Universals on Natural Language Determiners from a PAC-learnability Perspective

Giorgio Magri (magrigrg@gmail.com)
SFL UMR 7023 (CNRS, University of Paris 8) UiL-OTS (Utrecht University)
61 rue Pouchet, 75017 Paris, France Trans 10, 3512 JK Utrecht, The Netherlands

Abstract

A classical conjecture in generative linguistics is that universal restrictions on determiners in Natural Language (e.g. *monotonicity*, *invariance*, and *conservativity*) serve the purpose of simplifying the language acquisition task. This paper formalizes this conjecture within the PAC-learnability framework.

Keywords: Natural Language determiners; PAC-learnability.

Introduction

Determiners and quantifiers. According to Natural Language Semantics (Montague, 1973; Heim & Kratzer, 1978), a noun like *animal* and a predicate like *tall* denote *properties*, namely subsets of the *domain of quantification* $\mathbb{D}$. *Determiners* like *some*, *every*, *no* denote a function $\wp(\mathbb{D}) \to \wp(\wp(\mathbb{D}))$ that takes a property such as the one denoted by *animal* and returns a collection of properties, exemplified in (1).

$$\begin{aligned} some(animal) &= \{B \,|\, animal \cap B \neq \emptyset\} \\ every(animal) &= \{B \,|\, animal \subseteq B\} \\ no(animal) &= \{B \,|\, animal \cap B = \emptyset\} \end{aligned} \tag{1}$$

The meaning of the sentence *Some/every/no animal is tall* with the parse [[$_{\text{NP}}$ *Some/every/no animal*] *tall*] is derived *compositionally* as follows. The denotation of the NP is obtained by applying the property *animal* to the function denoted by the determiner. This yields a collection of properties as in (1), called a *generalized quantifier*. The sentence is then true iff the property *tall* belongs to that generalized quantifier.

Universal restrictions. Westerståhl (1989) formulates "the basic question facing anyone who studies Natural Language quantification" as follows: "Logically, the category of determiners is extremely rich. For example, even on a [domain $\mathbb{D}$] with two elements, there are $2^{16} = 65.536$ possible determiners. But, in natural languages, just a small portion of these are [lexicalized]. Which ones [...]? What are the *constraints* on determiner denotations in natural languages?". It has been suggested (Westerståhl, 1989; Gamut, 1991; Keenan & Stavi, 1986; Barwise & Cooper, 1981; Benthem, 1983) that all lexical (i.e., not syntactically complex) determiners in Natural Language satisfy the *monotonicity*, *invariance* and *conservativity* constraints defined below.

DEF 1 *A determiner* $\Delta : \wp(\mathbb{D}) \to \wp(\wp(\mathbb{D}))$ *is:*

- **(a)** monotone$^+$ *(*monotone$^-$*) iff* $B \in \Delta(A)$ *entails* $B' \in \Delta(A)$*, for any properties* $A, B, B' \subseteq \mathbb{D}$ *such that* $B \subseteq B'$ *(*$B \supseteq B'$*). It is* monotone *iff it is either monotone*$^+$ *or monotone*$^-$*;*
- **(b)** conservative *provided* $B \in \Delta(A)$ *holds iff* $A \cap B \in \Delta(A)$*, for any properties* $A, B \subseteq \mathbb{D}$*;*
- **(c)** invariant *provided* $B \in \Delta(A)$ *holds iff* $\pi(B) \in \Delta\big(\pi(A)\big)$*, for any permutation* π *over the domain of quantification* $\mathbb{D}$ *and any properties* $A, B \subseteq \mathbb{D}$*.*

Here are some examples. The property *dog* is a subset of the property *animal*. The determiner *some* is monotone$^+$ because *Some dog is tall* entails *Some animal is tall*. The determiner *no* is monotone$^-$ because *No animal is tall* entails *No dog is tall*. The determiners *some*, *every* and *no* are conservative because the sentence *Some/every/no dog is tall* is true iff *Some/every/no dog is a tall dog* is true. The latter equivalence does not hold for *only*: *Only dogs are tall* is stronger than *Only dogs are tall dogs*. And indeed *only* is not a determiner (rather an adverb), showing that the conservativity universal for lexical determiners has empirical bite.

The challenge of a learnability approach. Gamut (1991) points out that, although these universals are "a significant contribution to the characterization of the notion of *possible human language*" they do "not give any clue as to why this should be so. [...]. A formulation of a universal [...] is one thing; the explanation for it is something else." This is the problem addressed in this paper: how can these universal restrictions on the denotations of lexical determiners be *explained*? Generative linguistics has focused mainly on the conservativity universal, and has suggested that conservativity (and perhaps also the other universals) should be explained from a learnability perspective. For instance, Hunter and Lidz (2013) show that four- and five-year-olds fail to learn a novel nonconservative determiner but succeed in learning a comparable conservative determiner, consistently with the learnability hypothesis.[1] The question, then, is what might make conservativity and the other universals well suited from a learnability perspective.

Keenan and Stavi (1986) take on this issue. To start, they note that there are 2^{4^n} possible determiners over a domain of quantification $\mathbb{D}_n$ of cardinality n but that only 2^{3^n} are conservative. For instance, "in a model with only two individuals there are [...] $16^4 = 65.536$ determiners [...but] only 512 of these [...] are conservative! The constraint then is extremely strong [...]: the language learner does not have to seek the meaning of a novel determiner among all the logically possible [ones]. He only has to choose from among those ways which satisfy conservativity." This argument is rather weak. If the effectiveness of conservativity were to be measured in terms of its ability to prune the learner's search space, then the constraint is not at all "extremely strong" but rather quite weak, as it does not alter the asymptotic exponential growth of the search space as a function of the cardinality n of the domain of quantification. Furthermore, this argument does not explain why Natural Language enforces precisely conser-

[1] But see Fox (1999) for an approach to conservativity not based on learnability; and Piantadosi (2012) for modeling evidence that conservativity does not improve learnabaility.

vativity (together with the other universals) among the many alternative possible ways of restricting the search space.

Keenan and Stavi also show that the family of conservative determiners coincides with the closure of the set $\{some, every\}$ w.r.t. to the operations of conjunction, disjunction and adjectival restrictions. They then observe that this result "do[es] provide a basis for saying that the set of conservative determiners is cognitively apprehendable. Namely, we do in fact understand the denotations of simple determiners (*every* and *some*) [...]. And we do have a cognitive grasp of boolean operations and adjectival restriction, as [...] we use them in understanding meanings associated with essentially all categories." Again, the argument is rather weak. The determiner denoted by *only* is analogous to that denoted by *every*, just with the reversed set-inclusion (takes a property A and returns properties B such that $B \subseteq A$ rather than $A \subseteq B$). Thus, *only* is not conservative as noted above, and yet not in any way more complex than *every*. Why has Natural Language restricted determiners to the conservative ones rather than, say, to the closure of the set $\{some, only\}$?

Meeting the challenge with PAC-learnability. These objections are not meant to challenge Keenan and Stavi's learnability approach to universal restrictions on lexical determiners. Rather, they are meant to argue that this learnability approach needs to be cast within an explicit, formal learning framework, contrary to what done by Keenan and Stavi and much of the subsequent literature. This paper thus develops this learnability approach to universal restrictions on determiners within the *PAC learnability* paradigm. As argued by Natarajan (1991), "the PAC paradigm appears to be a good model of the natural learning process while lending itself to analysis" (but seeClark & Lappin, 2011 for discussion). It shows that the entire class of determiners is not learnable, not even according to a very weak PAC-learnability notion. This result provides support for the hypothesis that Universal Grammar enforces restrictions on lexical determiners to boost acquisition. It then shows that *monotonicity* has almost no effects on learnability. Thus, there are restrictions that do not affect learnability, reinforcing my point against Keenan and Stavi's naïve approach. Finally, it looks at *conservativity* and *invariance*, showing that those restrictions have a strong learnability effect, allowing PAC-learnability from malicious positive examples only.

No PAC-learnability without restrictions

Generalized quantifiers are not PAC-learnable. A *sample space* is a set $X = \bigcup_{n=1}^{\infty} X_n$ where $X_1, X_2 \ldots$ are finite (to avoid measurability problems) and disjoint sets, whose elements are called *examples*. A *concept class* on X is a set $\mathcal{C} = \bigcup_{n=1}^{\infty} \mathcal{C}_n$ where each $\mathcal{C}_n \subseteq \wp(X_n)$ is a collection of subsets of X_n, called *concepts*. A *sample* is an m-tuple $\mathbf{x} = (x_1, \ldots, x_m)$ in $X_n^m = X_n \times \ldots \times X_n$. A concept $c \in \mathcal{C}_n$ can be identified with its characteristic function $c : X_n \to \{0,1\}$. The *labels* assigned by a concept c to a sample $\mathbf{x}$ can be collected into the boolean vector $c(\mathbf{x}) = (c(x_1), \ldots, c(x_m)) \in \{0,1\}^m$. For any two concepts $c, h \in \mathcal{C}_n$ and any probability measure P over X_n, the *error* $e_{c,P}(h)$ of h w.r.t. c relative to P is the probability $P(c \bigtriangleup h)$of the symmetric difference between c and h. A concept class is *learnable* provided any concept in the class can be identified from a labeled sample with high *accuracy* (error smaller than ε) and with high *confidence* (larger than $1-\delta$), as formalized in Definition 2 (Valiant, 1984; Kearns & Vazirani, 1994; Kearns, 1999).

DEF 2 *A concept class $\mathcal{C}$ is* PAC-learnable *with* sample cardinality $m : (\varepsilon, \delta, n) \in (0,1) \times (0,1) \times \mathbb{N} \mapsto m(\varepsilon, \delta, n) \in \mathbb{N}$ *provided there exists a* learning function $\mathfrak{A}$ *of the form*

$$\mathfrak{A} : (\mathbf{x}, \mathbf{t}) \in X_n^m \times \{0,1\}^m \mapsto \mathfrak{A}(\mathbf{x}, \mathbf{t}) \in \mathcal{C}_n \tag{2}$$

such that for any $\varepsilon, \delta \in (0,1)$, any $n \in \mathbb{N}$, any concept $c \in \mathcal{C}_n$ and any probability measure P over X_n, the following condition holds with $m = m(\varepsilon, \delta, n)$

$$P^m \left\{ \mathbf{x} \in X_n^m \middle| e_{c,P}\Big(\mathfrak{A}\big(\mathbf{x}, c(\mathbf{x})\big)\Big) \geq \varepsilon \right\} \leq \delta \tag{3}$$

and furthermore $m(\cdot,\cdot,\cdot)$ grows polynomially in $\frac{1}{\varepsilon}$, $\frac{1}{\delta}$ and n.[2]

A subset $S \subseteq X_n$ is *shattered* by $\mathcal{C}_n$ provided $\{S \cap c \,|\, c \in \mathcal{C}_n\} = \wp(S)$. The concept class $\mathcal{C}_n$ has *Vapnik-Chervonenkis dimension* $VCD(\mathcal{C}_n)$ equal to $d \in \mathbb{N}$ provided there exists a shattered subset $S \subseteq X_n$ with cardinality d but no shattered subset $S \subseteq X_n$ with cardinality $d+1$. VCD controls the sample cardinality needed for PAC-learnability (Ehrenfeucht et al., 1989): no learning function $\mathfrak{A}$ satisfies condition (3) of PAC-learnability with sample cardinality $m = m(\varepsilon, \delta, n)$ smaller than $\frac{VCD(\mathcal{C}_n)-1}{32\varepsilon}$. With these preliminaries in place, let's now go back to generalized quantifiers, defined as follows.

DEF 3 *Consider the sample space $X = \bigcup_{n=1}^{\infty} X_n$ where $X_n = \wp(\mathbb{D}_n)$ is the collection of properties over a domain of quantification $\mathbb{D}_n$ of cardinality n. Consider the concept class $\mathcal{Q} = \bigcup_{n=1}^{\infty} \mathcal{Q}_n$ where $\mathcal{Q}_n = \wp(X_n)$ is the collection of generalized quantifiers $\mathbf{q} \in \wp(\wp(\mathbb{D}_n))$ over $\mathbb{D}_n$.*

By definition, $\mathcal{Q}_n$ shatters X_n. And X_n has cardinality 2^n. Hence, $VCD(\mathcal{Q}_n) = 2^n$. Any learning function for $\mathcal{Q}$ thus needs a sample cardinality $m(\cdot,\cdot,\cdot)$ larger than $\frac{2^n-1}{32\varepsilon}$ and therefore not polynomial in n. I thus conclude that:

RESULT 1 *The whole class $\mathcal{Q}$ of generalized quantifiers is not PAC-learnable.*

Not even uniformly weakly PAC-learnable. Let's weaken condition (3) to (4): the error is not required to be arbitrary small (i.e., smaller than ε) but just smaller than chance, with the polynomial $q(\cdot)$ controlling the improvement over chance.

DEF 4 *A concept class $\mathcal{C}$ is* weakly PAC-learnable *with sample cardinality $m : (\varepsilon, \delta, n) \in (0,1) \times (0,1) \times \mathbb{N} \mapsto \mathbb{N}$ provided there exists a* learning function *(2) such that for any $\delta \in (0,1)$,*

[2] The learning function $\mathfrak{A}$ is often also required to be computable in time polynomial in $\frac{1}{\varepsilon}$, $\frac{1}{\delta}$ and n (Valiant, 1984). In this paper, I ignore computational efficiency, and focus on a statistical perspective.

any $n \in \mathbb{N}$, any distribution P over X_n and any concept $c \in \mathcal{C}_n$, condition (4) holds with $m = m(\varepsilon, \delta, n)$

$$P^m\left\{\mathbf{x} \in X_n^m \middle| e_{c,P}\Big(\mathfrak{A}\big(\mathbf{x}, c(\mathbf{x})\big)\Big) \geq \frac{1}{2} - \frac{1}{q(n)}\right\} \leq \delta \quad (4)$$

and $m(\cdot,\cdot,\cdot)$ grows polynomially in $\frac{1}{\varepsilon}$, $\frac{1}{\delta}$ and n.

Weak and *strong* PAC-learnability are equivalent (Schapire, 1990), because of their distribution-independent nature (i.e., the learning function needs to succeed for any distribution P). Thus, weak PAC-learning is only interesting for fixed probability distributions. Consider the uniform distribution U. VCD also controls the sample cardinality for uniform weak PAC-learnability (Blumer et al., 1989): no learning function $\mathfrak{A}$ satisfies (4) with sample cardinality $m = m(\varepsilon, \delta, n)$ smaller than $\frac{VCD(\mathcal{C}_n)-1}{32\left(\frac{1}{2} - \frac{1}{q(n)}\right)}$. Reasoning as above, I thus obtain Result 2, that substantially strengthens Result 1.

RESULT 2 *The class $\mathcal{Q}$ of generalized quantifiers is not* weakly *PAC-learnable relative to the uniform distribution.*

Results 1 and 2 lend support to the initial conjecture that UG needs to enforce universal restrictions on the denotations of quantifiers in order to make the acquisition of quantifiers feasible. The rest of the paper thus investigates the learnability implications of these universals restrictions.

Monotonicity does not help with learnability

Monotonicity has a modest learnability effect. A universal restriction on quantifiers in Natural Language is that they be *monotone*, according to Definition 1a. I thus focus on the PAC-learnability of the class of monotone quantifiers.

DEF 5 *Consider the sample space $X = \bigcup_{n=1}^{\infty} X_n$ where $X_n = \wp(\mathbb{D}_n)$ is the collection of properties over a domain of quantification $\mathbb{D}_n$ of cardinality n. A generalized quantifier $\mathbf{q} \in \mathcal{Q}_n$ is* monotone$^+$ *provided that, if $B_1 \in \mathbf{q}$ and $B_1 \subseteq B_2$, then $B_2 \in \mathbf{q}$, for any two properties $B_1, B_2 \in X_n$; it is* monotone$^-$ *provided that, if $B_1 \in \mathbf{q}$ and $B_2 \subseteq B_1$, then $B_2 \in \mathbf{q}$; it is* monotone *provided it is either monotone$^+$ or monotone$^-$. $\mathcal{Q}^M = \bigcup_{n=1}^{\infty} \mathcal{Q}_n^M$ is the concept class of monotone quantifiers.*

Result 2 says that the entire class $\mathcal{Q}$ of generalized quantifiers is not weakly PAC-learnable w.r.t. the uniform distribution. A construction by Kearns et al. (1994) can be readapted to show that the subclass $\mathcal{Q}^M$ of monotone quantifiers is instead weakly PAC-learnable w.r.t. the uniform distribution. Monotonicity thus does lead to a learnability advantage. But the advantage is very modest, as it relies on the assumption of uniform distribution, which cannot be relaxed, by Result 3.

RESULT 3 *The class $\mathcal{Q}^M = \bigcup_{n=1}^{\infty} \mathcal{Q}_n^M$ of monotone quantifiers is not (weakly) PAC-learnable (w.r.t. arbitrary distributions).*

Proof. As recalled, it is sufficient to show that the Vapnik-Chervonenkis dimension $VCD(\mathcal{Q}_n^M)$ of monotone quantifiers grows super-polynomially in n. Assume n is even and let $S_n \subseteq X_n$ be the subset of properties of cardinality $\frac{n}{2}$. Thus:

$$VCD(\mathcal{Q}_n^M) \overset{(*)}{\geq} |S_n| \overset{(**)}{=} \binom{n}{\frac{n}{2}} = \sum_{k=0}^{n/2} \binom{\frac{n}{2}}{k}^2 \geq \sum_{k=0}^{n/2} \binom{\frac{n}{2}}{k} = \sqrt{2^n}$$

Where (*) holds because S_n is shattered by monotone quantifiers, as all properties in S_n have the same cardinality and thus cannot be in a subset relation. And (**) holds because $|S_n|$ is the number of subsets of cardinality $\frac{n}{2}$ out of n elements. □

A restriction that has no learnability effects. Given a concept class $\mathcal{C} = \bigcup_{n=1}^{\infty} \mathcal{C}_n$ over a sample space $X = \bigcup_{n=1}^{\infty} X_n$ and another concept class $\mathcal{C}' = \bigcup_{n=1}^{\infty} \mathcal{C}'_n$ over a possibly different sample space $X' = \bigcup_{n=1}^{\infty} X'_n$, $\mathcal{C}$ is *PAC-riducible* to $\mathcal{C}'$ according to Pitt and Warmuth (1990) iff there exists a polynomial $p(\cdot)$ and two *PAC-reduction maps*

$$R_1 : x \in X_n \mapsto R_1(x) \in X'_{p(n)} \qquad R_2 : c \in \mathcal{C}_n \mapsto R_2(c) \in \mathcal{C}'_{p(n)}$$

such that for any example $x \in X_n$ and any concept $c \in \mathcal{C}_n$ we have that $x \in c$ iff $R_1(x) \in R_2(c)$, i.e. the behavior of the transformed concept on the transformed examples is exactly the behavior of the original concept on the original examples. If $\mathcal{C}$ is PAC-reducible to $\mathcal{C}'$ and $\mathcal{C}'$ is PAC-learnable, then $\mathcal{C}$ is PAC-learnable too (Pitt & Warmuth, 1990). Let $\mathcal{Q}^{M,+} = \bigcup_{n=1}^{\infty} \mathcal{Q}_n^{M,+}$ be the concept class of monotone$^+$ quantifiers. Result 4 says that, despite the fact that $\mathcal{Q}^{M,+}$ is smaller than $\mathcal{Q}^M$, it has no learnability advantages. This result shows the importance of exploring the learnability implications of universal restrictions within an explicit learnability framework. The proof is based on a technique from Kearns et al. (1994) and Pitt and Warmuth (1990).

RESULT 4 *The subclass $\mathcal{Q}^{M,+}$ is PAC-reducible to the entire class $\mathcal{Q}^M$ of monotone quantifiers.*

Proof. For any n, order the elements in the domain of quantification $\mathbb{D}_n$ into a sequence, so that $\mathbb{D}_n = (d_1, \ldots, d_n)$. Let $p(n) = 2n$. Define the reduction map $R_1 : x \in X_n \mapsto R_1(x) \in X_{2n}$ as follows: for $i = 1, \ldots, n$, let $d_i \in R_1(x)$ iff $d_i \in x$; and for $i = n+1, \ldots, 2n$, let $d_i \in R_1(x)$ iff $d_{i-n} \notin x$. Then, $R_1(y) \subseteq R_1(x)$ entails $y = x$. Define next the reduction map $R_2 : \mathbf{q} \in \mathcal{Q}_n \mapsto R_2(\mathbf{q}) \in \mathcal{Q}_{2n}^{M,+}$ as follows: if $\mathbf{q} = \{x_1, \ldots, x_m\} \in \mathcal{Q}_n$, then $R_2(\mathbf{q}) \in \mathcal{Q}_{2n}^{M,+}$ is the monotone$^+$ quantifier that consists of the properties $R_1(x_1), \ldots, R_1(x_m)$ as well as of all their supersets. It is easy to check that $x \in \mathbf{q}$ iff $R_1(x) \in R_2(\mathbf{q})$. □

Conservativity and invariance help learnability

A determiner Δ is *conservative* (Definition 1b) provided that $B \in \Delta(A)$ iff $A \cap B \in \Delta(A)$, for any properties $A, B \subseteq \mathbb{D}$. Assume that Δ is furthermore *invariant* (Definition 1c). Thus, whether it is the case that $A \cap B \in \Delta(A)$ depends only on the cardinality of $A \cap B$. Thus, to learn from examples the denotations of quantified noun phrases projected by conservative and invariant determiners means to learn the concept class $\mathcal{Q}^{C,I}$ defined below, that is the focus of this section.

DEF 6 *Consider the sample space $X = \bigcup_{n=1}^{\infty} X_n$ where $X_n = \wp(\mathbb{D}_n)$ is the collection of properties over a domain of quantification $\mathbb{D}_n$ of cardinality n. Consider the concept class $\mathcal{Q}^{C,I} = \bigcup_{n=1}^{\infty} \mathcal{Q}_n^{C,I}$ where $\mathcal{Q}_n^{C,I}$ is the collection of those generalized quantifiers $\mathbf{q} \in \wp(X_n)$ that are* conservative *and* invariant*, namely satisfy the following implication: if $A \in \mathbf{q}$ and $|A| = |B|$, then $B \in \mathbf{q}$, for any properties $A, B \subseteq X_n$.*

Plain PAC-learnability. A learning function $\mathfrak{A}$ as in (2) is *consistent* provided that for any labeled sample $(\mathbf{x},\mathbf{t}) \in X_n^m \times \{0,1\}^m$, it returns a concept $\hat{c} = \mathfrak{A}(\mathbf{x},\mathbf{t})$ that classifies the examples $\mathbf{x} = (x_1,\ldots,x_m)$ according to the labels $\mathbf{t} = (t_1,\ldots,t_m)$, i.e. $\hat{c}(x_i) = t_i$. A consistent learning function satisfies the PAC-learnability condition (3) provided its sample cardinality m is large enough (Blumer et al., 1989):

$$m(\varepsilon,\delta,n) \geq \max\left\{\frac{4}{\varepsilon}\log\frac{2}{\delta}, \frac{8\mathrm{VCD}(\mathcal{C}_n)}{\varepsilon}\log\frac{13}{\varepsilon}\right\} \qquad (5)$$

The Vapnik-Chervonenkis dimension of the concept class $\mathcal{Q}^{C,I}$ of conservative and invariant quantifiers is $n+1$. The bound in (5) is thus compatible with a polynomial sample cardinality function m. I thus only need to construct a consistent learning function $\mathfrak{A}$ for $\mathcal{Q}^{C,I}$. Assume $\mathfrak{A}$ takes a labeled sample $(\mathbf{x},\mathbf{t}) \in X_n^m \times \{0,1\}^m$ and returns the generalized quantifier $\mathbf{q} = \mathfrak{A}(\mathbf{x},\mathbf{t})$ that contains the properties of a certain cardinality $i \in \{0,1,\ldots,n\}$ iff one of the properties in the sample $\mathbf{x}$ with a positive label in $\mathbf{t}$ has cardinality i. This learning function is obviously consistent. Hence:

RESULT 5 *The class $\mathcal{Q}^{C,I}$ of conservative and invariant quantifiers is PAC-learnable.*

Result 5 contrasts sharply with Results 1 and 3, lending support to a learnability explanation of the universals of invariance and conservativity. I now strengthen Result 5 by looking at more demanding PAC-learnability notions.

PAC-learnability from statistical information. According to PAC-learnability (Definition 2), a learning function has access to a labeled sample of a concept. I now consider a stronger notion of PAC-learnability, whereby the learning function has access only to *statistical information* about a concept. For instance, statistical information concerning a concept $c \subseteq X_n$ over the set X_n of patients of age n could be the probability w.r.t. a certain distribution P that a patient in c is overweight. This probability is $P\{x \in X_n \,|\, \Phi(x,c(x)) = 1\}$ where $\Phi(x,t) = 1$ iff $t = 1$ and x is overweight.

DEF 7 *A concept class $\mathcal{C}$ is* PAC-learnable *from the* exact statistical information *induced by* $\Phi_1,\Phi_2,\ldots : X \times \{0,1\} \to \{0,1\}$ *with sample cardinality* $m : (\varepsilon,n) \in (0,1)\times\mathbb{N} \to m(\varepsilon,n) \in \mathbb{N}$ *polynomial in* $\frac{1}{\varepsilon}, n$ *provided there exists a learning function*

$$\mathfrak{A} : \big(\varepsilon,n,\mathbf{p}\big) \in (0,1)\times\mathbb{N}\times[0,1]^m \mapsto \mathfrak{A}\big(\varepsilon,n,\mathbf{p}\big) \in \mathcal{C}_n \qquad (6)$$

such that for any $\varepsilon \in (0,1)$, any $n \in \mathbb{N}$, any concept $c \in \mathcal{C}_n$, any probability P over X_n, the following condition holds

$$e_{c,P}\Big(\mathfrak{A}\big(\varepsilon,n,(\hat{p}_1^c,\ldots,\hat{p}_m^c)\big)\Big) \leq \varepsilon \qquad (7)$$

where $\hat{p}_i^c$ is exact statistical information, namely:

$$\hat{p}_i^c = P\{x \in X_n | \Phi_i(x,c(x)) = 1\} \qquad (8)$$

Also, $\mathcal{C}$ is PAC-learnable from approximated statistical information *provided (7) holds with (8) replaced by*

$$|\hat{p}_i^c - P\{x \in X_n | \Phi_i(x,c(x)) = 1\}| \leq \tau$$

for some constant $\tau \in (0,1]$ with $\frac{1}{\tau} \leq m(\varepsilon,n)$.

The learning function (6) differs slightly from (2): the parameter δ has been suppressed because uninformative samples are already averaged out by the statistical information; the parameter ε is provided to the learning function; the parameter n is provided as well (that was not necessary in (2), because implicit in the sample $\mathbf{x} \in X_n^m$). We have that:

RESULT 6 *The class $\mathcal{Q}^{C,I}$ of conservative and invariant generalized quantifiers is PAC-learnable from approximated statistical information.*

Proof. Define the functions $\Phi_0,\Phi_1,\ldots : X \times \{0,1\} \to \{0,1\}$ by setting $\Phi_i(x,t) = 1$ iff $t = 1$ and $|x| = i$. Let $m(\varepsilon,n) \doteq n+1$ and $\tau = \tau(\varepsilon,n) \doteq \frac{\varepsilon}{3n}$. Define the learning function $\mathfrak{A}$ of the form (6) as follows: for any parameters ε and n and for any vector $\mathbf{p} \in [0,1]^m$ (whose $m = n+1$ components are indexed from 0 to n), let $\mathfrak{A}(\varepsilon,n,\mathbf{p})$ be the generalized quantifier in $\mathcal{Q}_u^{C,I}$ that contains properties of cardinality $i \in \{0,\ldots,n\}$ iff $p_i \geq \frac{2\varepsilon}{3n}$. For any $\mathbf{q} \in \mathcal{Q}_u^{C,I}$, I can thus bound as follows:

$$P\big\{x \in \mathbf{q} \;\big|\; x \notin \mathfrak{A}\big(\varepsilon,n,(\hat{p}_0^{\mathbf{q}}\ldots\hat{p}_m^{\mathbf{q}})\big)\big\} = \sum_{i=0}^{n} P\left\{x \in \mathbf{q} \;\middle|\; \begin{array}{l} |x| = i \\ \hat{p}_i^{\mathbf{q}} < \frac{2\varepsilon}{3n} \end{array}\right\}$$
$$= \sum_{i=0,\,\hat{p}_i^{\mathbf{q}} < \frac{2\varepsilon}{3n}}^{n} P\big\{x \in \mathbf{q} \;\big|\; |x| = i\big\} = \sum_{i=0,\,\hat{p}_i^{\mathbf{q}} < \frac{2\varepsilon}{3n}}^{n} P\big\{x \in X_n \;\big|\; \Phi_i(x,\mathbf{q}(x)) = 1\big\}$$
$$\leq \sum_{i=0,\,\hat{p}_i^{\mathbf{q}} < \frac{2\varepsilon}{3n}}^{n} \hat{p}_i^{\mathbf{q}} + \tau(\varepsilon,n) \leq \sum_{i=0,\,\hat{p}_i^{\mathbf{q}} < \frac{2\varepsilon}{3n}}^{n} \frac{2\varepsilon}{3n} + \frac{\varepsilon}{3n} \leq \varepsilon$$

If a property x of cardinality i does not belong to the quantifier $\mathbf{q}$, $\Phi_i\big(x,\mathbf{q}(x)\big) = 0$ for any x and $P\{x \in X_n | \Phi_i(x,\mathbf{q}(x)) = 1\} = 0$. Hence $\hat{p}_i^{\mathbf{q}} \leq \tau(\varepsilon,n) \leq \frac{\varepsilon}{3n}$, and the following quantity is zero.

$$P\big\{x \notin \mathbf{q} \;\big|\; x \in \mathfrak{A}\big(\varepsilon,n,(\hat{p}_0^{\mathbf{q}}\ldots\hat{p}_m^{\mathbf{q}})\big)\big\} = P\big\{x \notin \mathbf{q} \;\big|\; \hat{p}_i^{\mathbf{q}} \geq \tfrac{2\varepsilon}{3n}, i = |x|\big\}$$

As the error $e_{\mathbf{q},P}\big(\mathfrak{A}\big(\varepsilon,n,(\hat{p}_1^{\mathbf{q}},\ldots,\hat{p}_m^{\mathbf{q}})\big)\big)$ is the sum of the two terms just bounded, (7) holds. □

PAC-learnability from misclassified examples. According to PAC-learnability (Definition 2), the learning function is trained on a sample $\mathbf{x} = (x_1,\ldots,x_m) \in X_n^m$ together with the corresponding correct labels $c(\mathbf{x}) = (c(x_1),\ldots,c(x_m))$ assigned by a target concept c. I now consider a stronger notion of PAC-learnability, whereby some of the labels $c(x_i)$ are altered. Given $\xi \in \{0,1\}$ and $x \in X_n$, define $c(x,\xi) = c(x)$ iff $\xi = 1$. Assume that ξ is sampled according to a Bernoulli B_η with probability of success $\xi = 1$ equal to $\eta \in [0,1]$. The m-tuple $(c(x_1,\xi_1),\ldots,c(x_m,\xi_m))$ is denoted by $c(\mathbf{x},\xi)$.

DEF 8 *A concept class $\mathcal{C}$ is* PAC-learnable *from* misclassified examples *with* sample cardinality function $m : (\varepsilon,\delta,n,\eta) \in (0,1)\times(0,1)\times\mathbb{N}\times[0,\frac{1}{2}) \mapsto \mathbb{N}$ *if there is a* learning function

$$\mathfrak{A} : (\varepsilon,\eta,\mathbf{x},\mathbf{t}) \in (0,1)\times[0,\frac{1}{2})\times X_n^m \times \{0,1\}^m \mapsto \mathfrak{A}(\varepsilon,\eta,\mathbf{x},\mathbf{t}) \in \mathcal{C}_n$$

such that for any $\varepsilon,\delta \in (0,1)$, any $n \in \mathbb{N}$, any $\eta \in [0,\frac{1}{2})$, any concept $c \in \mathcal{C}_n$, and any probability P over X_n, we have

$$P^m \times B_\eta^m\left\{(\mathbf{x},\xi)\Big|e_{c,P}\Big(\mathfrak{A}\big(\varepsilon,\eta,\mathbf{x},c(\mathbf{x},\xi)\big)\Big) \geq \varepsilon\right\} \delta$$

and furthermore the sample cardinality $m(\cdot,\cdot,\cdot,\cdot)$ grows polynomially in $\frac{1}{\varepsilon}$, $\frac{1}{\delta}$, n and $1/\big(\frac{1}{2}-\eta\big)$.

The misclassification rate η cannot be larger than $\frac{1}{2}$, otherwise learning would be impossible. As the complexity of the learning task increases as η gets closer to the threshold $\frac{1}{2}$, the cardinality m of the sample is allowed to grow (polynomially) with $1/\left(\frac{1}{2}-\eta\right)$. The learning function $\mathfrak{A}$ is provided with the noise rate η and the accuracy parameter ε. PAC-learnability from statistical information is known to entail PAC-learnability from a misclassified sample (Kearns, 1998). From Result 6 we thus have:

RESULT 7 *The class $Q^{C,I}$ of conservative and invariant generalized quantifiers is PAC-learnable from misclassified examples.*

PAC-learnability from positive, malicious examples. According to PAC-learnability (Definition 2), when the learning function is trained on a target concept c, it is provided with a sample $\mathbf{x} = (x_1, \ldots, x_m) \in X_n^m$ that in general contains both positive examples $x_i \in c$ and negative examples $x_i \in \overline{c}$ (where $\overline{c}$ is the complement of c w.r.t. X_n). I now consider a stronger notion of PAC-learnability, whereby $\mathbf{x}$ is sampled w.r.t. a distribution concentrated on c so that the learning function receives only positive examples (Kearns et al., 1994).

DEF 9 *A concept class $\mathcal{C}$ is PAC-learnable* from positive examples only *with sample cardinality $m : (0,1) \times (0,1) \times \mathbb{N} \to \mathbb{N}$ provided there exists a* learning function $\mathfrak{A} : \bigcup_{n,m=1}^{\infty} X_n^m \to \bigcup_{n=1}^{\infty} \mathcal{C}_n$ *such that for any $\varepsilon, \delta \in (0,1)$, any $n \in \mathbb{N}$, any concept $c \in \mathcal{C}_n$ and any probability measures $P, \overline{P}$ concentrated over c and $\overline{c}$ respectively, condition (9) holds with $m = m(\varepsilon, \delta, n)$*

$$P^m \left\{ \mathbf{x} \in X_n^m \;\middle|\; \begin{array}{l} e_{c,P}(\mathfrak{A}(\mathbf{x})) \leq \varepsilon \\ e_{c,\overline{P}}(\mathfrak{A}(\mathbf{x})) \leq \varepsilon \end{array} \right\} \geq 1 - \delta \tag{9}$$

and furthermore the sample cardinality function $m(\cdot,\cdot,\cdot)$ grows polynomially in $\frac{1}{\varepsilon}$, $\frac{1}{\delta}$ and n.

Consider next a noisy variant of this framework, whereby the distribution P used to sample points from c is corrupted: with probability μ, the example x_i of the sample is chosen not according to the distribution P concentrated on c but according to a distribution Q_i over the entire X_n. The distribution Q_i can be chosen by a *malicious* adversary that knows the concept c, the distribution P, the learning strategy $\mathfrak{A}$.

DEF 10 *The concept class $\mathcal{C}$ is PAC-learnable* from positive examples only *with* malicious error rate $\mu : (\varepsilon, \delta, n) \in (0,1) \times (0,1) \times \mathbb{N} \mapsto [0, \frac{1}{2})$ *and* sample cardinality $m : (\varepsilon, \delta, n) \in (0,1) \times (0,1) \times \mathbb{N} \mapsto \mathbb{N}$ *if there is a* learning function

$$\mathfrak{A} : (\varepsilon, \mu, \mathbf{x}) \in (0,1) \times [0, \frac{1}{2}) \times X_n^m \mapsto \mathfrak{A}(\varepsilon, \mu, \mathbf{x}) \in \mathcal{C}_n \tag{10}$$

such that for any $\varepsilon, \delta \in (0,1)$, any $n \in \mathbb{N}$, any $\eta \in [0, \mu(\varepsilon, \delta, n))$, any concept $c \in \mathcal{C}_n$, any distributions P and $\overline{P}$ concentrated over c and $\overline{c}$, any additional m distributions $Q_1, \ldots, Q_m$ over X_n, condition (11) holds, where $m = m(\varepsilon, \delta, n)$, $\tilde{P}_k = (1-\mu)P + \mu Q_k$ and $\bigotimes$ is measure-product:

$$\bigotimes_{k=1}^{m} \tilde{P}_k \left\{ \mathbf{x} \in X_n^m \;\middle|\; \begin{array}{l} e_{c,P}(\mathfrak{A}(\varepsilon, \mu, \mathbf{x})) \leq \varepsilon \\ e_{c,\overline{P}}(\mathfrak{A}(\varepsilon, \mu, \mathbf{x})) \leq \varepsilon \end{array} \right\} \geq 1 - \delta \tag{11}$$

and furthermore the sample cardinality function $m(\cdot,\cdot,\cdot)$ grows polynomially in $\frac{1}{\varepsilon}$, $\frac{1}{\delta}$, and n

PAC-learnability from *misclassified examples* (Definition 8) allows the error rate η to vary arbitrarily between 0 and $\frac{1}{2}$. In the more demanding case of PAC-learnability with *malicious error*, η is only required to vary between 0 and the malicious error-rate $\mu(\varepsilon, \delta, n) < \frac{1}{2}$. I now show that:

RESULT 8 *The class $Q^{C,I}$ of conservative and invariant quantifiers is PAC-learnable from positive examples only with sample cardinality m and malicious error rate μ as follows:*

$$m(\varepsilon, \delta, n) \geq \frac{24n}{\varepsilon}\left((n+1) + \ln\frac{4}{\delta}\right), \quad \mu(\varepsilon, \delta, n) \leq \frac{\varepsilon}{8n} \tag{12}$$

The proof rests on the following result (Kearns & Li, 1993). Suppose the sample cardinality m is large, as in (13).

$$m(\varepsilon, \delta, n) \geq \frac{24}{\varepsilon}\log\left(\frac{4|C_n|}{\delta}\right) \tag{13}$$

Suppose furthermore that for some $\varepsilon, \delta \in (0,1)$ and $\eta \in [0, \frac{\varepsilon}{4})$, the learned concept $\mathfrak{A}(\varepsilon, \mu, \mathbf{x})$ assigns a positive label to many of the examples x_i in the sample $\mathbf{x} = (x_1, \ldots, x_n)$, as in (14).

$$\bigotimes_{k=1}^{m} \tilde{P}_k \left\{ \mathbf{x} \;\middle|\; |\{x_i | x_i \notin \mathfrak{A}(\varepsilon, \eta, \mathbf{x})\}| \leq \frac{\varepsilon}{2} m \right\} \geq 1 - \frac{\delta}{2} \tag{14}$$

Then, $\mathfrak{A}$ has small error relative to P in the sense that:

$$\bigotimes_{k=1}^{m} \tilde{P}_k \left\{ \mathbf{x} \;\middle|\; e_{c,P}(\mathfrak{A}(\varepsilon, \eta, \mathbf{x})) \leq \varepsilon \right\} \geq 1 - \frac{\delta}{2} \tag{15}$$

Proof. Consider the following learning function: $\mathfrak{A}(\varepsilon, \mu, \mathbf{x})$ is the generalized quantifier in $Q_{\mathfrak{u}}^{C,I}$ that contains the properties of cardinality $i \in \{0, \ldots, n\}$ iff the sample $\mathbf{x}$ contains at least $\frac{\varepsilon}{4n} m$ properties with cardinality i. As $\mathfrak{A}$ does not depend on μ, I will write just $\mathfrak{A}(\varepsilon, \mathbf{x})$. Obviously $|Q^{C,I}| = 2^{n+1}$, so that (12) entails (13). For any sample $\mathbf{x} = (x_1, \ldots, x_m)$, the quantifier $\mathfrak{A}(\varepsilon, \mathbf{x})$ classifies as negative at most $\frac{\varepsilon}{4n} m(n+1) < \frac{\varepsilon}{2} m$ of the m examples in $\mathbf{x}$, so that (14) holds too. By the result mentioned above, we thus have for any quantifier $\mathbf{q} \in Q^{C,I}$:

$$\bigotimes_{k=1}^{m} \tilde{P}_k \left\{ \mathbf{x} \in X_n^m \;\middle|\; e_{\mathbf{q},P}(\mathfrak{A}(\varepsilon, \mathbf{x})) \leq \varepsilon \right\} \geq 1 - \frac{\delta}{2}$$

The following chain of inequalities finally proves (11).

$$\begin{aligned}
&\bigotimes_{k=1}^{m} \tilde{P}_k \left\{ \mathbf{x} \in X_n^m \;\middle|\; e_{\mathbf{q},\overline{P}}(\mathfrak{A}(\varepsilon, \mathbf{x}))0 \right\} \\
&\leq \bigotimes_{k=1}^{m} \tilde{P}_k \left\{ \mathbf{x} \;\middle|\; \exists x \notin \mathbf{q} \text{ s.t. } x \in \mathfrak{A}(\varepsilon, \mathbf{x}) \right\} \\
&\leq \sum_{i=0}^{n} \bigotimes_{k=1}^{m} \tilde{P}_k \left\{ \mathbf{x} \;\middle|\; \begin{array}{l} \exists x \notin \mathbf{q} \text{ s.t.} \\ |x| = i, x \in \mathfrak{A}(\varepsilon, \mathbf{x}) \end{array} \right\} \\
&= \sum_{i=0}^{n} \bigotimes_{k=1}^{m} \tilde{P}_k \left\{ \mathbf{x} \;\middle|\; \begin{array}{l} \text{there is } x \notin \mathbf{q} \text{ s.t. } |x| = i \text{ and} \\ \text{the sample } \mathbf{x} \text{ contains at least} \\ \frac{\varepsilon}{4n} m \text{ properties of cardinality } i \end{array} \right\} \\
&\leq \sum_{i=0}^{n} \bigotimes_{k=1}^{m} \tilde{P}_k \left\{ \mathbf{x} \;\middle|\; \begin{array}{l} \mathbf{x} \text{ contains at least} \\ \frac{\varepsilon}{4n} m \text{ properties in } \overline{\mathbf{q}} \\ \text{of cardinality } i \end{array} \right\}
\end{aligned}$$

$$\leq \sum_{i=0}^{n} \bigotimes_{k=1}^{m} \tilde{P}_k \underbrace{\left\{ \mathbf{x} \middle| \begin{array}{l} \mathbf{x} \text{ contains} \\ \text{at least } \frac{\varepsilon}{4n} m \\ \text{properties in } \overline{\mathbf{q}} \end{array} \right\}}_{(*)} \leq \sum_{i=0}^{n} e^{-m\varepsilon/24n} \leq \frac{\delta}{2}$$

In the penultimate step, I have noted that the probability (*) is the probability that m Bernoulli trials each with a probability of success $\mu \leq \frac{\varepsilon}{8n}$ overall yield at least $\frac{\varepsilon}{4n} m$ successes, which is bound by $e^{-m\varepsilon/24n}$ through *Chernoff inequality*. □

Could a more sophisticated learning function than the one considered in the preceding proof lead to a stronger result? namely a higher noise tolerance or a smaller sample cardinality? No learning function from positive examples only for a concept class $\mathcal{C} = \bigcup_{n=1}^{\infty} C_n$ can tolerate a rate of malicious error $\mu(\varepsilon, \delta, n) \geq \frac{\varepsilon}{VCD(C_n)-2}$, as shown in Kearns and Li (1993). The concept class $Q^{C,I}$ has Vapnik-Chervonenkis dimension $VCD(Q_n^{C,I}) = n+1$. Hence, the largest tolerable rate of malicious error for $Q^{C,I}$ is of the order of $\frac{\varepsilon}{n}$. The learning function in the proof above thus tolerates the largest possible rate of malicious error.

Furthermore, no learning function satisfies the plain PAC-learnability condition (3) with sample cardinality $m(\varepsilon, \delta, n)$ smaller than $\frac{VCD(C_n)-1}{32\varepsilon}$, as recalled above. And Blumer et al. (1989) show that $m(\varepsilon, \delta, n)$ cannot be smaller than $\frac{4}{\varepsilon} \log \frac{2}{\delta}$ either. PAC-learnability thus requires:

$$m(\varepsilon, \delta) \geq \max \left\{ \frac{4}{\varepsilon} \log \frac{2}{\delta}, \frac{VCD(n) - 1}{32\varepsilon} \right\} \tag{16}$$

Thus, the learning function in the preceding proof meets the demanding condition of PAC-learnability *from positive malicious examples* while using a sample cardinality (12) that asymptotically exceeds only by a factor n the lower bound (16) needed for *plain* PAC-learnability.

Conclusion

I have looked at the conjecture informally made in the recent linguistic literature that universal restrictions on Natural Language determiners serve the purpose of simplifying the learning task. To start, I have looked at the *monotonicity* universal, and I have shown that it contributes only little to simplifying the learning task. This result shows the importance of investigating the conjectured link between universals and learnability within an explicit, formal learnability framework. I have then focused on the universals of *conservativity* and *invariance*. And I have provided support for the conjecture that they crucially simplify the learning task, by showing that the class $Q^{C,I}$ of conservative and invariant quantifiers has the property that the simplest and most straightforward learning strategy, namely the one considered in the proof of Result 8, is the optimal one, namely the one that tolerates the largest tolerable rate of malicious error. Furthermore, the class $Q^{C,I}$ has the property that the presence of noise (even malicious noise) does not require any substantial increase of the sample cardinality compared to the noise-free case.

Acknowledgments

I would like to thank G. Chierchia. This work was supported by a Marie Curie Fellowship (PIEF-GA-2011-301938).

References

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, *4*, 159–219.

Benthem, J. van. (1983). Determiners and logic. *Linguistics and Philosophy*, *6.8*, 447–47.

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1989). Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM*, *36.4*, 929–965.

Clark, A., & Lappin, S. (2011). *Linguistic nativism and the poverty of the stimulus*. Wiley-Blackwell.

Ehrenfeucht, A., Haussler, D., Kearns, M., & Valiant, L. (1989). A general lower bound on the number of examples needed for learning. *Information and Computation*, *82.3*, 247–251.

Fox, D. (1999). Reconstruction, binding theory, and the interpretation of chains. *Linguistic Inquiry*, *30.2*, 157–196.

Gamut, L. T. F. (1991). *Logic, language and meaning*. Chicago and London: The University of Chicago Press.

Heim, I., & Kratzer, A. (1978). *Semantics in generative grammar*. Blackwell Textbooks in Linguistics.

Hunter, T., & Lidz, J. (2013). Conservativity and learnability of determiners. *Journal of Semantics*, *30.3*, 315–334.

Kearns, M. (1998). Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, *45.6*, 983–10061.

Kearns, M. (1999). PAC-learning. In R. A. Wilson & F. C. Keil (Eds.), *The MIT encyclopedia of cognitive sciences.* Cambridge, MA: The MIT Press.

Kearns, M., & Li, M. (1993). Learning in the presence of malicious errors. *Journal on Computing*, *22.4*, 807–837.

Kearns, M., Li, M., & Valiant, L. (1994). Learning boolean formulas. *Journal of the ACM*, *41.6*, 1298–1328.

Kearns, M., & Vazirani, U. (1994). *An introduction to computational learning theory*. Cambridge, MA: The MIT Press.

Keenan, E. L., & Stavi, J. (1986). A semantic characterization of natural language determiners. *Linguistics and Philosophy*, *9*, 253–326.

Montague, R. (1973). The proper treatment of quantification in ordinary english. In P. Suppes, J. Moravcsik, & J. Hintikka (Eds.), *Approaches to natural language.* Reidel.

Natarajan, B. K. (1991). *Machine learning. A theoretical approach*. Morgan Kaufmann Publishers.

Pitt, L., & Warmuth, M. (1990). Prediction-preserving reducibility. *Journal of Computer and System Science*, *41.3*, 430–467.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5.2*, 197–227.

Valiant, L. (1984). A theory of the learnable. *Communications of the ACM*, *27.11*, 1134–1142.

Westerståhl, D. (1989). Quantifiers in formal and natural languages. In D. Gabbay & F. Guenthner (Eds.), *Handbook of philosophical logic* (Vol. 4, pp. 1–131). Reidel Publishing.