aside

Some of the models we will talk about are built by hand in both attributes and values of the independent vector and aim to predict a quantitative dependent attribute, mostly some kind of frequency, typological frequencies or corpus frequencies, for example. We will call these *theory-driven counting models*. Theory-driven counting models assume that frequencies are not accidents of usage or performance, but linguistic facts that systematically covary with abstract representational properties. They are part of the grammar (Merlo 1994; see Yang 2015 for discussions in acquisition, see also Diessel and Hilpert 2016 for a critical account; footnote for details).[3]

A very well-known point of view maintains that frequencies are not part of grammar or the cognitive system (Chomsky, 1965). The only area in which frequency plays a role is in the notion of markedness. Markedness has been equated in generative grammar to the cost of structure building or movement operations (Cinque, 2005, 2013) or to default parameter setting (Travis 1984; Baker 2002; Yang 2003; see also Burnett et al. 2018 for a quantitative analysis). This point of view is not supported by current practical successes on language tasks based on large amounts of text and their quantitative properties, as we will discuss in the third section.

Theory-driven counting models put forward a point of view on frequency that differs both from the functionalist approach— an approach where frequency is the cause and linguistic phenomena are the effect (cf. Ibbotson 2013) — and also differs from the traditional formal grammar disregard for quantitative aspects of formal representations. Frequency is neither the independent variable in the explanation nor irrelevant to language.

[3]Connections between linguistic constructs and frequency counts have been shown before. It has been shown, for example, that subcategorisation frequencies in corpora are correlated to grammaticality judgements (Merlo, 1994), that deep principles of verbal lexicon organisation, namely verb classes, show robust statistical regularities within and across languages (Merlo and Stevenson, 2001; Merlo et al., 2002; Samardžić and Merlo, 2018). These results have been followed by many other investigations in verb classification, sense disambiguation, propositional annotation, distributional lexical semantics, among other topics (Palmer et al., 2005; Schulte im Walde, 2006; Abend et al., 2008; Baroni and Lenci, 2010). Frequency distributions and their relation with some notions of acquisition or processing complexity have also been explained as an effect of pressure for efficient communication (Dryer, 1992, 2009; Hawkins, 1994, 2004; Gibson, 1998; Tily et al., 2011; Fedzechkina et al., 2012; Zipf, 1949; Reali and Christiansen, 2007). Despite these observations, the status of frequency counts in a theory of grammar is hard to pin down. Functionalists and formal grammarians are in agreement in assuming that frequencies are an expression of language use. They then treat the relationship between grammar and frequency in a very different way. Functionalist approaches have addressed the relationship between frequency and grammar by assuming that usage shapes grammar, and that frequency of use is the cause of some prominent linguistic effects, especially related to change or as the zero-coded, most flexible, typologically flexible element(see Bybee 2007, or Haspelmath 2006, among many others)

1.3 Tour of the paper

Some of the models we present here formulate solutions to linguistic problems, the linguists' problems, in a quantitative or computational way. In so doing, they provide novel insights beyond what more theoretical approaches have discovered. Some other models have been developed in computational linguistics and natural language processing for other goals, mostly technological, but by studying their solutions and the computational properties that led to these solutions, we sometimes find intriguing answers to linguistic questions.

Section 2 presents two types of theory-driven counting models. The first type comprises fully predefined models that define both attributes and values by hand (section 2.1). They mostly deal with typological and parametric issues. Section 2.2. illustrates partially predefined models where attributes are defined theoretically, but values are collected by corpus inspection. Current large-scale, syntactically-annotated resources support these investigations of the correlation between quantitative linguistic occurrences and much more specific, abstract linguistic representations and operations. We report here a coherently developed set of results in this vein in the generative tradition. This work has been collected under the moniker of Quantitative Computational Syntax (Merlo 2015b, 2016; Merlo and Ouwayda 2018; Samo and Merlo 2019, 2021; Gulordava and Merlo 2020; Merlo and Samo 2022, see also van Craenenbroeck and van Koppen (forthcoming) for its contextualisation in a larger view of quantitative studies). We also briefly survey some other quantitative approaches to dialectal variation and cross-linguistic parameters settings.

Section 3 presents models where attributes and their values are both inductively defined, the most salient characteristics of neural networks. Moving from hand-defined theory-driven counting models to architectural explanations with artificial neural networks (henceforth ANN) requires a change in the researcher's point of view in dealing with the computational model. Hand-defined models, whether fully or partially hand-defined, are better treated as the researcher's artifact and it makes sense to take their performance as the demonstration of the inspiring underlying theory. Models in the form of current ANNs define their own internal representation, so they are better conceptualized as natural objects. Natural objects preexist the creation of the researcher and the consequence of internal modifications cannot necessarily be studied deductively. This shift in point of view is reflected in the large literature on the interpretability of ANN, which will be discussed here as a particularly interesting example of how linguistic theory can interact with complex computational architectures. Interpretability research uses the comparative method: the linguistic theory and the computational model are developed independently and are then compared to see if the computational model can prove a mechanistic embodiment of the linguistic theory.

2 Theory-driven counting models

This section presents two types of counting models, organized based on the amount of the researcher's intervention in the design of the model. The modelling component interacts with the theory in two different ways, in our examples. In some cases, it gives rise to simple computational learning models, for example decision trees, that support the comparison of theories or the investigation of costs of operation proposed by the theory. In other cases, the counting investigations are the prerequisite for successive computational models. For example, our counting results show that intervention effects are present in the statistics of the training corpora and that lack of effects in the computational models (some results in section 3, for example) cannot be the effect of lack of signal in the input.

				Dryer's Languages	Dryer's Genera	Cinque's 05 Languages	Cinque's 13 Languages
Dem	Num	Adj	N	74	44	V. many	300
Dem	Adj	Num	N	3	2	0	0
Num	Dem	Adj	N	0	0	0	0
Num	Adj	Dem	N	0	0	0	0
Adj	Dem	Num	N	0	0	0	0
Adj	Num	Dem	N	0	0	0	0
Dem	Num	N	Adj	22	17	Many	114
Dem	Adj	N	Num	11	6	V. few (7)	35
Num	Dem	N	Adj	0	0	0	0
Num	Adj	N	Dem	4	3	V. few (8)	40
Adj	Dem	N	Num	0	0	0	0
Adj	Num	N	Dem	0	0	0	0
Dem	N	Adj	Num	28	22	Many	125
Dem	N	Num	Adj	3	3	V. few (4)	37
Num	N	Dem	Adj	5	3	0	0
Num	N	Adj	Dem	38	21	Few (2)	180
Adj	N	Dem	Num	4	2	V. few (3)	14
Adj	N	Num	Dem	2	1	V. few	15
N	Dem	Num	Adj	4	3	Few (8)	48
N	Dem	Adj	Num	6	4	V. few (3)	24
N	Num	Dem	Adj	1	1	0	0
N	Num	Adj	Dem	9	7	Few (7)	35
N	Adj	Dem	Num	19	11	Few (8)	69
N	Adj	Num	Dem	108	57	V. many (27)	411

Table 1: Attested word orders of Universal 20 and their counts: the first two columns report counts of genera and languages in Dryer's sample, the last two columns report counts from Cinque (2005) and from a large sample (Cinque, 2013, p.c.).

2.1 Fully hand-defined counting models

The models presented in this section cover several topics but they are all similar in the use of vectorial representations to describe the data, vectors where both the elements and their values are defined on theoretical grounds.

2.1.1 Measuring costs of word orders

One of the most easily observable distinguishing features of human languages is the order of words: the position of the verb in the sentence or the respective order of the modifiers of a noun, for example. Word orders vary greatly cross-linguistically and they can also show some variation within a given language (Cinque, 2005; Cysouw, 2010; Steedman, 2011; Culbertson et al., 2012; Culbertson and Smolensky, 2012).

A proposal in line with the minimalist attention to costs of operations is found in Cinque (2005), where Greenberg's Universal 20 is derived from independently motivated syntactic operations, organised in a derivational explanation.

Greenberg's universal 20 identifies four main elements in the noun phrase: nouns (N), demonstratives (Dem), numerals (Num) and adjectives (Adj) and states, explicitly or implicitly, many typological properties of NPs.[4] First, it states that not all the logically possible orders

[4] **Greenberg's Universal 20** When any or all the items (demonstrative, numeral, and descriptive adjective) precede the noun, they are always found in this order. If they follow, the order is exactly the same or its exact opposite.

NoPP: Uses NP movement without pied piping
XPNP: Uses NP movement with pied-piping of the [XP [NP]] type
NPXP: Uses NP movement with pied-piping of the [NP[XP]] type
Par: Involves partial movement
Split: Uses NP-splitting movement
NPless: Requires movement of a phrase not containing the NP

Word Order	Step 1	Step 2
DEM NUM A N	No movements	
DEM NUM N A	NP above Adj	No more movement
	[NP[XP]]-Move	Partial movement
N DEM NUM A	NP above Dem	
	No-Pied-Piping	

	NoPP	XPNP	NPXP	Par	Split	NPless	Freq
DEM NUM A N	0	0	0	0	0	0	300
DEM NUM N A	0	0	1	1	0	0	114
N DEM NUM A	1	0	0	0	0	0	48

Figure 1: Top panel: Different binary movement features. Bottom panel: Sequence of movement steps to generate a word order and resulting vectors of feature values and corresponding frequencies.

are attested. Second, it states that two postnominal orders but only one prenominal order are attested, thereby establishing an asymmetry between prenominal and postnominal distributions of word orders. Third, it states that of the three possible orders, the prenominal order is the mirror image of one of the postnominal orders. The frequencies of occurrence of the combinatorially possible orders of the four elements on the Noun Phrase have been discussed in several publications (Dryer 2006, Cinque 2005, 2013) and the counts put forth in these publications are shown in Table 1.[5] The different sets of counts show substantial agreement between counts and genera and across different samples. Cinque puts forth a proposal where the distribution of observed and unobserved word orders is generated by a derivation where the actually attested orders, and none of the unattested ones are derivable from a single universal base order, and from independent conditions on phrasal movement.

To build a model, Merlo (2015b)'s proposes a vectorization method to translate Cinque's derivational explanation of Universal 20 into vectorial representations (see also the replication in Futrell et al. 2017). Merlo and Ouwayda (2018) develops this method further: the derivational operations are represented as binary properties that a word order either has or does not have, each of them a feature in a vector (Figure 1).

The results are interesting. Merlo and Ouwayda's use a linear regression to automatically rank the costs of different syntactic movements within Cinque's proposal and investigate some proposals on partial and complete movement. This investigation of movement suggests that the nature of the movement is important, while the importance of harmonic specification of functional categories, i.e. whether the movement is partial or complete, is more context-dependent. A second contribution of the paper is the investigation whether the base order DEM NUM ADJ N is the best predictor of the typological facts. Comparing different syntactic proposals on the position of numerals in the noun phrase, it is found that a merge position of

[5] In the models that we present later, we use Cinque's most recent counts. We report here Dryer's counts by languages and by genera to illustrate the fact that taking into account density of linguistic neighbourhood in the sample (i.e. counting genera rather than single languages) does not fundamentally change the typological distribution of counts. In Dryer's definition "a genus is a group of languages whose relatedness is fairly obvious without systematic comparative analysis and which even the most conservative "splitter" would accept." Examples are such subfamilies of Indo-European as Germanic, Slavic, and Romance languages.

Cysouw

	Noun edge	NA-Adjacency	Dem edge	Freq
DEM NUM A N	1	1	1	300
DEM NUM N A	1	0	1	114
N DEM NUM A	0	1	0	48

Dryer

	Sym1	Sym2	Asym	Harmony	U18	Freq
DEM NUM A N	1	1	1	1	1	300
DEM NUM N A	1	1	1	1	0	114
N DEM NUM A	0	0	1	1	1	48

Figure 2: Vectors of features values representing different theories and corresponding frequencies.

numerals higher than adjectives has better results in both methods.

2.1.2 Comparing theories of Universal 20

One of the advantages of vectorisation of theoretical proposals is that they become comparable. Several other authors have attempted to reconcile word order principles with typological observations by a system of costs and constraints that generate statistical universals also in a typological tradition (Cysouw, 2010; Dryer, 2006, 2009). In one proposal, three characteristics are used as predictive variables: hierarchical structure, noun-adjective order, and whether the noun is at the phrase boundary (Cysouw, 2010). In another model, a factorial explanation is proposed based on general principles of symmetry and harmony (Dryer, 2006, 2009).[6] These models are already expressed as systems of declarative principles. Each word order is then directly encodable as vectors of features, each feature being the principle at work, encoded as a binary feature that can take a positive value if the principle is active and a negative value if it is not. Figure 2 shows the encoding of the different proposals for three different word orders.

Taking the generalising ability of the proposals on unseen data as our measure of explanatoriness, we can compare these different approaches. Vectorial representations can be used with several machine learning methods. Merlo (2015b) discusses Bayesian methods. The comparison can also be performed through the very simple method of decision trees. Decision trees produce a very easily understandable output and are based on the notion of information gain encoded in the use of every attribute. The topology of the tree is then directly readable as the informativity of each feature: the highest the node in the tree the more informative. In Figure 3 and Figure 4, we show the decision trees that are output by Cysouw's and Dryer's encodings respectively.

The results are revealing. The three binary features proposed in Cysouw do not provide as much information (the predictive performance is 90% compared to 97% of Dryer's). In Dryer's model, the principle which encodes universal 18 is the most informative. Looking at the confusion matrices, we notice that all three models do better on the classification of frequent orders than rare orders, but Cysow's has more difficulty with rare orders. Cysouw's limited ability to generalise might indicate that principles developed specifically to explain Universal 20 are not good expressions of the general notion of complexity. For example, many

[6]The factors comprise two symmetry principles that describes the closeness of the modifiers to the noun: Symmetry Principle 1 describes the preference of the adjective and numeral to occur closer to the noun than the Determiner, and symmetry principle 2 states that the adjective tends to occur closer to the Noun than the Numeral, when they occur on the same side of the Noun. Dryer also uses a principle of asymmetry that captures the main observation that prenominal modifiers exhibit fewer alternatives than post-nominal modifiers (also observed by Cinque); a principle of intra-categorial harmony, which says that all modifiers tends to occur on the same side of the noun; and Greenberg's universal 18 (see footnote 9 for a definition of U18).

```
N-edge = Y
    Dem-edge = Y
        NA-adjacency = Y: VF (103.0/2.0)
        NA-adjacency = N: R (7.0)
    Dem-edge = N
        NA-adjacency = Y: F (12.0/1.0)
        NA-adjacency = N: R (11.0/3.0)
N-edge = N
    NA-adjacency = Y: F (73.0/13.0)
    NA-adjacency = N: R (8.0/2.0)

Number of leaves :  6
Size of the tree : 11
```

Class	P	R	F
VF	0.98	1.00	0.99
F	0.83	1.00	0.91
R	0.81	0.60	0.69
No	0.00	0.00	0.00

Class	VF	F	R	No
VF	101	0	0	0
F	0	71	0	0
R	0	12	21	0
No	0	2	5	0

Figure 3: Cysow token-wise decision tree, precision, recall and F scores, and confusion matrix.

```
U-18 = Y
    Sym1 = Y
        Sym2 = Y: VF (101.0)
        Sym2 = N: R (9.0)
    Sym1 = N
        Asym = Y
            Sym2 = Y: F (15.0/4.0)
            Sym2 = N: R (4.0)
        Asym = N: No (4.0)
U-18 = N
    Harmony = Y
        Sym1 = Y
            Sym2 = Y: F (60.0)
            Sym2 = N: R (3.0)
        Sym1 = N: R (4.0/1.0)
    Harmony = N
            Asym = Y: R (12.0)
            Asym = N: No (2.0)

Number of leaves :  10
Size of the tree : 19
```

Class	P	R	F
VF	1.00	1.00	1.00
F	0.95	1.00	0.97
R	0.91	0.87	0.90
No	1.00	0.57	0.73

Class	VF	F	R	No
VF	101	0	0	0
F	0	71	0	0
R	0	4	31	0
No	0	0	3	4

Figure 4: Dryer token-wise decision tree, precision, recall and F scores, and confusion matrix.

languages in the world exhibit an SVO sentential order, where the main element, the verb, is medial. It appears, then, that a general preference for being at the edge of the phrase is not

likely. Dryer's proposals provides a more principled models whose performance is also better, although at the cost of more assumptions.

2.1.3 Modelling micro and macro-variation

The detection of micro- and macro-variability plays an important role in another piece of work where entire theories are encoded as vectors, the line of research developed in the so-called Parametric Comparison Method (Longobardi 2003, 2018; Guardiano et al. 2020; see Crisma et al. 2020 for a recent description). Here the vectorial representations encode parameters, in terms of syntactic rules and not languages or structures. Each parameter is binary and its value is determined by the boolean answer to a YES/NO question (e.g., 'Does a (set of) structure(s)/interpretation(s) so-and-so occur in language L?', Crisma et al. 2020, 111) and the answer YES is assumed only on the basis of positive evidence.[7] Parameters are encoded then as the values they receive in languages in order to find patterns of interdependence. Implicitly, languages (among which local varieties, cf. Teramano, an Italian dialect, in Table 2) are thus encoded as values of parameters.

Parameter	Teramano	Italian	Spanish	French
FSN	+	+	+	+
FNN	-	+	+	-
FGT	-	-	-	-

Table 2: Based from Figure 1 in the supplementary files of Ceolin et al. 2020 (see also Guardiano and Longobardi 2016; Longobardi 2018), showing 3 parameters and 4 languages. FSN = number spread to N; FNN = number on N; CGR = a grammaticalized temporality.

The methodology aims to find patterns of dependencies across parameters (e.g., hierarchical relationships, cf. Biberauer and Roberts 2017) and similarities across languages. For example, Kazakov et al. (2017) adopted a symbolic machine learning algorithm, decision trees to find dependencies in a table of parameters (producing dependency graphs, Figure 5, left panel). In a similar spirit, Ceolin et al. (2020), aimed to find similarities across and within languages and parameters (Figure 5, right panel) exploring Jaccard distances (Jaccard 1901, from 0 to 1: the closer to 1, the more similar two lists/entities).

Along similar lines, formal-theoretical analysis and feature-based methods are used to map micro- and macro-variation across languages, dialects and idioms. Rules of grammar (e.g., word orders, values of parameters) and languages/varieties are integrated into a quantitative analysis with inferential and predictive statistics, alongside geographical and/or social variables (see details in van Craenenbroeck and van Koppen, forthcoming). In these pieces of work, the dependent variable is either the geographic dimension or the binary presence or absence of a given property to be predicted. Some relevant examples of micro-variability in the Germanic and Romance syntactic landscape are presented in Van Craenenbroeck et al. (2019) and Pescarini (2022).

Van Craenenbroeck et al. (2019) map sets of verb clusters across varieties of Dutch to identify main tendencies and correlations. Specifically, Van Craenenbroeck et al. (2019) analyses the distributions of combinations of two and three verb clusters, exemplified in Table 3. The authors extract data from the Syntactic Atlas of Dutch Dialects (SAND, Barbiers et al. 2005) and turn each cluster into a vectorial representation whose features indicate the presence or absence of a particular cluster in a given variety (indicated by the Boolean operators in Table 3, NA if the datapoint is missing).

[7] Each parameter is identified by a progressive number and by a combination of three capital letters (e.g. see labels in Table 2; the see details of the methodology in Longobardi 2018, 523).

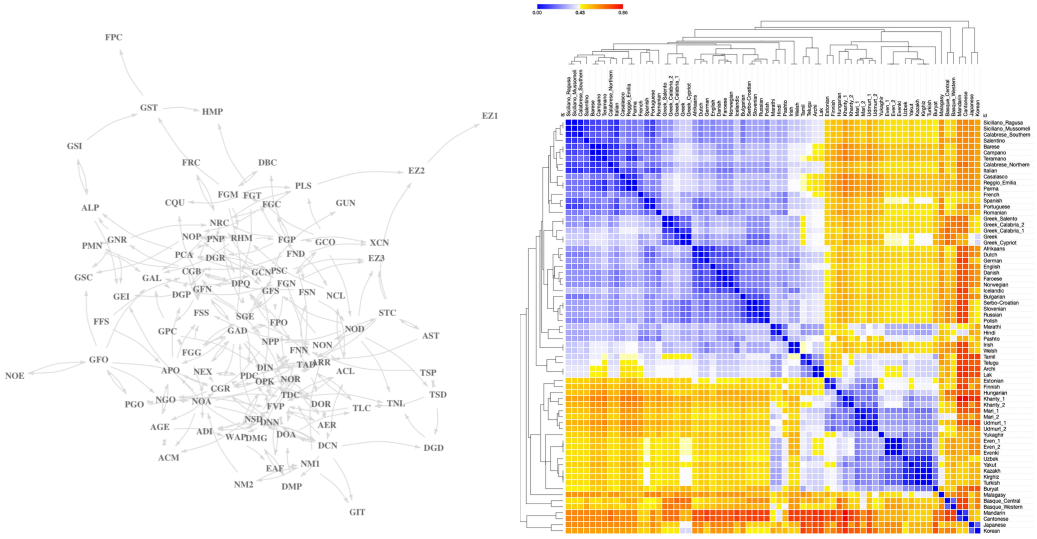

Figure 5: Left panel: full dependency graph of interdependence of parameters (from Kazakov et al. 2017, Figure 2, p. 34. Right panel: Heatmap of syntactic Jaccard distances between the 69 languages of the sample, calculated on 94 parameters (from Ceolin et al. 2020, 7, Fig.1).

	Example	Varieties			
	ik vind dat iedereen...				
Cluster	'I find that everyone'	Ml	L	Wt	Oo
MUST_CAN_SWIM	*moet kunnen zwemmen*	0	0	0	0
MUST_SWIM_CAN	*moet zwemmen kunnen*	0	0	0	1
(*)CAN_SWIM_MOD	*kunnen zwemmen moet*	1	0	0	0

Table 3: Data extracted from Van Craenenbroeck et al. (2019), ex. 6, p. 338 and Table 2 p. 341. * = ungrammatical in Dutch; Ml = Midsland; L = Lies; Wt = West-Terschelling; Oo = Oosterend.

	Persons					
Variety	1	2	3	4	5	6
Olivone (Ticino)	1	1	1	1	1	1
Moncalvo (Piedmont)	0	1	1	0	0	0
Valmacca (Piedmont)	1	1	1	0	0	1
Breme (Lombardy)	1	1	1	1	1	1

Table 4: Based on (Pescarini, 2019, 265, Fig.2). Varieties correspond to the 187 datapoints surveyed by Manzini and Savoia (2005), columns present the presence (optional/mandatory, 1) or absence (0) of subject clitic forms for each Person (1-6).

Van Craenenbroeck et al. (2019) explore machine learning and statistical techniques (e.g. k-nearest neighbour classification, correspondence analysis) to identify main tendencies and correlations, in grammatical and extra-grammatical terms. For example, Van Craenenbroeck et al. (2019) reduced verb clusters' variability in Dutch dialects to three grammatical parameters, in line with theoretical considerations in Barbiers et al. (2018), and how these are distributed across varieties.

In a similar vein, Pescarini (2019) investigates the distribution (in terms of presence/absence) of dedicated subject clitics in Northern Italian dialects. Pescarini (2019)'s conclusion is that microvariation in the inventory of clitics cannot derive only from external factors (e.g. contact, sociolinguistics dynamics), but a representational model in terms of principled features can account for relevant patterns of defectivity/syncretims. Interestingly, the correlation between linguistic and geographical distances with respect to the inventory of subject clitics in Northern Italian dialects, is "surprisingly low" (Pescarini, 2019, 275). A similar conclusion emerge when Pescarini (2022) compares models containing both geographical and grammatical information framing negative marking strategies in Central Romance varieties.

Pescarini (2022) gathers data from the Linguistic Atlas of France (ALF) and the Italo-Swiss Atlas (AIS). These data contain information on the type of negation (e.g. negation derived from Latin non and negation derived from various kind of elements, e.g. French *pas* 'step'). In this case every observation (i.e., a reply to a questionnaire) is encoded with (i) geographical and sociolinguistic information, (ii) the presence/absence of the two types of negation and (iii) additional syntactic elements (e.g., veridicality). Pescarini (2022) applies multiple linear regression finding correlations. He finds, among other results, that (i) negative polarity items correlate positively (but not for all negative polarity items) with the presence of negation derived from Latin non and (ii) that preverbal quantifiers disfavour negative concord with negation of the French *pas* type, while adverbial PPs do allow it. These conclusions are in line with the syntactic literature on Gallo-Romance.

Conclusions This first set of models shows how to perform precise theory and model comparison. In this practice, one does not develop new theories, but simply takes existing theories and distills their more perspicuous features in a formal, uniform representation, the vectorial representation. This method allows one to compare proposals and draw conclusions on their performance and their ability to generalise and learn. This formalisation also allows the researcher to be precise about notions of complexity, simplicity, and similarity, fundamental notions in theory development and assessment.

2.2 Partially predefined counting models

As already indicated, an important element within the Minimalist program is that syntactic operations have costs in terms of computational resources, and, thus natural language syntax

is designed in order to minimize such costs (see details in Chomsky 1995). Empirical results on locality has shown that natural languages do prefer local simplicity of configurations (cf. Rizzi 2013).

2.2.1 Intervention Locality

The theory of locality in A′-constructions has played a major role in the conceptual development of generative grammar, whether in terms of islands (Ross, 1967), superiority conditions (Chomsky, 1973) or barriers (Chomsky, 1986a). Later developments, such as Relativized Minimality (Rizzi 1990, 2004) and Minimal link conditions, (Chomsky 1995) take into account the length of dependencies between two syntactic links (e.g. a moved element and its gap) and the nature of the syntactic constituents involved in the dependency. Specifically, the nature of the structure is encoded in terms of morpho-syntactic features in featural Relativized Minimality (henceforth, fRM Rizzi 1990, 2004; Starke 2001; Rizzi 2013). The different geometry of morpho-syntactic features creates asymmetries across constructions and languages.

According to fRM, a local relation is disrupted by the intervention of an element with certain properties which makes the intervener a potential participant in that local relation. Consider the examples in (1). The example shows the asymmetry between different types of relative clauses (Friedmann et al. 2009; Sanfelici et al. 2014 *inter alia*), in which the fronted relativized element - generated in the verb argument structure and moved to a CP position - may cross intervening syntactic elements (in bold), as in (1b,c) or not, as in (1a) .[8]

(1)
a. The professor$_{\text{XP, sg}}$ that <the professor$_{\text{XP, sg}}$> helps the student$_{\text{XP, sg}}$.

b. The student$_{\text{XP, sg}}$ that **the professor**$_{\text{XP, sg}}$ helps <the student$_{\text{XP, sg}}$>.

c. The student$_{\text{XP, sg}}$ that **they**$_{\text{pro, pl}}$ help <the student$_{\text{XP, sg}}$>.

Despite all three sentence types being grammatical, degrees of syntactic complexity emerge, quantified in ungrammaticality or slower parsing time in adults, difficulties in acquisition and in populations with language impairments (Grillo 2008; Friedmann et al. 2009; Belletti et al. 2012; Sanfelici et al. 2014; Villata et al. 2016 among many others). Due to the lack of interveners, (1a) is described as easier to parse than (1b) and (1c) - the subject moves to its landing position without encountering blocking constituents in its way. An additional asymmetry is that (1c) is easier than (1b). fRM models also predict that parsing improves across populations of speakers, if the two elements in a relation are dissimilar in the values of features (for instance, maximality of projection and number). The results of extensive experimental research have demonstrated the effect of dissimilarity in values of a selected set of features, which may vary according to languages (e.g. gender in Hebrew, but not in Italian, cf. Belletti et al. 2012).

The varying degrees of syntactic complexity lend themselves to investigation in corpora. Practically, a theory-driven corpus-based investigation is possible because the investigated A′-constructions are grammatical clauses, so they are retrievable in large-scale datasets. Conceptually, the core intuition underlying a theory-driven approach to corpus frequencies is that we need to compare the frequency counts we see in a corpus with those that the theory would predict. The representational nature of fRM makes certain predictions clear. Specifically,

[8]Morphosyntactic features are indicated in pedex (Maximality, values: XP = maximal projection, pro = pronoun; Number, values: sg = singular, pl = plural).

one can expect (i) that no intervention is easier than intervention, (ii) that intervention created by elements bearing mismatching features is easier than intervention with elements with matching features and, finally, (iii) that languages may vary in the computation of syntactic locality. The notion of easy or hard is reflected in corpus counts.

In this spirit, Samo and Merlo (2019) investigated locality constraints in object relative clauses (cf. 1b and 1c) in English and Italian. Their analysis was based on a comparison of theoretically expected counts, built on the distribution of features (maximality, number, and animacy) in canonical clauses , and observed counts in actually retrieved object relatives (see Table 5 for a representational model of occurrences of object relative clauses in English). The predictions of fRM were confirmed: intervention with dissimilar elements occurs more frequently than expected and in Italian, the feature number plays a role in the computation of locality. Furthermore, the feature animacy also creates an effect, as predicted by a broader view of locality (exemplified in Table 6).

Match			Relative head			Intervener			Sentence
type	num	an	type	num	an	type	num	an	
0	0	0	XP	sg	in	head	pl	an	*the foreign investment* that **they** need to help
0	1	0	XP	pl	in	head	pl	an	*the fees* that **they** charge
1	0	0	XP	sg	in	XP	pl	an	*a luxury* that **only rich countries** can afford
1	0	1	XP	sg	an	XP	pl	an	*a better person* that **people** are wanting to hire
1	1	0	XP	sg	in	XP	sg	an	*a realist technique* which **French novelist Marcel Proust** later named
1	1	1	XP	sg	in	XP	sg	in	*a format* that **Access** recognizes

Table 5: Examples of OR clauses in several featural configurations in English. The examples show the values of the features and if they match (1) or not (0) between head of the relative clause (in italics, in the examples) and the intervener (in bold). The examples are naturally occurring clauses extracted from the UD corpora. From Samo and Merlo 2019, Table 4, p. 6.

Along similar lines, Samo and Merlo (2021) investigate clefts in three languages (English, French and Italian) which vary in terms of usage of such structures. Clefts are another case of A′-constructions, which syntactically share the structure of relative clauses, involving a movement of a constituent in a higher layer of the syntactic tree. In their results, subject and object clefts confirm the different acceptability levels found in experimental settings: object clefts are less frequent than expected in intervention configuration, while subject clefts are roughly as frequent as expected (Figure 6). Samo and Merlo (2021) also found an interesting finer-grained effect: the size of the effect is proportional to the number of features that give rise to the intervention effect.

2.2.2 Word order

The papers on word order in section 2.1 used the vectorization method to study entire theories or languages, providing a feature-based method to represent word orders in individual languages, based on primitives that represent types of movement and structural representations. This representation supports comparisons across theories and validations of costs of operations. This section instead has introduced a different method, which predicts corpus counts and has applied it to problems in the theory of locality. It should however also be noted that methods and objects of study are interchangeable and word order too can be studied by exploring corpus predictions, as done in Gulordava and Merlo (2020). In this paper, quantitative evidence is presented about Universal 18.[9] The work shows that corpus data confirm a dispreference for

[9] "When the descriptive adjective precedes the noun, the demonstrative, and the numeral, with overwhelmingly more than chance frequency, do likewise."

Match condition
English

HRel	Int	Exp	Obs	p	Binomial-p	z-p
XP	XP	123.0	**108**	0.490	0.033	0.033
sing	sing	128.7	132	0.511	0.341	0.341
plur	plur	20.3	22	0.081	0.382	0.393
anim	anim	51.4	**20**	0.205	0.000	<.001

Match condition
Italian

HRel	Int	Exp	Obs	p	Binomial-p	z-p
XP	XP	164.3	**149**	0.620	0.031	0.031
sing	sing	131.4	138	0.496	0.218	0.219
plur	plur	22.7	34	0.860	0.011	0.008
anim	anim	41.3	**23**	0.156	0.001	0.001
inam	inam	46.6	**27**	0.176	0.001	0.001

Mismatch condition
English

HRel	Int	Exp	Obs	p	Binomial-p	z-p
XP	head	120.5	**135**	0.480	0.383	0.038
XP	null	7.5	0	0.030	0.001	*n.v.*
sing	plur	47.4	49	0.219	0.203	0.202
plur	sing	53.2	40	0.189	0.131	0.132
anim	inam	3.90	0	0.015	0.022	*n.v.*
inam	anim	182.1	**211**	0.725	<.001	<.001

Mismatch condition
Italian

HRel	Int	Exp	Obs	p	Binomial-p	z-p
XP	head	13.3	**29**	0.050	<.001	<.001
XP	null	87.5	**101**	0.330	0.045	0.044
sing	plur	46.2	**59**	0.174	0.025	0.022
plur	sing	64.7	48	0.244	0.009	0.010
anim	inam	11.7	0	0.044	<.001	<.001
inam	anim	165.4	**229**	0.624	<. 001	<.001

Table 6: Expected counts (Exp) and observed counts (Obs) in Samo and Merlo (2019), between matching and mismatching conditions betweeen the relativized head (HRel) and the intervening subject (Int). Binomial test: Binomial p indicates the probability of the observed counts under a binomial distribution. z-p is the statistical significance of the binomial probability as the (one-tailed) probability of exactly the observed, or greater/smaller counts than the expected counts. *n.v.* indicates that conditions are not met for a valid calculation of statistical significance. Results confirming the hypotheses are in bold. Adapted from Samo and Merlo 2019, Table 6, p. 8

the word order combination where adjectives precede but numerals follow the nouns (Adj-N and N-Num). It then investigates if this dispreference is better explained as a constraint expressed at the level of the dominant orders (for example, as different coefficients of combinations of the rule in the grammar, as proposed by Culbertson and Smolensky 2012) or at the level of individual structures, as a single distributional probability of a given structure (as proposed by Cinque 2005, or by the Final-over-Final-Constraint, cf. Sheehan et al. 2017). Corpus counts support the latter interpretation.

2.3 Conclusions

In the quest for balance between the far-sightedness of theoretical constructs and the empirical richness of large amounts of data, it pays to enrich the explanatory power of the data we use. The data used in grammaticality judgements in formal grammar is nominal and

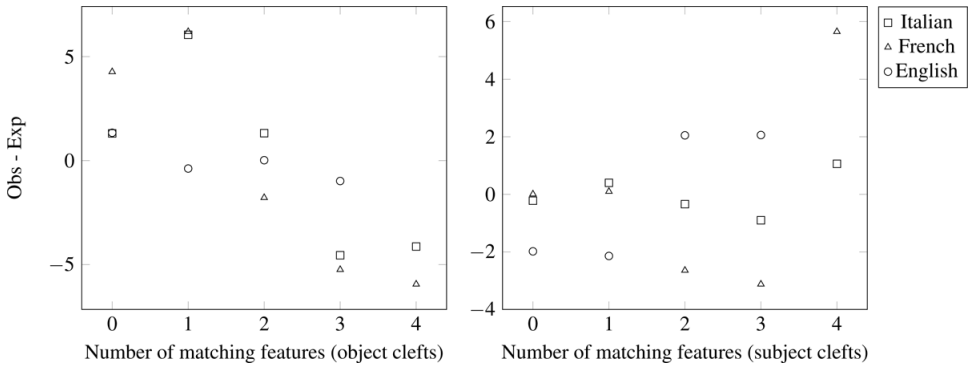

Figure 6: Difference between observed and expected counts as a function of the number of matching features in object clefts (left panel, Pearson $r = –0.79$, $p < .05$) and in subject clefts (right panel, Pearson $r = 0.41$, $p = 0.12$). From Samo and Merlo 2021, Figure 2, p. 21

sometimes ordinal. Nominal data are the least expressive as they cannot be compared or ordered, and few statistical techniques can be applied to them. Quantitative data support more elaborate theories, which take into account some non-categorical facts about language. Because quantitative data is more expressive, they also support theories that have fewer a priori assumptions, without losing explanatoriness or predictiveness.

In this section, we have presented simple counting models and simple learning models applied to well-studied linguistic phenomena: movement operations in the typological explanation of word orders, model selection for the adjective-order preference, relative clauses and clefts and their degrees of acceptability, dialectal variation. These relatively simple approaches add a quantitative aspect to qualitative linguistic observations, and provide more concrete generative mechanisms to the notion of markedness and gradient of acceptability, by adding numerical parameters to the formal notions proposed by the linguistic theory.

Counting models have evolved with time into probabilistic models. A Bayesian probabilistic model has the ability to incorporate both uncertain observations and theoretical biases to provide a functional explanation, an explanation that abstracts away from the actual mechanisms to calculate probabilities and centers instead on the different components of the problem (see Lau et al. 2017 and references therein). We will not discuss these kinds of models, for reasons of space, but in the next section we illustrate instead the current trends in neural network architectures and their interpretability.

3 Complex architectural models

The recent resurgence of neural networks and deep learning and their increasingly astonishing successes have renewed interest in architectural and mechanistic explanations. An insightful historical perspective of the counterpoint between grammar-based approaches and distributed neural network approaches is provided by the pioneer of neural networks in natural language processing (Henderson, 2020). Henderson (2020) explores the history of applying neural networks to natural language understanding tasks, arguing that variable binding is one of the main areas of difference between the distributed representations of neural networks and the algebraic approaches of more traditional language processing methods. The paper highlights the novel instantiation of variable binding in attention-based models (Vaswani et al. 2017), arguing that, for this reason, transformers are not sequence models but induced-structure models and hold promise for more structured linguistic problems. A good summary of the issues close to the heart of linguists is provided in Linzen and Baroni (2021). Modern deep neural

networks achieve impressive performance in engineering applications that require extensive linguistic skills. This success has stimulated interest in investigating whether these models can induce human-like grammatical knowledge from raw data and contribute to the long-standing debates about the innate structure necessary for language acquisition (see details in the recent contribution in Wilcox et al. 2022). There is a growing body of work of this nature, where different architectures are investigated to probe their abilities to represent and acquire linguistic phenomena and cast light on problems raised in more linguistic venues. We review some here.

A neural network is built by computational units. A unit, inspired by a single neuron, takes a vector of inputs, performs some computation on it and produces an output. The computation performed is usually a nonlinear function (for instance, a sigmoid or a rectified linear function) applied to the dot product of the input vector with a vector of weights (and a threshold bias).

The power of neural networks derives from combining these simple units into many layers and different patterns. The intermediate layers of the network, sandwiched between the observable input and output, are called hidden layers. These layers encode a hidden representation, the weights of the intermediate layers, a representations that is unobservable and that arises autonomously by training the network. So, in this sense, compared to the models we have discussed in the previous section, both the features and their values are induced by a more generic learning procedure. Most of these neural network models are trained by optimizing a language modelling task. A language modelling task is traditionally defined as the task of predicting the next word. Recently, it has been extended to bidirectional training tasks, and also to word masking tasks, hence predicting masked words in the context of the sentence, analogously to a cloze test.

The simplest form of neural network is a feed-forward neural network, where units are organized in layers that have no cycles. Recursive Neural Networks (RNNs) have cycles in their structure, where the output of one step is fed back as an input to the next step. This allows RNNs to process sequential data, like language, where inputs are fed one at a time, because they can maintain a hidden state that encodes the previous information. RNNs suffer from technical shortcomings, in particular they tend to show a recency bias, among others. LSTMs (Hochreiter and Schmidhuber, 1997) are a special kind of RNNs that have a more complex structure, with gates, to control how much information is allowed to enter, leave, or be forgotten by the unit. This way, LSTMs can learn long-term dependencies. Another useful concept is the notion of encoder-decoder architecture, where a network encodes the input and a decoder network produces the output. For example, the encoder encodes sentences in French and the decoder output the sentence in English. Here again, the problem is the bottleneck created by the sequential structure of the encoder. The solution resides in the notion of attention. Transformers (Vaswani et al., 2017) can process all inputs in parallel and use an attention mechanism to selectively focus on different parts of the input data allowing the architecture to detect and capture long-range dependencies in the data. Finally, Variational Autoencoder (Kingma et al. 2016; see also Kingma et al. 2019; Lin et al. 2020; Vahdat and Kautz 2020; Henderson and Fehr 2023) consist of two main parts: an encoder network that maps the input data to a latent representation, and a decoder network that maps the latent representation back to the input space. The encoder network learns to approximate the true posterior distribution of the latent variables, while the decoder network learns to reconstruct the input data from the latent representation. For more details on the different architectures, see the excellent Jurafsky and Martin (2023), third edition.

3.1 RNNs and LSTMs and agreement

As we indicated in the introduction, despite their practical success and impressive performances, neural networks as models of language remain fundamentally opaque. A whole new trend of research on the interpretability of these models has developed. Recall that the process of investigation is voluntarily different from what we have seen in the previous section. Here, we deal with much more complex models, but more importantly, we deal with generic models, whose architectures have not been developed specifically for a specific task or problem. The genericity of the architecture has the advantage that its formal properties are well-studied and are of general application beyond special cases. It has the disadvantage that exactly what the network has learnt needs to be studied *a posteriori*.

To cast light on what linguistic information is learned and encoded in these representations, several pieces of work have recently studied core properties of language in syntax (Linzen et al., 2016; Bernardy and Lappin, 2017; Gulordava et al., 2018; Linzen and Leonard, 2018; van Schijndel and Linzen, 2018; Lakretz et al., 2022; Wilcox et al., 2022), semantics (Herbelot and Ganesalingam, 2013; Erk, 2016; Lenci, 2018; Rambelli et al., 2023) as well as morphology (Cotterell and Schütze, 2015; Kann, 2020; Knigawka, 2022). Results are at present rather mixed on whether RNNs and the representations they learn have human-like properties. Many papers have recently investigated the grammatical abilities of these networks, in particular investigating if such generic models can learn structural properties of language and long-distance relations.

One fruitful line of work aims to correlate ANN-induced representations to linguistic properties, namely to the property that core linguistic rules are structure-dependent. The main phenomenon studied to exemplify this property is subject-verb number agreement. Initial work had shown that ANNs do not really learn the structure-dependency of this construction on their own, but that they can learn to predict English subject-verb agreement if provided with explicit supervision (Linzen et al., 2016). In follow-up work, Bernardy and Lappin (2017) have shown that RNNs are better at modeling long-distance agreement if the model can be trained with a large vocabulary and the rest of the words are replaced by their POS to highlight structural patterns. Both these studies have reported that neural networks are able to perform complex tasks on subject-verb agreement (mainly in English) with the presence of linearly intervening subjects (e.g. *the parents of the student *is/are*). On the other hand, other work on agreement has also shown that neural networks might perform poorly in very rare structures, such as nested agreement dependencies following object relative clause (*The cat that the dogs chase runs/*run*) (Marvin and Linzen, 2018).

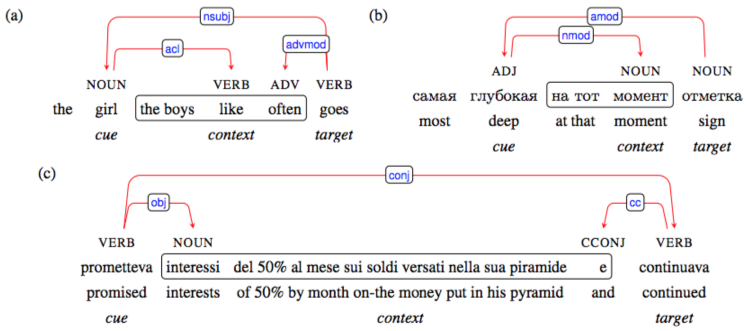

Figure 7: Examples of agreement constructions and "separating"/"intervening" materials in English (a), Russian (b) and Italian (c). From Gulordava et al. 2018, 1197, Figure 1.

While this early work seemed to indicate that RNNs showed shortcomings on core properties of structural and distant grammatical relations, much cited later work has demonstrated that stronger techniques can yield more positive results (Gulordava et al., 2018). Gulordava et al. (2018) explore the LSTMs capacity to track abstract hierarchical structure, by predicting long-distance number agreement in various constructions in four languages (English, Hebrew, Italian, Russian). Some examples are shown in Figure 7. They also show that LSTMs perform well in grammatically well-formed, but non-sensical sentences (in English, Russian, Hebrew, and Italian), whose accuracy is comparable with a control group of human speakers. Their results suggest that neural networks can learn hierarchical grammatical phenomena and not just shallow patterns, despite being trained only to predict the linearly next word.

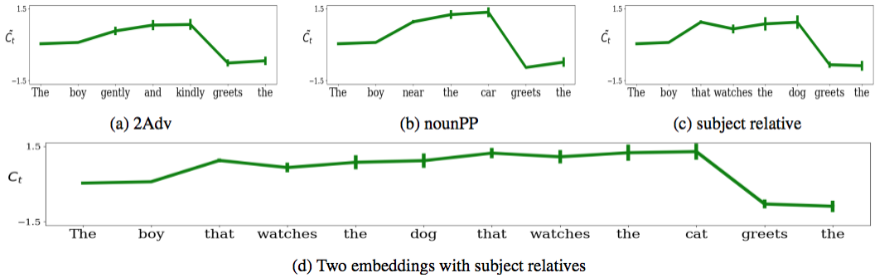

Figure 8: Cell activity of syntax unit while processing various syntactic constructions. Values averaged across all stimuli. Error bars represent standard deviations (details in Lakretz et al. 2019). From Lakretz et al. 2019, Figure 3, p. 17.

These impressive results were obtained by using the neural networks as a 'black box': the algorithm is studied, as we indicated in the introduction, as if it were a natural object whose internal mechanisms cannot be observed. Recent work by Lakretz et al. (2021) studies networks in more detail, as a 'white box', looking at single neurons and thereby getting a glimpse of the internal working of the neural network. They find that individual neurons encode linguistically meaningful features very saliently and their behaviour over time corresponds to the expected propagation of subject-verb number agreement information. These fine-grained results argue in favour of these architectures' ability to learn long agreement patterns. Some examples of the level of activation of relevant neural network units are shown in Figure 8.

Finally, in a recent paper on transformers, Li et al. (2023) investigates deeper representational issues, by contrasting two kinds of agreement, subject-verb agreement and past-participle agreement in French. They argue, based on theoretical accounts, that these superficially similar kinds of agreement, involve in fact very different abstract operations and demonstrate that transformers do reflect this difference in their representations.

Two main observations can be gathered from these works that could be of interest for theoretical syntacticians. First of all, a fundamental outcome is that the neural networks encode the hierarchical structures that are required to correctly perform subject-verb agreement at a distance. They are therefore fundamentally different in their internal representations from simple n-gram language models. Secondly, this hierarchical knowledge emerges from linear operations. These neural networks are trained only to predict the next word or the masked word and are never provided with any explicit hierarchical information. Similarly, hierarchical structures emerge also in non-sensical input (cf. Gulordava et al. 2018). This shows that hierarchical information can emerge by induction in complex internal representations. Consequently, this result is related to the issue of how much data might be required to learn structural configurations, and the debate on the poverty of the stimulus (Chomsky 1965; see Wilcox et al. 2022 for details).

3.2 LSTM and Transformers and locality

As seen in the previous section, the recent widespread and strong interest in ANNs has spurred detailed investigations of the distributed representations they use, learn and generate and specifically if they exhibit properties similar to those characterizing human languages. Besides structure-dependence of rules, like number agreement, another core, defining property of human languages is the property of long-distance dependencies.[10] Human languages exhibit the ability to interpret discontinuous elements distant from each other in the string as if they were adjacent. Similarly to the literature on agreement, probing different aspects of long-distance dependencies, so far divergent results have been reported on these constructions.

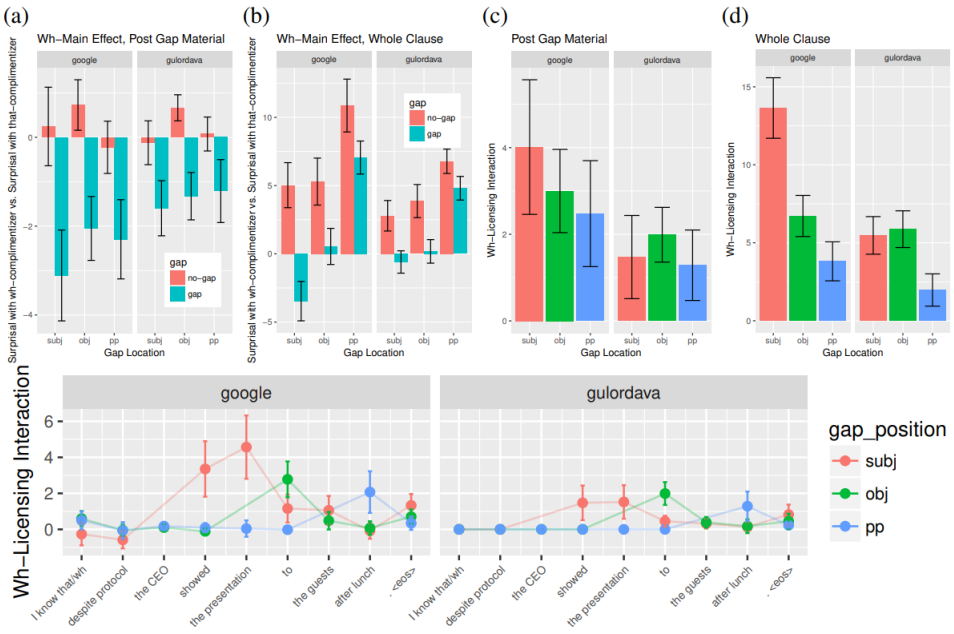

Figure 9: Surprisal (a-d) of wh-licensors by syntactic positions. Bottom chart displays wh-licensing interaction summed across all regions. From Wilcox et al. 2018, 215, Figure 1.

While some experiments have shown that ANNs can learn the main descriptive properties of long-distance dependencies in English (Wilcox et al., 2018, 2022), work attempting to replicate finer-grained human judgments for French have failed (Merlo and Ackermann, 2018), while other work on English has found mixed results (Chowdhury and Zamparelli, 2018).

Wilcox et al. (2018, 2022) investigate whether neural language models like LSTMs learn (English) filler-gap dependencies. They put forward an approach in which they treat neural networks like psycholinguistic subjects, testing them on paradigms developed to encode the properties of long-distance dependencies and measuring their surprisal at predicting the next word given the preceding context. They demonstrate that ANN have acquired this knowledge by assessing their sensitivity to "island" constraints. Specifically, they test separately the properties of filler-gap dependencies, for example, the fact that they obey a uniqueness

[10]To clarify, we distinguish here long dependencies from long-distance dependencies: the term long-distance dependencies is used here in its usual technical sense that refers to discontinuous constructions where two elements in the string receive the same interpretation. Long-distance dependency constructions are *wh*-questions, relative clauses, right-node raising, among others Rimell et al. (2009); Nivre et al. (2010); Merlo (2015a). While the terminology used in the NLP literature is sometimes confusing, it is clear that not all long-distance dependencies are actually long, for example, subject-oriented relative clauses, and not all long dependencies are long-distance dependencies, for example, long subject-verb agreement as studied in Linzen et al. (2016); Bernardy and Lappin (2017); Gulordava et al. (2018) is usually not considered a long-distance dependency.

constraint (only one gap per filler), but also the unboundedness and hierarchical constraints implicated in the dependency. Even more intriguingly, they demonstrate that the model has acquired island constraints by showing that its expectation for a filler-gap dependency is reduced within an island environment.

For example, the examples in (2) act as test sentences that measure whether the uniqueness property of filler-gap dependencies is detected (from Wilcox et al. 2018, 212, ex. 2a-d).

(2) a. I know that the lion devoured a gazelle at sunrise. [no wh-licensor, no gap]
b. *I know what the lion devoured a gazelle at sunrise. [wh-licensor, no gap]
c. *I know that the lion devoured ___ at sunrise. [no wh-licensor, gap]
d. I know what the lion devoured ___ at sunrise. [wh-licensor, gap]

Figure 9 shows the main results of the paper: the main interaction between gapped and non-gapped conditions and the fact that the effect occurs independently of where the gap is located. The bottom panel also very interestingly shows that the increase in surprisal is right where we would expect it in both tested models.

Islands are not the only constraints on long-distance dependencies; other factors can reduce the ability to establish such a relation. For example, as was already discussed in the previous section, it has been shown that this ability is blocked if a similar, but extraneous, element intervenes between the discontinuous components. Under exhaustive and precise conditions, it can be shown that word embeddings and the similarity spaces they define do not encode the properties of intervention similarity in long-distance dependencies and that therefore they fail to represent this core linguistic notion (Merlo and Ackermann, 2018; Merlo, 2019). This result has been obtained with context-free word embeddings, such as word2vec (Mikolov et al., 2013), and with symmetric but also asymmetric similarity operators (such as those illustrated in Henderson and Popa 2016). Other syntactic structures have been explored, such as filler-gap dependencies, auxiliary fronting or other island effects, also with mixed results (Chowdhury and Zamparelli, 2018; Marvin and Linzen, 2018; Warstadt et al., 2019; McCoy et al., 2020).

A positive result comes instead from object *it*-clefts. As discussed in detail in Samo and Merlo (2021) (see previous section), object *it*-clefts represent complex structures that also occur rarely in corpora. Despite their rarity, it can be shown that their corpus counts exhibit cross-linguistic locality effects, because they disfavour matching morpho-syntactic features between the fronted cleft object and the intervening subject (see also section 2.2). This means that in large-scale corpora similarity between the cleft object and the intervening subject is avoided. Thus, we expect that computational models sensitive to these statistics might show a dispreference for matching and a preference for mismatching configurations as predicted from a theory of locality. In a recent paper, Samo and Merlo (2023) tackle this research question with ANNs trained on French. They observe a gradation of surprisal effects that vary with the number of matching features. This shows that the representations of neural network models are sensitive to morpho-syntactic features (type/ number/person and number/gender) in intervention configuration.

In conclusion, these studies show that the internal representations of NN are sensitive to constraints on long-distance dependencies. Following the same line of argument as the literature on agreement discussed above, Wilcox et al. (2022) claim that their results provide empirical evidence against the Argument from the Poverty of the Stimulus for the investigated structures. While the representational abilities of these networks are impressive, the debate is, however, for the moment, still open. It must be recalled that these networks are usually trained on sizes of training material that are several orders of magnitude greater than the data available to a child. Research on scaling down the training needs of neural networks is ongoing and requires investigating methods that learn the underlying rules more explicitly, as we discuss in the next section.

3.3 Modelling the generating principles

The current reported success of machine learning architectures is based on computationally expensive algorithms and prohibitively large amounts of data that are available for only a few, non-representative languages (Bender et al. 2021). They have been shown not to generalise well (Belinkov and Bisk 2018; Belinkov and Glass 2019, see also also Sinha et al. 2021; Wallat et al. 2021; Chaves and Richter 2021). Generalisation in NLP has been traditionally defined in a very narrow way, as extension from a set of data points to new data points of exactly the same nature (i.i.d. assumption) (Schölkopf, 2019). But recent approaches to generalisation have shifted attention to out-of-distribution generalisation, be it across languages, across domains, or new unseen combinatorial tokens and structures Hupkes et al. (2022).

A strong motivation and inspiration is to look at human behaviour on generalisation. It is conjectured that one likely reason why people generalise better is that they have a strong prior bias, grounded in the actual structure of the problem. A large body of literature on experimental work has demonstrated that the human mind is predisposed to extract regularities and generate rules from data, in a way that is distinct from the patterns of activation of neural networks (Sablé-Meyer et al., 2021).

One possible approach to develop more robust methods that require fewer training data, then, is to pay more attention to the decomposition of complex observations, discovering the factors in the generative process that gives rise to the data (Schölkopf et al., 2012). To study how to discover the underlying problem structure, recent research has developed the notion of disentanglement. A disentangled representation can be defined as one where single latent units in the neural network are sensitive to changes in single generative factors, while being relatively invariant to changes in other factors (Bengio et al., 2013).

Novel tasks and curated data for specific linguistic problems have been proposed as a method to learn more disentangled linguistic representations, that reflect the underlying linguistic rules of grammar. To this goal, Merlo et al. (2022); Merlo (2023a,b); Merlo et al. (2023) propose a new linguistic task. The process to solve the task would be very similar to what linguists do when they discover new regularities and new paradigms on novel data descriptions. The Blackbird's language matrices (BLMs) task is a linguistic analog to Raven's visual progressive matrices (Raven 1938, see also Merlo et al. 2023). The multiple-choice task consists in finding the right continuation sentence that completes a sequence of sentences. The sentences are superficially very different but, in fact, they are constructed to be the extensional expression of a linguistic phenomenon and its properties. The data is produced according to a specific generative process. This process controls the compositionality of the data and the expression of the underlying rules, and it manipulates different levels of complexity (e.g. size, lexical variation, structure variation), so that it can be tested how well a system is able to generalize. To generalize, the system must discover the underlying generative rules and learn with different amounts of training data.

The linguistic phenomenon is presented as an incomplete sequence of sentences (*context*), deliberately designed to follow given linguistic rules and given linguistic properties. BLMs correspond to problems where only one answer satisfies the constraints defined by the given context. The construction of BLM datasets is described in more detail in An et al. (2023) (subject-verb agreement in French) and Samo et al. 2023 (verb alternation in English).

For example, Samo et al. (2023) describe the linguistic insights for the creation of the BLM dataset for the English *spray/load* verb alternation (Levin 1993, e.g. *The girl sprays paint onto the wall* vs. *The girl sprays the wall with paint*). This linguistic phenomenon involves an alternation in which a verb belonging to a a selected class (spray/load; see Levin 1993) combines three arguments to describe an event wherein an Agent causes the motion a Theme to a designated Loc(ation). In such alternation both Theme and Loc can represent

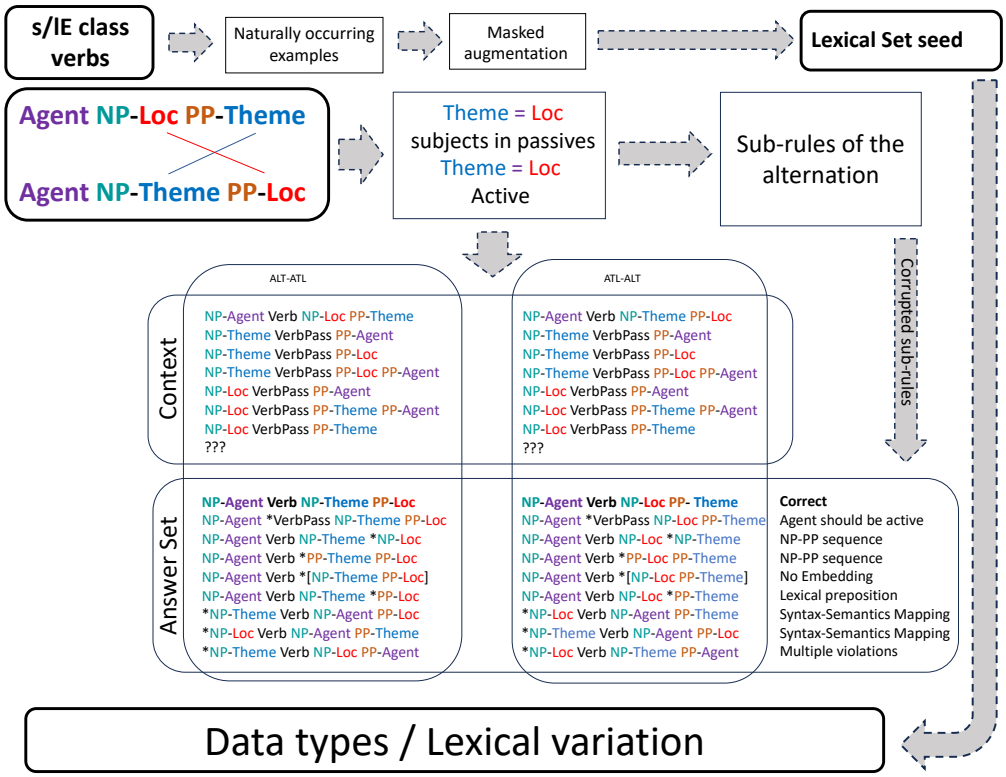

Figure 10: BLM process to construct data to learn the spray/load verb alternation in English (s/lE). The dataset consists of a sequence of sentences with a specific structure as input (see details in Samo et al. 2023). Candidate answers follow a specific structure, each representing a violation of one of the (sub-)sets of rules.

the direct object of the syntactic architecture. In one alternant, the THEME constitutes the direct object of the verb while LOC is preceded by a preposition (PP) (*The girl$_{Agent}$ sprays paint$_{Theme}$ onto the wall$_{Loc}$*). In the other alternant, LOC represents the direct object while the THEME is a PP (*The girl$_{Agent}$ sprays the wall$_{Loc}$ with paint$_{Theme}$*).

The BLM dataset presented in Samo et al. (2023) exposes the model to common morpho-syntactic properties of the arguments THEME and LOC. For instance, both arguments can be passivized, suggesting that they are potential direct objects of the verb. To effectively learn this alternation, the model needs to be capable of generalizing from shared,common properties.

The construction of the dataset is illustrated in Figure 10. In this task, an instance consists of sequences of sentences with specific attributes. To predict the correct answer as the next element of the sequence, a model must correctly detect the generative model used to produce the dataset. Two exploratory baselines based on commonly used architectures show that despite the simplicity of the phenomenon, it is a complex problem for deep learning systems, and in particular that even the correct solutions do not manage to identify the underlying rules (i..e the models learned are not disentangled).

The BLM task has revealed that correct predictions in previously studied linguistic problems do not yet stem from a deep understanding of the generative factors that define these problems. For example, recall that previous work on agreement, described in section 3.2, has demonstrated that RNN capture the phenomenon of subject-verb agreement well, reaching in some languages human-level accuracy in predicting the correct verb form (Linzen et al. 2016; Gulordava et al. 2018; see Linzen and Baroni 2021 for an overview). However, in their exploration of the BLM methodology, An et al. (2023) discover that when more complex predictions are introduced, which require understanding of the factors that compose the subject-verb agreement rule, the results are much less compelling.

Thus, this work demonstrates that this language task and the data that instantiate it provide a new challenging testbed to understand generalization and abstraction on those problems that looked solved in the black box approaches described above.

3.4 Conclusions

Current technological advances in generative NLP, such as ChatGPT and other language models, have captured the attention of the world, and alerted the lay user to the great capabilities of NLP tools and models. Beyond these fascinating technological successes lies an uncharted land of investigation on the inner workings of these models, their extension to more and more languages with fewer and fewer available data, models of language acquisition, comparison to human neural data, integration of pragmatics and reasoning and language use, multimodality and many other areas of investigation where linguistic theories and linguistic knowledge will be crucial to separate the superficial solutions that mimic input data from more human-like solutions where true structure and meaning emerge from simple exposure to language stimuli.

4 Conclusions

This chapter has presented several methods and discussed issues related to the theoretically-inspired computational modelling of language. We believe that computational modelling is an important development in the study of language that has great capabilities and potential, and that, we conjecture, once adopted will very rapidly spread in the linguistic community.

We have organised the chapter in a progression of induction of both the model structure and the values of the models parameters. In minimalist terms, we view this progression as an

increase in ability to develop complex mechanistic explanations of the induction of internal I-language representations from externalised and estensional E-xpressions of language.

We have discussed models created by hand. These models need to be built with prior knowledge and assumptions drawn from theoretical linguistics. This type of modelling leads us to gain insights into syntactic variability at the E-language level and test the abstract rules and principles that govern the dimension of the I-language, such as the simplest operation of *move* (see for example the vectorization of Universal 20, in section 2.1). These studies also provide quantitatively motivated measures of the learnability of different theories on a specific phenomenon, contributing to model comparison and model selection. Specifically, the concept of learnability in simulation-based research is strictly connected to the notion of prediction, for example, in terms of a prediction of class-membership (e.g. defined typological class, but also classes of frequencies) in theory-driven counting models (section 2; Merlo 2015b; Merlo and Ouwayda 2018).

The notion of prediction and its applications also represent a core element in large language models (section 3; see also Mahowald et al. 2023, 8-12). Artificial neural networks learn patterns of information from large amounts of data, defining attributes and their values based on the E-language data they are exposed to. Theoretical studies can provide very valuable information in their evaluation. As the results emerging from the works discussed in section 3 (cf. Gulordava et al. 2018; Wilcox et al. 2022 *inter alia*), ANNs and large language models, in general, can mark asymmetries between minimal pairs of sentences (for example, sentences with and without gaps), implicitly signalling differences to a deeper level (e.g. movement). Importantly, these models are shown not to depend on pure linear configuration. In parallel with human language, these models do not rely on what can be labelled as "the simplest algorithm" (Chomsky 2023, 353). The exploration of these models can tease apart the rules underlying the output asymmetries (cf. subsection 3.3.). From a syntactician's perspective, language models trained with different datasets, properly labelled, might represent a "population of speakers" acting as participants of "psycholinguistic" experiments.

Crucially, these models still make a fundamental distinction between the Extensional, observational language they are trained on and the Internal representations that are induced. These internal abstract representations comprise two distinct conceptual aspects: on the one hand, the predefined architecture, which is generic across tasks, language phenomena and language diversity; on the other hand, the specific, learnt abstract representation. [11]

Far from being the purely statistically-driven models that detractors claim them to be, these complex language models provide us with thought-provoking mechanistic explanations of how fundamental properties of language can be learnt, and how grammars can be organised. For example, transformers give us a glimpse of how simple operations, such as language modelling, that is simply predicting the next word, can interact with complex a priori architectures to solve fundamental learning and processing problems, such as object induction or soft representations of long-distance dependencies in a distributed representation. They also show, so far poorly understood, emerging properties of zero-shot learning, within and across languages.

We think that future avenues of research that will bring better convergence towards human-like models lie in reducing data needs through the development of dedicated curated data and tasks for the exploration of how to induce rule-like behaviour from distributed representations.

[11] In this respect, we entirely agree with the point of view that claims that the architectural primitives of the model are to be considered the 'theory' (Baroni, 2022, 2). Following Baroni (2022, 7) "It is more appropriate, instead, to look at deep nets as linguistic theories, encoding non-trivial structural priors facilitating language acquisition and processing. More precisely, we can think of a deep net architecture, before any language-specific training, as a general theory defining a space of possible grammars, and of the same network trained on data from a specific language as a grammar, that is, a computational system that, given an input utterance in a language, can predict whether the sequence is acceptable to an idealized speaker of the language (e.g., Chomsky 1986b; Sag et al. 2003; Müller 2020" (page 7)

References

Abend, O., R. Reichart, and A. Rappoport (2008). A supervised algorithm for verb disambiguation into verbnet classes. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pp. 9–16.

An, A., C. Jiang, M. A. Rodriguez, V. Nastase, and P. Merlo (2023, May). BLM-AgrF: A new French benchmark to investigate generalization of agreement in neural networks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia, pp. 1363–1374. Association for Computational Linguistics.

Baker, M. C. (2002). *The atoms of language*. New York: Basic Books.

Barbiers, S., H. Bennis, G. De Vogelaer, M. Devos, and M. van der Ham (2005). *Syntactische Atlas van de Nederlandse Dialecten, Deel I.* Amsterdam: Amsterdam University Press.

Barbiers, S., H. Bennis, and L. Dros-Hendriks (2018). Merging verb cluster variation. *Linguistic Variation 18*(1), 144–196.

Baroni, M. (2022). On the proper role of linguistically oriented deep net analysis in linguistic theorising. In S. Lappin and J.-P. Bernardy (Eds.), *Algebraic Structures in Natural Language*, pp. 1–16. CRC Press.

Baroni, M. and A. Lenci (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics 36*(4), 673–721.

Belinkov, Y. and Y. Bisk (2018). Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations (ICLR)*.

Belinkov, Y. and J. Glass (2019). Analysis methods in neural language processing: A survey. *Transaction of the ACL 7*, 49–72.

Belletti, A., N. Friedmann, D. Brunato, and L. Rizzi (2012). Does gender make a difference? Comparing the effect of gender on children's comprehension of relative clauses in Hebrew and Italian. *Lingua 122*(10), 1053–1069.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell (2021). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623.

Bengio, Y., A. Courville, and P. Vincent (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence 35*(8), 1798–1828.

Bernardy, J.-P. and S. Lappin (2017). Using deep neural networks to learn syntactic agreement. *Linguistic Issues in Language Technology 15*(2), 1–15.

Biberauer, T. and I. Roberts (2017). Parameter setting. In A. Ledgeway and I. Roberts (Eds.), *The Cambridge Handbook of Historical Syntax*, pp. 134–162. Cambridge: Cambridge University Press.

Burnett, H., H. Koopman, and S. A. Tagliamonte (2018). Structural explanations in syntactic variation: The evolution of english negative and polarity indefinites. *Language Variation and Change 30*(1), 83–107.

Bybee, J. (2007). *Frequency of Use and the Organisation of Language*. Oxford University Press.

Ceolin, A., C. Guardiano, M. A. Irimia, and G. Longobardi (2020). Formal syntax and deep history. *Frontiers in psychology 11*, 488871.

Chaves, R. P. and S. N. Richter (2021). Look at that! bert can be easily distracted from paying attention to morphosyntax. *Proceedings of the Society for Computation in Linguistics 4*(1), 28–38.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1973). Conditions on transformations. In S. R. Anderson and P. Kiparsky (Eds.), *A festschrift for Morris Halle*. New York: Holt, Rinehart and Winston.

Chomsky, N. (1986a). *Barriers*, Volume 13. MIT Press (MA).

Chomsky, N. (1986b). *Knowledge of language: Its nature, origin, and use*. Westport, CT: Praeger.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Chomsky, N. (2023). Genuine explanation and the strong minimalist thesis. *Cognitive Semantics 8*(3), 347 – 365.

Chowdhury, S. A. and R. Zamparelli (2018). RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING'18)*, pp. 133–144. Association for Computational Linguistics.

Cinque, G. (2005). Deriving Greenberg's universal 20 and its exceptions. *Linguistic Inquiry 36*(3), 315–332.

Cinque, G. (2013). *Typological Studies: Word Order and Relative Clauses*. New York/London: Routledge.

Cotterell, R. and H. Schütze (2015). Morphological word-embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Denver, Colorado, pp. 1287–1292. Association for Computational Linguistics.

Crisma, P., C. Guardiano, and G. Longobardi (2020). Syntactic diversity and language learnability. *Studi e Saggi Linguistici 58*, 99–130.

Culbertson, J. and P. Smolensky (2012). A Bayesian model of biases in artificial language learning: The case of a word-order universal. *Cognitive Science*, 1–31.

Culbertson, J., P. Smolensky, and G. Legendre (2012). Learning biases predict a word order universal. *Cognition*, 306–329.

Cysouw, M. (2010). Dealing with diversity: towards an explanation of NP word order frequencies. *Linguistic Typology 14*(2), 253–287.

Diessel, H. and M. Hilpert (2016). Frequency effects in grammar. In *Oxford research encyclopedia of linguistics*.

Dryer, M. S. (1992). The Greenbergian word order correlations. *Language 68*, 81–138.

Dryer, M. S. (2006). The order demonstrative, numeral, adjective and noun: an alternative to Cinque. http://exadmin.matita.net/uploads/pagine/1898313034_cinqueH09.pdf.

Dryer, M. S. (2009). The branching direction theory of word order correlations revisited. In *Universals of language today*, pp. 185–207. Springer.

Erk, K. (2016). What do you know about an alligator when you know the company it keeps? *Semantics and Pragmatics 9*(17), 1–63.

Fedzechkina, M., T. F. Jaeger, and E. L. Newport (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences 109*(44), 17897–17902.

Friedmann, N., A. Belletti, and L. Rizzi (2009). Relativized relatives: Types of intervention in the acquisition of A-bar dependencies. *Lingua 119*(1), 67–88.

Futrell, R., R. Levy, and M. Dryer (2017). A statistical comparison of some theories of np word order. *arXiv preprint arXiv:1709.02783*.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition 68*(1), 1–76.

Grillo, N. (2008). *Generalized Minimality*. Ph. D. thesis, Utrecht Institute of Linguistics, OTS.

Guardiano, C., P. Crisma, G. Longobardi, and G. Cordoni (2020). Formal syntax as a phylogenetic method. In R. D. Janda, B. D. Joseph, and B. S. Vance (Eds.), *The Handbook of Historical Linguistics, Volume II*, pp. 145–182. Hoboken: Wiley/Blackwell Publishers.

Guardiano, C. and G. Longobardi (2016). Parameter Theory and Parametric Comparison. In I. Roberts (Ed.), *The Oxford Handbook of Universal Grammar*, pp. 377–398. Oxford University Press.

Guest, O. and A. E. Martin (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science 16*(4), 789–802.

Gulordava, K., P. Bojanowski, E. Grave, T. Linzen, and M. Baroni (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1195–1205. Association for Computational Linguistics.

Gulordava, K. and P. Merlo (2020). Computational quantitative syntax. In *Romance Languages and Linguistic Theory 16: Selected papers from the 47th Linguistic Symposium on Romance Languages (LSRL), Newark, Delaware*, Volume 16, pp. 109. John Benjamins Publishing Company.

Haspelmath, M. (2006). Against markedness (and what to replace it with). *Journal of linguistics 42*(1), 25–70.

Hawkins, J. A. (1994). *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.

Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Henderson, J. (2020). The unstoppable rise of computational linguistics in deep learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 6294–6306. Association for Computational Linguistics.

Henderson, J. and F. J. Fehr (2023). A VAE for transformers with nonparametric variational information bottleneck. In *The Eleventh International Conference on Learning Representations*.

Henderson, J. and D. Popa (2016). A vector space for distributional semantics for entailment. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, pp. 2052–2062. Association for Computational Linguistics.

Herbelot, A. and M. Ganesalingam (2013). Measuring semantic content in distributional vectors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Sofia, Bulgaria, pp. 440–445. Association for Computational Linguistics.

Hochreiter, S. and J. Schmidhuber (1997). Long short-term memory. *Neural computation 9*(8), 1735–1780.

Hupkes, D., M. Giulianelli, V. Dankers, M. Artetxe, Y. Elazar, T. Pimentel, C. Christodoulopoulos, K. Lasri, N. Saphra, A. Sinclair, D. Ulmer, F. Schottmann, K. Batsuren, K. Sun, K. Sinha, L. Khalatbari, M. Ryskina, R. Frieske, R. Cotterell, and Z. Jin (2022). State-of-the-art generalisation research in NLP: a taxonomy and review. *CoRR*.

Ibbotson, P. (2013). The scope of usage-based theory. *Frontiers in psychology 4*, 255.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société vaudoise des sciences naturelles 37*, 547–579.

Jurafsky, D. and J. H. Martin (2023). Speech and language processing (3rd ed. draft).

Kann, K. (2020). Acquisition of inflectional morphology in artificial neural networks with prior knowledge. In *Proceedings of the Society for Computation in Linguistics 2020*, New York, New York, pp. 144–154. Association for Computational Linguistics.

Kazakov, D., G. Cordoni, A. Ceolin, M.-A. Irimia, S.-S. Kim, D. Michelioudakis, N. Radkevich, C. Guardiano, and G. Longobardi (2017). Machine learning models of universal grammar parameter dependencies. In *Proceedings of the Workshop Knowledge Resources for the Socio-Economic Sciences and Humanities associated with RANLP 2017*, Varna, pp. 31–37. INCOMA Inc.

Kingma, D. P., T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling (2016). Improved variational inference with inverse autoregressive flow. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 29. Curran Associates, Inc.

Kingma, D. P., M. Welling, et al. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning 12*(4), 307–392.

Knigawka, Ł. (2022). Constructing a derivational morphology resource with transformer morpheme segmentation. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, Potsdam, Germany, pp. 104–109. KONVENS 2022 Organizers.

Koslow, S. H. and M. F. Huerta (2013). *Neuroinformatics: an overview of the human brain project*. Psychology Press.

Lakretz, Y., T. Desbordes, D. Hupkes, and S. Dehaene (2022). Can transformers process recursive nested constructions, like humans? In *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea, pp. 3226–3232. International Committee on Computational Linguistics.

Lakretz, Y., D. Hupkes, A. Vergallito, M. Marelli, M. Baroni, and S. Dehaene (2021). Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*.

Lakretz, Y., G. Kruszewski, T. Desbordes, D. Hupkes, S. Dehaene, and M. Baroni (2019). The emergence of number and syntax units in LSTM language models. 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)*, Minneapolis, Minnesota, pp. 11–20. Association for Computational Linguistics.

Lau, J. H., A. Clark, and S. Lappin (2017). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science 41*(5), 1202–1241.

Lenci, A. (2018). Distributional models of word meaning. *Annual review of Linguistics 4*, 151–171.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Li, B., G. Wisniewski, and B. Crabbé (2023). Assessing the capacity of transformer to abstract syntactic representations: A contrastive analysis based on long-distance agreement. *Transactions of the Association for Computational Linguistics 11*, 18–33.

Lin, Z., G. I. Winata, P. Xu, Z. Liu, and P. Fung (2020). Variational transformers for diverse response generation. *CoRR abs/2003.12738*.

Linzen, T. and M. Baroni (2021). Syntactic structure from deep learning. *Annual Review of Linguistics 7*(1), 195–212.

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

Linzen, T. and B. Leonard (2018). Distinct patterns of syntactic agreement errors in recurrent networks and humans. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Longobardi, G. (2003). Methods in parametric linguistics and cognitive history. In P. Pica and J. Rooryck (Eds.), *Linguistic Variation Yearbook*, pp. 101–138. Amsterdam: John Benjamins.

Longobardi, G. (2018). Principles, parameters, and schemata: A radically underspecified ug. *Linguistic Analysis*, 517–558.

Mahowald, K., A. A. Ivanova, I. A. Blank, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko (2023). Dissociating language and thought in large language models: a cognitive perspective.

Manzini, M. R. and L. M. Savoia (2005). *I dialetti Italiani e Romanci. Morfosintassi generativa.* Alessandria: Edizioni dell'Orso.

Marvin, R. and T. Linzen (2018). Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1192–1202. Association for Computational Linguistics.

McCoy, R. T., E. Grant, P. Smolensky, T. L. Griffiths, and T. Linzen (2020). Universal linguistic inductive biases via meta-learning. *arXiv preprint arXiv:2006.16324*.

Merlo, P. (1994). A corpus-based analysis of verb continuation frequencies for syntactic processing. *Journal of Psycholinguistic Research 23*, 435–457.

Merlo, P. (2015a). Evaluation of two-level dependency representations of argument structure in long-distance dependencies. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pp. 221–230.

Merlo, P. (2015b). Predicting word order universals. *Journal of Language Modelling 3*(2), 317–344.

Merlo, P. (2016). Quantitative computational syntax: some initial results. *IJCoL. Italian Journal of Computational Linguistics 2*(2-1).

Merlo, P. (2019). Probing word and sentence embeddings for long-distance dependencies effects in French and English. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Florence. Association for Computational Linguistics.

Merlo, P. (2023a). Blackbird language matrices (BLM), a new task for rule-like generalization in neural networks: Can Large Language Models pass the test? In *Findings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Merlo, P. (2023b). Blackbird language matrices (blm), a new task for rule-like generalization in neural networks: Motivations and formal specifications. *ArXiv cs.CL 2306.11444*.

Merlo, P. and F. Ackermann (2018). Vectorial semantic spaces do not encode human judgments of intervention similarity. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018*, Brussels, Belgium, pp. 392–401.

Merlo, P., A. An, and M. A. Rodriguez (2022). Blackbird's language matrices (blms): a new benchmark to investigate disentangled generalisation in neural networks. *arXiv2205.10866*.

Merlo, P., C. Jiang, G. Samo, and V. Nastase (2023). Blackbird Language Matrices Tasks for Generalization. In *Proceedings of the GenBench workshop*.

Merlo, P. and S. Ouwayda (2018). Movement and structure effects on universal 20 word order frequencies: A quantitative study. *Glossa: a journal of general linguistics 3*(1).

Merlo, P. and G. Samo (2022). Exploring T3 languages with quantitative computational syntax. *Theoretical Linguistics 48*(1-2), 73–83.

Merlo, P., G. Samo, V. Nastase, and C. Jiang (2023). Building structured synthetic datasets: The case of Blackbird Language Matrices (BLMs). In *Proceedings of the Ninth Italian Conference on Computational Linguistics (Clic-It 2023)*.

Merlo, P. and S. Stevenson (2001). Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics 27*(3), 373–408.

Merlo, P., S. Stevenson, V. Tsang, and G. Allaria (2002). A multi-lingual paradigm for automatic verb classification. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, PA, pp. 207–214.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. *CoRR abs/1301.3781*.

Müller, S. (2020). *Grammatical theory:: From transformational grammar to constraint-based approaches*. Berlin: Language Science Press.

Nivre, J., L. Rimell, R. McDonald, and C. Gómez Rodríguez (2010). Evaluation of dependency parsers on unbounded dependencies. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, Beijing, China, pp. 833–841.

Palmer, M., D. Gildea, and P. Kingsbury (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics 31*, 71–105.

Pescarini, D. (2019). Microvariation and microparameters. some quantitative remarks. *Quaderni di Linguistica e Studi Orientali 5*, 255–277.

Pescarini, D. (2022). A quantitative approach to microvariation: negative marking in central romance. *Languages 7*(2), 87.

Rambelli, G., E. Chersoni, M. S. G. Senaldi, P. Blache, and A. Lenci (2023). Are frequent phrases directly retrieved like idioms? an investigation with self-paced reading and language models. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, Dubrovnik, Croatia, pp. 87–98. Association for Computational Linguistics.

Raven, J. C. (1938). Standardization of progressive matrices. *British Journal of Medical Psychology 19*, 137–150.

Reali, F. and M. H. Christiansen (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language 57*(1), 1–23.

Rimell, L., S. Clark, and M. Steedman (2009). Unbounded dependency recovery for parser evaluation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, pp. 813–821. Association for Computational Linguistics.

Rizzi, L. (1990). *Relativized minimality.* The MIT Press.

Rizzi, L. (2004). Locality and left periphery. In A. Belletti (Ed.), *Structures and beyond: The cartography of syntactic structures*, Volume 3, pp. 223–251. Oxford, New York: Oxford University Press Oxford.

Rizzi, L. (2013). Locality. *Lingua 130*, 169–186.

Ross, J. R. (1967). *Constraints on variables in syntax.* ERIC.

Sablé-Meyer, M., J. Fagot, S. Caparos, T. van Kerkoerle, M. Amalric, and S. Dehaene (2021). Sensitivity to geometric shape regularity in humans and baboons: A putative signature of human singularity. *Proceedings of the National Academy of Sciences 118*(16).

Sag, I. A., T. Wasow, and E. Bender (2003). *Syntactic Theory: a formal introduction* (Second Edition ed.). University of Chicago Press.

Samardžić, T. and P. Merlo (2018). Probability of external causation: An empirical account of cross-linguistic variation in lexical causatives. *Linguistics 56*(5), 895–938.

Samo, G. and P. Merlo (2019). Intervention effects in object relatives in english and italian: a study in quantitative computational syntax. In *Proceedings of SyntaxFest*, Paris,France.

Samo, G. and P. Merlo (2021). Intervention effects in clefts: a study in quantitative computational syntax. *Glossa: a journal of general linguistics 6*(1).

Samo, G. and P. Merlo (2023). Distributed computational models of intervention effects: a study on cleft structures in French. In C. Bonan and A. Ledgeway (Eds.), *It-clefts: Empirical and Theoretical Surveys and Advances*. De Gruyter.

Samo, G., V. Nastase, C. Jiang, and P. Merlo (2023). Blm-s/le: A structured dataset of english spray-load verb alternations for testing generalization in llms. In *Findings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Sanfelici, E., I. Caloi, and C. Poletto (2014). Subject object asymmetries in relative clauses: An investigation into three new empirical domains. *Quaderni di lavoro ASIt n 18*, 127–160.

Schulte im Walde, S. (2006). Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics 32*(2), 159–194.

Schölkopf, B. (2019). Causality for machine learning. Technical report, arXiv:1911.10500v2.

Schölkopf, B., D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij (2012). On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK.

Sheehan, M., T. Biberauer, I. Roberts, and A. Holmberg (2017). *The final-over-final condition: A syntactic universal*, Volume 76. MIT Press.

Sinha, K., R. Jia, D. Hupkes, J. Pineau, A. Williams, and D. Kiela (2021). Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv preprint arXiv:2104.06644* .

Starke, M. (2001). *Move dissolves into Merge: A theory of locality*. Ph. D. thesis, Unviersity of Geneva.

Steedman, M. (2011). Greenberg's 20th: The view from the long tail. unpublished manuscript, University of Edinburgh.

Tily, H., M. Frank, and F. Jaeger (2011). The learnability of constructed languages reflects typological patterns. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pp. 1364–1369.

Travis, L. (1984). *Parameters and Effects of Word Order Variation*. Ph. D. thesis, MIT, Cambridge, MA.

Vahdat, A. and J. Kautz (2020). Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems 33*, 19667–19679.

van Craenenbroeck, J. and M. van Koppen (forthcoming). Quantitative approaches to syntactic variation. In N. C. Sjef Barbiers and M. Polinsky (Eds.), *The Cambridge Handbook of Comparative Syntax*. Cambridge: Cambridge University Press.

Van Craenenbroeck, J., M. van Koppen, and A. van den Bosch (2019). A quantitative-theoretical analysis of syntactic microvariation: Word order in dutch verb clusters. *Language 95*(2), 333–370.

van Schijndel, M. and T. Linzen (2018). Modeling garden path effects without explicit hierarchical syntax. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 30. Curran Associates, Inc.

Villata, S., L. Rizzi, and J. Franck (2016). Intervention effects and relativized minimality: New experimental evidence from graded judgments. *Lingua 179*, 76–96.

Wallat, J., J. Singh, and A. Anand (2021). Bertnesia: Investigating the capture and forgetting of knowledge in bert. *arXiv preprint arXiv:2106.02902*.

Warstadt, A., Y. Cao, I. Grosu, W. Peng, H. Blix, Y. Nie, A. Alsop, S. Bordia, H. Liu, A. Parrish, S.-F. Wang, J. Phang, A. Mohananey, P. M. Htut, P. Jeretic, and S. R. Bowman (2019). Investigating BERT's knowledge of language: Five analysis methods with NPIs. 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)*, Hong Kong, China, pp. 2877–2887. Association for Computational Linguistics.

Wilcox, E., R. Levy, T. Morita, and R. Futrell (2018). What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 211–221. Association for Computational Linguistics.

Wilcox, E. G., R. Futrell, and R. Levy (2022). Using computational models to test syntactic learnability. *Linguistic Inquiry*, 1–88.

Yang, C. (2003). *Knowledge and Learning in Natural Language*. Oxford University Press.

Yang, C. (2015). For and against frequencies. *Journal of Child Language 42*(2), 287–293.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Boston: Addison-Wesley.