Could Machine Learning Shed Light on Natural Language Complexity?

Leonor Becerra-Bonache
Laboratoire Hubert Curien
Jean Monnet University
Saint-Etienne, France
leonor.becerra@univ-st-etienne.fr

M. Dolores Jiménez-López
Research Group on Mathematical Linguistics
Universitat Rovira i Virgili
Tarragona, Spain
mariadolores.jimenez@urv.cat

Abstract

In this paper, we propose to use a subfield of machine learning –grammatical inference– to measure linguistic complexity from a developmental point of view. We focus on relative complexity by considering a child learner in the process of first language acquisition. The relevance of grammatical inference models for measuring linguistic complexity from a developmental point of view is based on the fact that algorithms proposed in this area can be considered computational models for studying first language acquisition. Even though it will be possible to use different techniques from the field of machine learning as computational models for dealing with linguistic complexity –since in any model we have algorithms that can learn from data–, we claim that grammatical inference models offer some advantages over other tools.

1 Introduction

Complexity has become an important concept in several scientific disciplines (biology, physics, chemistry, philosophy, psychology and sociology) (Mitchell, 2009). There has been a lot of research on complexity and complex systems in the natural sciences, economics, social sciences and, now, also increasingly in linguistics. From McWhorther's (2001) pioneering work, there have been many seminars, conferences, articles, monographs (Dahl, 2004; Kusters, 2003) and collective volumes (Miestamo et al., 2008; Sampson et al., 2009; Newmeyer and Preston, 2014) that have dealt with linguistic complexity and have challenged the so-called *equi-complexity dogma*. In fact, we can say that, nowadays, complexity figures prominently in linguistics.

However, despite the interest it has generated, there is no agreement in the literature on the definition of *complexity*. In a recent article, Pallotti (2015) underlines the polysemy of the term complexity in the linguistic literature and summarizes the different notions of complexity in this field by referring to three main meanings:

- *Structural complexity*, a formal property of texts and linguistic systems having to do with the number of their elements and their relational patterns.
- *Cognitive complexity*, having to do with the processing costs associated with linguistic structures.
- *Developmental complexity*, the order in which linguistic structures emerge and are mastered in second (and, possibly, first) language acquisition.

The above three meanings cover the two conceptions that, according to Crystal (1997), the concept has in linguistics, where 'complexity refers to both the internal structuring of linguistic units and psychological difficulty in using or learning them'. This distinction is directly reflected in the two main types of complexity found in the literature (Miestamo, 2006; Miestamo, 2009a; Miestamo, 2009b):

- The *absolute complexity* approach that defines complexity as an objective property of the system and it is measured in terms of the number of parts of the system, the interrelations between the parts

Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity,
pages 1–11, Osaka, Japan, December 11-17 2016.

or the length of the description of the phenomenon. It is a usual complexity notion in cross-linguistic typology studies (McWhorter, 2001; Dahl, 2004).

- The *relative complexity* approach that takes into account the users of language and identifies complexity with difficulty/cost of processing, learning or acquisition. This type of approach is very common in the fields of sociolinguistics and psycholinguistics (Kusters, 2003).

Differences in the definition led to abundance of complexity measures. Edmonds (1999), for example, identifies forty-eight different metrics used in natural and social sciences. In linguistics, there is no conventionally agreed metric for measuring the complexity of natural languages. The measures proposed are varied and could be grouped into two blocs:

- *Measures of absolute complexity*. The number of categories or rules, length of the description, ambiguity, redundancy, etc. (Miestamo, 2009a).

- *Measures of relative complexity*. When the relative approach of complexity is adopted the problem that has to be faced is the answer to the question: Difficult/costly to whom? This means that it is necessary to determine what kind of task –learning, acquisition, processing– must be considered and, of course, what type of user must be taken into account –speaker, listener, child, adult–. The problem, here, is to choose a task and a type of user representative for the definition of complexity and to properly motivate this choice. The complexity of L2 learning (Kusters, 2003) or the complexity of processing (Hawkins, 2009) are examples of measures proposed in this field.

Some researchers have attempted to apply the concept of complexity used in other disciplines in order to find useful tools to calculate linguistic complexity. Information theory, for example, offers two formalisms that might be appropriate for measuring linguistic complexity:

- *Shannon information entropy* that captures the average number of bits of information necessary to specify the state of a random variable or system described by a probability model (Bane, 2008).

- *Kolmogorov complexity* that measures the informativeness of a string as the length of the algorithm required to describe that string (Juola, 2009). This measure can be applied to measuring language complexity in such a way that the longer the description of a linguistic structure, the more complex it is (Dahl, 2004; Juola, 2009; Miestamo, 2009a).

Other than information theory, computational models (Blache, 2011), or the theory of complex systems (Andrason, 2014) are examples of areas that provide measures to quantitatively evaluate linguistic complexity.

In this paper, we propose to use a subfield of Machine Learning –*Grammatical Inference*– to measure linguistic complexity from a developmental point of view. We focus on relative complexity by considering a child learner in the process of first language acquisition. Therefore, we need a computational model for first language acquisition. This is why we have chosen grammatical inference models, since they deal with idealized learning procedures for acquiring grammars on the basis of exposure to evidence about languages (D'Ulizia et al., 2011).

The paper is organized as follows. Firstly, we present an overview of research on linguistic relative complexity. Secondly, we briefly discuss the relevance of grammatical inference models for explaining natural language acquisition. Thirdly, we present different grammatical inference algorithms that may deal with linguistic complexity. Finally, we conclude with some remarks and possible directions for future work.

2 Relative Complexity

Even though complexity is a central notion in linguistics, until recently, it has not been widely researched in the area (Sinnemäki, 2011). During the twentieth century, linguistic complexity was supposed to be invariant. Linguists, from very different theoretical schools, have agreed that all natural languages must

be equally complex. However, the validity of this claim has rarely been subjected to systematic cross-linguistic investigation.

In the last fifteen years, the interest on linguistic complexity has led researchers to challenge the equi-complexity dogma by addressing the study of complexity from different points of views. In general, recent work on language complexity takes an *absolute* perspective of the concept while the *relative* complexity approach –even though considered as conceptually coherent– has hardly begun to be developed.

In general, researchers agree that it is more feasible to approach complexity from an objective or theory-oriented viewpoint than from a subjective or user-related viewpoint. To approach complexity from the relative point of view constraints the researcher to face many problems:

- What does complex mean?: More difficult, more costly, more problematic, more challenging?
- Different situations of language use (speaking, hearing, L1 acquisition, L2 learning) differ as to what is difficult and what is easy.
- Some linguistic phenomena can be difficult for a certain group of language users while facilitating the task of another group, so we have to answer the question 'complex to whom?'
- A user-based approach would require focusing on one user-type over the others or defining an idealized user-type. How do we decide which type of language use (and user) is primary?

According to Miestamo (2006), there will always be some conflict between definitions of complexity based on different types of users, and no general user-type-neutral definition is possible. This is problematic for a relative approach to complexity. Absolute definitions of complexity avoid these problems.

Among the relative complexity metrics that have been proposed, we can refer to the following ones: L2 acquisition complexity (Trudgill, 2001a); redundancy-induced complexity (Trudgill, 1999; McWhorter, 2001); irregularity-induced complexity (Trudgill, 2001b; McWhorter, 2012); incomplete dependency hypothesis (Gibson, 1998); dependency locality theory (DLT) (Gibson, 2000); structural depth (Ferreira, 1991; Abney and Johnson, 1991; Schuler, 2009); time of acquisition and degree of acquisition, etc.

Studies that have adopted a relative complexity approach have showed some preferences for L2 learners (Kusters, 2003). This is, from the following three different questions that could be answered, researchers have preferably chosen the first one:

1. Second-language learning. Do some language take longer for the adult learner to learn than others?
2. Language use. Are some languages more difficult to use than others?
3. Language acquisition: Do some languages take longer for the child to acquire than others?

However, as pointed out by Miestamo (2006), if we aim to reach a general definition of relative complexity, the primary relevance of L2 learners is not obvious. In fact, they could be considered the least important of the four possible groups that may be considered –speakers, hearers, L1 learners, L2 learners.

Taking into account that some of the ideas that backup the equi-complexity dogma are based on the process of language acquisition –learning a first language is something every child does successfully in every society, in every language, independently of the type of education and intelligence level; all children acquire language in the same way, regardless of the language they learn; children progress through the same stages in language acquisition regardless the language–, we think that studies on developmental complexity may check differences among languages by considering child first language acquisition. Due to the problems that methods for studying language acquisition (observational and experimental) may set out to the study of linguistic complexity, we defend that computational modeling of the process of language acquisition may be considered an important complementary tool that -by avoiding practical problems of analyzing authentic learner productions data– will make possible to consider children (or their simulation) as suitable candidates for evaluating the complexity of languages.

3 Grammatical Inference Relevance for Natural Language Acquisition

Within the field of Machine Learning (Olivas et al., 2009) –that focus on the development of techniques that allow computers to learn–, *Grammatical Inference* (GI) deals with the learning of grammars and languages from data (de la Higuera, 2010). This subfield of machine learning was born in the 1960s and since then has attracted the attention of researchers working on different fields (formal languages, automata theory, computational linguistics, information theory, pattern recognition, and many others).

The relevance of GI models for measuring linguistic complexity from a developmental point of view is based on the fact that the computational models developed in this area can be useful for studying first language acquisition. In fact, the initial theoretical foundations of GI were given by E.M.Gold (1967), who tried to formalize the process of natural language acquisition.

According to Pearl and Goldwater (2016), language acquisition is a problem of induction: the child learner is faced with a set of specific linguistic examples and must infer some abstract linguistic knowledge that allows the child to generalize beyond the observed data, i.e., to both understand and generate new examples. Likewise, GI is a task where the goal is to learn or infer a grammar (or some device that can generate, recognize or describe strings) for a language and from all sorts of information about this language. GI consists, therefore, of finding the grammar or automaton for a language of which we are given an indirect presentation through strings, sequences, trees, terms or graphs (de la Higuera, 2010).

GI can, therefore, provide computational models for natural language acquisition. The use of formal or computational tools to give a description of the machinery necessary to acquire a language has been recognized as an important strategy within the field of language acquisition (Frank, 2011). In general, it is recognized that computational models can shed new light on language acquisition processes (Wintner, 2010). Even though, using computational tools for studying language is as old as the onset of Artificial Intelligence, over the last twenty-five years the progress in machine learning techniques has resulted in the emergence of a wider range of computational models that are much more powerful and robust than their predecessors (Alishahi, 2011).

Using computational tools, and therefore GI algorithms, for studying natural language acquisition offers many methodological advantages. Following Alishahi (2011) and Pearl (2010), we can highlight the following ones:

- *Explicit assumptions*. When implementing a computational model, every assumption of the input data and the learning mechanism has to be specified.
- *Controlled input*. Computational models offers the possibility to manipulate the language acquisition process and see the results of that manipulation. The researcher has full control over all the input data.
- *Observable behavior*. The impact of every factor in the input or the learning process can be directly studied in the output of the model. The performance of two different mechanisms on the same data set can be compared against each other.
- *Testable predictions*. Novel situations or combinations of data can be simulated and their effect on the model can be investigated.

Besides the enumerated advantages, one of the main benefits of computational models of language acquisition for determining relative linguistic complexity is the type of questions these formalisms could answer. According to Pearl (2010), language acquisition research is concerned with three different questions: *what* children know, *when* they know it, and *how* they learn it. While theoretical research deals with the knowledge that children acquire and experimental work provides information regarding the age at which the child acquires particular linguistic knowledge, computational modeling can explain how the child learns a language. Computational models, therefore, can be used to explain the *process* of natural language acquisition, because models are meant to be simulations of the child's acquisition mechanism.

Being tools for explaining the *process* of natural language acquisition, computational models in general, and GI algorithms in particular, are potential good tools to deal with developmental linguistic complexity.

4 Grammatical Inference Algorithms and Linguistic Complexity

Even though it would be possible to use different techniques from the field of machine learning to study linguistic complexity, we consider that GI models offer some advantages over other tools.

The first advantage is their motivation. As we have said, Gold (1967) introduced his model of identification in the limit with the ultimate goal of explaining the learning process of natural language:

> The study of language identification described here derives its motivation from artificial intelligence. The results and the methods used also have implications in computational linguistics, in particular the construction of discovery procedures, and in psycholinguistics, in particular the study of child learning (...). I wish to construct a precise model for the intuitive notion "able to speak a language" in order to be able to investigate theoretically how it can be achieved artificially (Gold, 1967).

Secondly, GI models have advantages over grammar induction tools, since whereas in grammar induction what really matters is the data and the relationship between the data and the induced grammar, in GI the actual learning process is what is central and is being examined and measured, not just the result of the process (de la Higuera, 2010).

Thirdly, an important advantage of GI tools is that they allow us to reproduce the learning context of first language acquisition. In fact, in any GI problem we have a teacher that provides data to a learner, and a learner (or learning algorithm) that from that data must identify the underlying language. This process has some similarities with the process of language acquisition where instead of a teacher and a learner, we have an adult and a child. In general, all the models in GI simulate learners who are developing monolingual L1 learning from monolingual data. However, modeling can be extended to other scenarios when the appropriate input data are available.

Importantly, GI models are grounded theoretically and empirically, as required in any computational model for language acquisition (Pearl, 2010). Theoretical grounding includes a description of the knowledge learners have and how it is represented. In GI models, the learner –this is, the machine– has no previously knowledge about the language. It has just the capacity –algorithm– to learn, but no linguistic structure previously storaged in order to facilitate the process. The machine represents, therefore, the child that has to acquire a language by just being exposed to this language. Empirical grounding includes using realistic data as input, measuring the model's learning behavior against children's learning behavior, and incorporating psychologically plausible algorithms into the model. In GI, the learner is exposed to language. The learner -like the child- can received positive and negative data as well as corrections. The model counts the needed number of interactions for the machine to achieve a good level of performance in a specific domain of the language. This could be seen as equivalent to calculate the child's cost/difficulty to acquire a language. There are different ways to present results, depending on what the model is testing. Useful measures are *recall*, *precision* and *F-score*.

By taking into account the above advantages, we claim that GI can provide a good tool for measuring linguistic complexity. We claim that models in this research area are potentially suitable for measuring the developmental complexity of languages, this is, the complexity understood in terms of cost and/or difficulty in language acquisition.

In what follows, we briefly outline the functioning of two novel models in the field of GI, in order to show their potential usefulness in the study of linguistic complexity.

4.1 Angluin and Becerra-Bonache's Model

Angluin and Becerra-Bonache (2010; 2011) introduced a novel model in GI inspired by studies on children's language acquisition. While the main part of studies in GI reduce the learning problem to the acquisition of the syntax, and omit any semantic information during the learning process, Angluin and Becerra-Bonache (2010; 2011) proposed a GI model that takes into account semantics during the language learning process.

In this model, the teacher and the learner interact in a sequence of situations by producing sentences that denote an object in each situation. These interactions are developed in the following way: First, a

situation is randomly generated and it is presented to the teacher and the learner; then, the learner tries to produce a sentence that designs one of the objects in this situation; after that, the teacher produces a random sentence that designs one of the objects in this situation and, finally, the learner analyzes the teacher's sentence and updates its current grammar for the language as appropriate.

Given any situation, the learner's goal is to produce correct sentences that denote one object in this situation. Semantics is formalized by using first order logic. In the model, a *situation* is composed of some objects and some of their properties and relations and it is represented as a finite set of ground atoms over some constants and predicates symbols. A *meaning* is a finite sequence of variable atoms, this is an expression formed by applying a predicate symbol to the correct number of variables as arguments. It is assumed that each utterance in the target language is assigned a unique meaning. An *utterance* is a finite sequence of words over a finite alphabet W of words. An utterance is *denoting* in a situation, if it uniquely picks out the objects it refers to in a situation. The *linguistic competence* of the teacher is represented by a finite state transducer. This transducer is used by the teacher for comprehension and production of utterances. Initially, the teacher has the meaning transducer for the target language, but the learner has no language-specific knowledge (i.e., the learner has not access to this transducer and it does not have any information about the target language). Regarding the *learning task*, the goal of the learner is to learn a grammar for the language that will enable to produce all and only the denoting utterances for any given situation. Although the learner's representation is referred as a grammar, it does not take the form of a classical grammar from formal languages. The learner's grammar has three main components: 1) Weighted co-occurrence graph; 2) general forms; 3) decision trees. Detailed information about the model can be found in Angluin and Becerra-Bonache (2010).

To evaluate the model it was used a simplification of the Feldman's *Miniature Language Acquisition task* (Stolcke et al., 1994) that consists on learning a sublanguage from sentences-picture pairs that involve geometric figures. The model was tested with limited sublanguages of ten different natural languages: English, German, Greek, Hebrew, Hungarian, Mandarin, Russia, Spanish, Swedish, Turkish.

In the experiments, each situation had two objects and one binary relation between them (above/below, to the left to/to the right to). Every object had three attributes: *form, color* and *size*. The attribute of shape had six possible values (circle, square, triangle, star, ellipse, hexagon), that of color had six possible values (red, orange, yellow, green, blue, purple), and that of size three possible values (big, medium, small). Thus, there were 108 different objects and 23,328 different situations. The total number of possible meanings was 113,064.

Two different measures were used to evaluate the performance of the learner: *correctness* and *completeness*. The *correctness* is the sum of the probabilities of the learner's sentences that are in the set of sentences that denote correctly one object. The *completeness* is the fraction of sentences that denote correctly one object and appear in the set of learner sentences. A learner achieved a level p of performance if correctness and completeness were at least p (this is a more stringent measure of performance than the F-score). In the experiments, teacher and learner interacted until the learner achieves a level $p = 0.99$. Table 1 shows the number of interactions that were necessary to achieve this level of performance in all the languages taken into account. Each entry is the median of 10 trials. The learner was evaluated after receiving 100 sentences from the teacher.

The results distinguished two different groups: 1) Greek and Russian that need at least 3.400 interactions with the teacher; and 2) the rest of languages that need at most 1.000 interactions.

In essence, the model developed by Angluin and Becerra-Bonache (2010; 2011) calculates the number of interactions that are necessary to achieve a good level of performance in a given language by using a unique algorithm to learn any of the languages analyzed. The model shows that not all the languages need the same number of linguistic interactions to reach the same level of performance. Even though we can think that these differences may be due to computational reasons (i.e., the size of the target machines), it has been shown in Angluin and Becerra-Bonache (2016) that these differences are because of linguistic reasons. For example, Mandarin has an alphabet (i.e., number of words) of half of the size of that for Greek, while its transducer has similar size to the Greek one, but Mandarin required fewer interactions to reach a high level of performance. Hence, this example shows that the differences in the results obtained

Level	0.60	0.70	0.80	0.90	0.95	0.99
English	200	200	300	400	500	700
German	200	300	300	400	550	800
Greek	400	500	700	1500	2200	3400
Hebrew	200	300	400	500	650	900
Hungarian	200	300	350	450	550	750
Mandarin	200	200	300	400	500	700
Russian	450	500	850	1750	2350	3700
Spanish	200	300	350	500	600	1000
Swedish	200	300	300	400	600	1000
Turkish	200	200	300	400	550	800

Table 1: Number of interactions that the learner needs to reach different levels of performance. Extracted from Angluin and Becerra-Bonache (2010) .

are due to linguistic rather than computational reasons, providing some evidence of the different level of linguistic complexity of the analyzed natural languages.

Therefore, this GI model may be a potential adequate tool to measure the linguistic complexity in *relative* terms. In fact, the unique algorithm used in the model could be equivalent to the innate capacity that allows humans to acquire a language. Moreover, the learner –this is, the machine– has no previously knowledge about the language. The machine represents, therefore, the child that has to acquire a language by just being exposed to this language. To count the needed number of interactions for the machine to achieve a good level of performance in a specific domain of the language may be equivalent to calculate the child's cost/difficulty to acquire a language. Finally, to show that with the same algorithm not every language requires the same number of interactions may be interpreted (in terms of complexity) as an evidence to defend that the difficulty/cost to acquire different languages is not the same and, therefore, languages differ in relative complexity.

4.2 Models based on Inductive Logic Programming Techniques

Another interesting model in GI that may be adequate to calculate linguistic complexity is the one introduced in Becerra-Bonache et. al (2015) and improved in Becerra-Bonache et al. (2016).

Inspired by Angluin and Becerra-Bonache (2010; 2011), Becerra-Bonache et al. (2015) developed a new system based on Inductive Logic Programming techniques. This system learns from pairs consisting of a *sentence* and the *context* in which this sentence has been produced. A *sentence* is represented as a sequence of words (n-grams) and for the context they use a first-order logic based representation. In contrast to other approaches, a *context* is a description of what the learner can see in the world, and not a set of candidate meanings for that utterance; the system constructs all candidate meanings itself. The model assumes that sentences are relevant, i.e., they never refer to something outside the context. The meaning of an n-gram is defined as whatever is in common among all contexts where the n-gram can be used. The algorithm incrementally learns the meaning of specific n-grams by using Inductive Logic Programming techniques. Becerra-Bonache et al. (2015) experimentally demonstrate that the proposed model can explain the gradual learning of simple concepts and language structure. The system was also tested with a toy dataset based on the Feldman's task (it contains simple noun phrases that referred only to the color, shape, size and relative position of simple objects). Experiments with three different languages (English, Dutch and Spanish) showed that the system learns a language model that can easily be used to understand, generate and translate utterances.

An improvement of the model introduced in Becerra-Bonache et al. (2015) is presented in Becerra-Bonache et al. (2016). In this paper, a system that deals with more realistic contexts (provided in the form of images) and work in noisy environments is introduced. The system learns from pairs (S,I) where S is a sentence telling something about a part of an image I. After a basic preprocessing step, each image I is transformed into a scene Sc, by using a first-order logic based representation. Therefore, the input of the learner is a dataset made up of pairs (S,Sc) where S is a sentence related to a particular scene Sc. The input pairs have similar properties to those of the inputs received by children (Fazly et al., 2010): a) *alignment ambiguity*: it is not explicitly indicated in the input which words refer to which

meaning; b) *referential uncertainty*: the description of a context may contain elements that are not in the corresponding sentence; c) *noise*: a sentence may refer to things that are not present in the context. The main improvements with respect to the work presented in Becerra-Bonache et al. (2015) are the following: 1) the system can better learn the meaning of words; 2) it can learn from noisy environments; 3) it can generate relevant sentences for a given scene, rather than just any sentence. Moreover, a series of experiments based on a more realistic dataset, called *Abstract Scenes Dataset* (Zitnick and Parikh, 2013), were conducted; this dataset contains clip-art pictures of children playing outdoors and sentences that describe these images. The goal of those experiments were to study the ability of the model to generate relevant sentences for a given scene and to learn the meaning of words.

We claim that those models may be used in the research on linguistic complexity. They are models that focused on the learning process. Moreover, they do not require any prior language-specific knowledge and learn incrementally. Therefore, they present features to, somehow, ensure that the model is actually about acquisition, rather than simply about what behavior a computational algorithm is capable of producing. They use realistic data and psychologically plausible algorithms that include features like gradual learning, robustness to noise in the data, and learning incrementally.

5 Conclusions

In this paper, we have proposed to use GI algorithms to measure the relative complexity of natural languages by considering the process of children first language acquisition. GI algorithms allow us to calculate the *cost* –in terms of the number of interactions– to reach a good level of performance in a given language. Therefore, GI offers the possibility to measure the *difficulty* of acquiring different natural languages.

The adequacy of GI models for calculating linguistic complexity is based on the fact that algorithms proposed in this area are computational models of language acquisition that use real data in the learning process. As any computational simulation, GI allows the researcher to perform some manipulations that could be not possible to carry out with children. Therefore, GI may provide data to calculate linguistic complexity that would be difficult to be obtained by observing the process of language acquisition through experimental research. Unlike psycholinguistic experiments with real children, GI models can avoid the problem of the influence of external (and non linguistic) factors that can conditioning the process of language acquisition. GI algorithms allow to reproduce the same context and features for the learning of any language. By analyzing the language acquisition process in different children in order to get data on linguistic complexity differences, we could not assure that every child have the same capacity, motivation, inputs, etc. If, on the contrary, we analyze the same speaker/child acquiring different languages, we could not assure that the acquisition of one language is not conditioned by the acquisition of the other language (even in bilingual acquisition processes). All these problems can be avoided by machine learning, since the computational simulation allows to reproduce exactly the same state/environment/requirements for the acquisition of any language.

As pointed out by Pearl (2010), the main disadvantage of modeling is that we can never be absolutely sure our model is really showing how acquisition works in children's minds. In fact, the reader could object that GI do not reproduce the process of natural language acquisition. We claim that, in order to defend the usefulness of GI models in the study of linguistic complexity, it not necessary to defend an identity between the processes of inferring a grammar and acquiring a natural language. We just say that an analogy can be established between those two processes. Moreover, in the case that not even the analogy is tenable, we still claiming that machine learning techniques can offer to linguistic complexity studies efficient computational resources that can measure complexity differences among languages adequately through objective and controlled means.

In the literature on linguistic complexity, it is common to read that there is no reason to believe that all languages are equally complex. However, no definite method has been proposed up to now to measure the relative complexity of languages. If measures and techniques from different disciplines have been used to calculate absolute complexity, why not to resort to machine learning to find useful tools for calculating developmental complexity.

We are working on the development of objective and meaningful methods, based on GI, to calculate linguistic complexity. GI models can be seen as an alternative to the methods that have been used so far. They present the following advantages: their *interdisciplinarity*, they combine ideas from linguistics with computational models; their *motivation*, they are based on how humans acquire language; their *results*, they provide quantifiable experimental results; and their ability to perform *cross-linguistic analysis*.

Acknowledgements

This research has been supported by the Ministerio de Economía y Competitividad under the project number FFI2015-69978-P (MINECO/FEDER) of the Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia, Subprograma Estatal de Generación de Conocimiento.

References

S. Abney and M. Johnson. 1991. Memory requirements and local ambiguities of parsing strategies. *Journal of Psycholinguistic Research*, 20(3):233–250.

A. Alishahi. 2011. *Computational Modeling of Human Language Acquisition*. Morgan and Claypool Publishers, Toronto.

A. Andrason. 2014. Language complexity: An insight from complex-system theory. *International Journal of Language and Linguistics*, 2(2):74–89.

D. Angluin and L. Becerra-Bonache. 2010. A model of semantics and corrections in language learning. Technical report, Yale University.

D. Angluin and L. Becerra-Bonache. 2011. Effects of meaning-preserving corrections on language learning. In *Proceedings of the 15th International Conference on Computational Natural Language Learning, CoNLL 2011*, pages 97–105. Portland.

D. Angluin and L. Becerra-Bonache. 2016. A model of language learning with semantics and meaning preserving corrections. *Artificial Intelligence*, 242:23–51.

M. Bane. 2008. Quantifying and measuring morphological complexity. In Ch. Chang and H. Haynie, editors, *Proceedings of the 26th West Coast Conference on Formal Linguistics*, pages 69–76. Cascadilla Proceedings Project, Somerville.

L. Becerra-Bonache, H. Blockeel, M. Galván, and F. Jacquenet. 2015. A first-order-logic based model for grounded language learning. In *Advances in Intelligent Data Analysis XIV - 14th International Symposium, IDA 2015, Saint Etienne, France, October 22-24, 2015, Proceedings*, pages 49–60.

L. Becerra-Bonache, H. Blockeel, M. Galván, and F. Jacquenet. 2016. Relational grounded language learning. In *ECAI 2016 - 22nd European Conference on Artificial Intelligence, 29 August-2 September 2016, The Hague, The Netherlands*, pages 1764–1765.

P. Blache. 2011. A computational model for linguistic complexity. In G. Bel-Enguix, V. Dahl, and M.D. Jiménez-López, editors, *Biology, Computation and Linguistics. New Interdisciplinary Paradigms*, pages 155–167. IOS Press, Amsterdam.

D. Crystal. 1997. *The Cambridge Encyclopedia of Language*. Cambridge University Press, Cambridge.

O. Dahl. 2004. *The Growth and Maintenance of Linguistic Complexity*. John Benjamins, Amsterdam.

C. de la Higuera. 2010. *Grammatical Inference: Learning Automata and Grammars*. Cambridge University Press, Cambridge.

A. D'Ulizia, F. Ferri, and P. Grifoni. 2011. A survey of grammatical inference methods for natural language learning. *Artificial Intelligence Review*, 36(1):1–27.

B. Edmonds. 1999. *Syntactic Measures of Complexity*. Ph.D. thesis, University of Manchester.

A. Fazly, A. Alishahi, and S. Stevenson. 2010. A probabilistic computational model of cross-situational word learning. *Cognitive Science*, 34(6):1017–1064.

F. Ferreira. 1991. Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language*, 30(2):210–233.

M. C. Frank. 2011. Computational models of early language acquisition. unpublished manuscript.

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

E. Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, and W. O'Neil, editors, *Image, language, brain: Papers from the first mind articulation project symposium*, pages 95–126. MIT Press, New York.

E.M. Gold. 1967. Language identification in the limit. *Information and Control*, 10:447–474.

J. Hawkins. 2009. An efficiency theory of complexity and related phenomena. In G. Sampson, D. Gil, and P. Trudgill, editors, *Language Complexity as an Evolving Variable*, pages 252–268. Oxford University Press, Oxford.

P. Juola. 2009. Assessing linguistic complexity. In M. Miestamo, K. Sinnemäki, and F. Karlsson, editors, *Language Complexity: Typology, Contact, Change*, pages 89–108. John Benjamins, Amsterdam.

W. Kusters. 2003. *Linguistic Complexity: The Influence of Social Change on Verbal Inflection*. LOT, Utrecht.

J. McWhorter. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology*, 6:125–166.

J. McWhorter. 2012. *Linguistic simplicity and complexity: Why do languages undress?* Mouton de Gruyter, Berlin.

M. Miestamo, K. Sinnemäki, and F. Karlsson. 2008. *Language Complexity: Typology, Contact, Change*. John Benjamins, Amsterdam.

M. Miestamo. 2006. On the feasibility of complexity metrics. In K. Krista and M.M Sepper, editors, *Finest Linguistics. Proceedings of the Annual Finish and Estonian Conference of Linguistics*, pages 11–26. Tallinna Ülikooli Kirjastus, Tallinn.

M. Miestamo. 2009a. Grammatical complexity in a cross-linguistic perspective. In M. Miestamo, K. Sinnemäki, and F. Karlsson, editors, *Language Complexity: Typology, Contact, Change*, pages 23–42. John Benjamins, Amsterdam.

M. Miestamo. 2009b. Implicational hierarchies and grammatical complexity. In G. Sampson, D. Gil, and P. Trudgill, editors, *Language Complexity as an Evolving Variable*, pages 80–97. Oxford University Press, Oxford.

M. Mitchell. 2009. *Complexity: A Guided Tour*. Oxford University Press, New York.

F.J. Newmeyer and L.B. Preston. 2014. *Measuring Grammatical Complexity*. Oxford University Press, Oxford.

E. Olivas, J.D.M. Guerrero, M.M. Sober, J.R.M. Benedito, and A.J.S. Lopez. 2009. *Handbook Of Research On Machine Learning Applications and Trends: Algorithms, Methods and Techniques - 2 Volumes*. Information Science Reference. IGI Publishing, Hershey, PA.

G. Pallotti. 2015. A simple view of linguistic complexity. *Second Language Research*, 31:117–134.

L. Pearl and S. Goldwater. 2016. Statistical learning, inductive bias, and Bayesian inference in language acquisition. In J. Lidz, W. Snyder, and J. Pater, editors, *Oxford Handbook of Developmental Linguistics*. Oxford University Press, Oxford.

L. Pearl. 2010. Using computational modeling in language acquisition research. In E. Blom and S. Unsworth, editors, *Experimental Methods in Language Acquisition Research*, pages 163–184. John Benjamins, Amsterdam.

G. Sampson, D. Gil, and P. Trudgill. 2009. *Language Complexity as an Evolving Variable*. Oxford University Press, Oxford.

W. Schuler. 2009. Positive results for parsing with a bounded stack using a model-based right-corner transform. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pages 344–352, Stroudsburg, PA, USA. Association for Computational Linguistics.

K. Sinnemäki. 2011. *Language Universals and Linguistic Complexity. Three Case Studies in Core Argument Making*. Ph.D. thesis, University of Helsinki.

A. Stolcke, J.A. Feldman, G. Lakoff, and S. Weber. 1994. Miniature language acquisition: A touchstone for cognitive science. *Cognitive Science*, 8:686–693.

P. Trudgill. 1999. Language contact and the function of linguistic gender. *Poznan Studies in Contemporary Linguistics*, 35:133–152.

P. Trudgill. 2001a. Contact and simplification: historical baggage and directionality in linguistic change. *Linguistic Typology*, 5(2/3):371–374.

P. Trudgill. 2001b. Linguistic and social typology: The Austronesian migrations and phoneme inventories. *Linguistic Typology*, 8:305–320.

S. Wintner. 2010. Computational models of language acquisition. In A. Gelbukh, editor, *CICLing 2010*, volume LNCS 6008, pages 86–99. Springer, Berlin.

C.L. Zitnick and D. Parikh. 2013. ringing semantics into focus using visual abstraction. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 3009–3016. Portland.