Research on Language and Computation **00:** 1–16, 2006.

On the Denotations of Anaphors

EDWARD L. KEENAN (ekeenan@ucla.edu)
Department of Linguistics, UCLA
3125 Campbell Hall, UCLA, Los Angeles, CA 90095-1543

July 7, 2006

1. Introduction

In the past 25 years we have learned much about the semantic nature of DPs (formerly NPs) by directly interpreting them as generalized quantifiers. See Keenan (1996) and Keenan and Westerståhl (1997) for overviews of this work. For example, we can provide rigorous and empirically reasonable (not perfect) answers to questions that arise independently in generative grammar, such as: (1) The subject DPs which license negative polarity items in the predicate are those that denote monotone decreasing functions. (2) The DPs which occur naturally in the post *of* position in plural partitives, as in *two of those students*, are those that denote principal filters. (3) The DPs which occur context neutrally in Existential There Ss are (boolean compounds of) those built from intersective Determiners (Dets).

This work has also led to new semantic generalizations, unsuspected and, with one or two exceptions, unformulable without direct interpretation: (1) Lexical (syntactically simple) DPs denote monotone functions, almost always monotone increasing; (2) DPs built from lexical Dets almost always denote monotone functions, usually monotone increasing; (3) DPs built from proportionality Dets (*most, two out of three, every third*) are not in general definable in first order logic, nor are they sortally reducible (*Most poets daydream* has no paraphrase of the form *most individuals*$_x$ followed by a boolean compound of *poet*(x) and *daydream*(x)). (4) Overwhelmingly natural language (NL) Dets denote conservative, domain independent functions.

Our goal in this paper is to provide a comparable direct interpretation of *anaphors*, such as *himself* in (1b).

(1) a. Not a single patient criticized every doctor
    b. Not a single patient criticized himself
    c. Not a single patient criticized his doctor

In (1a) the interpretation of the object DP *every doctor* is referentially independent of the interpretation of the subject DP *not a single patient*. The *P*1 (one place predicate) *criticized every doctor* is interpreted as the set of objects which bear the *CRITICIZE* relation to every doctor. That *P*1 denotation does not change when we replace the subject DP with others — *John, most of the patients*, etc. What we predicate of John in *John criticized every doctor* is exactly what we predicate of Bill in *Bill criticized every doctor*.

But not so in (1b). The person understood to be criticized in *John criticized himself* is not the same as the one understood to be criticized in *Bill criticized himself.* What we predicate of John in *John criticized himself* is, in effect, the property expressed independently by *criticized John.* Obviously, this is not what we predicate of Bill in *Bill criticized himself.* So there is an interpretative dependency of the predicate on the subject denotation. In fact since *criticized* is denotationally constant just *himself* is interpretatively dependent. It is this dependency we characterize below.

Ss like (1c) are ambiguous according as *his* (*doctor*) is interpreted as referentially dependent on the subject DP or not. It might be interpreted as dependent, like *himself* in (1b), but equally it might refer to a fixed individual, say John's doctor, in a context in which we have been discussing John. So *his doctor* is may be referentially dependent but is not obligatorily so.

Below we propose that anaphorically interpreted expressions are ones which denote functions of a certain sort, a sort that subsumes the generalized quantifiers (GQs) denoted by referentially independent DPs such as *John, no student, most of John's students*, etc. Direct interpretation deepens our understanding of NL anaphors, leading us to notice new properties they have as well as contributing to solutions to old problems, such as where their antecedents may lie. Here are four specific merits of this approach, not all of which we can enter in detail here.

[1] A definition of *anaphor* in terms of denotations is syntax independent and hence applicable cross linguistically. It enables us to formulate tests to decide if *ang sarili* in Tagalog, *kendisi* in Turkish, *cagi-casin-ul* in Korean are anaphors. And this puts us on methodologically surer ground in evaluating language general hypotheses concerning anaphors, such as:

- Do all languages present lexical anaphors? — Surely not: Old English (Keenan 2003), Tongan (Dukes 1996).
- Do languages present lexical anaphoric Dets (lexical expressions which combine with Ns to form anaphoric DPs)? — Likely so, Norwegian *sin*, Latin *suus*, and Russian *svoi.*
- Are lexical anaphors syntactically invariant (Keenan and Stabler 2003)? — Likely so: In any given grammar lexical anaphors have syntactically

distinctive properties so that systematic interchange with non-anaphors may fail to preserve grammaticality.

- Is the anaphor-antecedent relation asymmetric (meaning we cannot interchange anaphor and antecedent preserving meaning and grammaticality)? — Likely so (Keenan 1993).
- Are anaphors always (asymmetrically) c-commanded by their antecedents? — Arguably not, as in case marking languages like Korean and voice marking languages like Batak and Tagalog (Keenan and Stabler 2003). (These last two queries require a language independent definition of antecedent of).

[2] We provide a semantic basis for expecting (if not predicting) that anaphors will combine syntactically with referentially independent DPs to form complex anaphors:

(2) a. Each pupil criticized both himself and the teacher (Boolean compounds)

b. One worker criticized everyone but himself/no one but himself (Exception Phrases)

c. Zakanya ta kula de kwikwiyo-n kanta
Lioness she watch of cub-of herself
'The lioness watched over her own cub' (Possessors; Hausa, Brenda Clark, pc)

Languages with DP anaphors often permit them in possessor position — Hindi, Japanese, Georgian, Chinese, Hebrew, Korean, Uzbek, Hausa — though English does not: **John loves himself's mother.* Keenan (2003) offers an historical account of this latter fact.

[3] We provide a semantic basis for predicting that if expressions which are interpreted anaphorically as objects (of verbs) are either ungrammatical as subjects of lexical $P1$s or they are not interpreted as anaphors there. Most commonly they are interpreted deictically. Thus at least part of standard Binding Theory is semantically motivated (see Büring 2005 for a recent overview). The standard English **Himself laughed* is covered by this claim, as is Japanese *zibun* in (3). (Keenan 1988; N. Akatsuka, pc):

(3) a. Hanako-ga zibun-o utagatte-iru
Hanako-NOM zibun-ACC doubts
'Hanako doubts herself' or 'Hanako doubts Speaker'

b. Zibun-ga Hanako-o utagatta-iru
zibun-nom Hanako-acc doubts-ASP
Speaker doubts Hanako, *Hanako doubts herself.

[4] We show that anaphoric DPs impose strictly weaker conditions on their possible denotations than referentially independent DPs, with the

result that they allow massively more possible denotations. Then a generalization in Keenan (1987) supports that lexical anaphors (*himself, kanta*, etc.) exhibit little or no variability in denotation whereas lexical non-anaphoric DPs (*John, Mary*, etc.) can denote basically denote any object in the domain of a model.

2. DP denotations

We consider first the denotations of referentially independent DPs, such as *John, every student, most students, not more than half the students, more students than teachers*, ... They are usually assigned type $((e,t),t)$ and relative to a domain E of entities (always assumed here to have at least two elements to avoid degenerate cases) are interpreted as elements of $[P(E) \to \{0,1\}]$, the set GQ_E of *generalized quantifiers* over E. Here $P(E)$ is the power set of E, $\{0,1\}$ the (boolean) set of truth values with $0 = False$ and $1 = True$. In general $[X \to Y]$ is the set of functions with domain X and codomain Y. Here is a sample GQ (generalized quantifier): for every subset A of E, $EVERY(A)$ is that GQ which maps each subset B of E to 1 iff $A \subseteq B$, that is, iff each A is a B. Similarly, writing $|X|$ for the cardinality of the set X,

(4) $SOME(A)(B) = 1$ iff $A \cap B \neq \emptyset$
$NO(A)(B) = 1$ iff $A \cap B = \emptyset$
$MOST(A)(B) = 1$ iff $2 \cdot |A \cap B| > |A|$
$(THE\ TEN)(A)(B) = 1$ iff $|A| = 10$ and $A \subseteq B$
$(MOST\ OF\ THE\ TEN)(A)(B) = 1$ iff $|A| = 10$ and $MOST(A)(B) = 1$

So here DPs semantically map $P1$ denotations (subsets of E, here called *properties* or *unary relations*) to truth values ($P0$ = Sentence denotations). But clearly this is insufficiently general, as DPs combine with $P2$s (two place predicates) to form $P1$s, such as *criticize every doctor*, and more generally with $Pn{+}1$'s to form Pn's. So they should in general map $Pn{+}1$ denotations, $n{+}1$-ary relations, to Pn denotations, n-ary relations. This is in fact how DP denotations are given in Keenan and Westerståhl (1997). Here we consider just binary relations.

Given a referentially independent DP such as *every poet* we know how to compute the set denoted by the $P1$ which results from combining it as an object with a $P2$, such as *admire*. *Admire every poet* denotes the set of objects x which bear the $ADMIRE$ relation to every poet. That is, the set of objects x which the GQ denoted by *every poet* is true of

Definition 1
For R a binary relation over E and $a \in E$, write aR for $\{b \mid (a,b) \in R\}$. Then, for each generalized quantifier F, set $F(R) =_{df} \{a \mid F(aR) = 1\}$

Note that the value F assigns to the binary relation R is uniquely determined by the value that it assigns to the unary relations (which aR is). So to define such a function it suffices, as before, to give its value on the unary relations. In this way we treat ordinary DPs as denoting functions, still called *generalized quantifiers*, mapping unary relations to zero-ary ones (truth values), and binary relations to unary ones, satisfying the condition in the definition above. It is then obvious that generalized quantifiers satisfy the *Extensions Condition* (EC):

> **Extensions Condition** (EC):
> For F a GQ, R and S binary relations, and $a, b \in E$, if $aR = bS$ then $a \in F(R)$ iff $b \in F(S)$[1]

To check that a DP X satisfies the EC, check that:

(5) If the individuals John criticized are exactly those that Bill praised then *John criticized* X and *Bill praised* X have the same truth value.

For example to check that *most of Peter's cousins* (built from a Det that is not even first order definable) denotes a GQ, imagine a situation in which John criticized exactly the people that Bill praised. Then clearly *John criticized most of Peter's cousins* and *Bill praised most of Peter's cousins* have the same truth value – both true or both false. Thus *most of Peter's cousins* is a GQ denoting DP. So it can be interpreted as the sole argument of a lexical $P1$.

The test in (5) is one that can be effected under normal elicitation procedures. (Of course with naive speakers we cannot elicit metalinguistic judgments like "Do A and B have the same truth value?". But we can check that A entails B by asking things like: Look, suppose A. Then B, right? If your speaker agrees you can infer that A entails B. Then similarly check that B entails A and then infer that A and B are always true together).

The test in (5) also helps us realize how strong the EC is. It says that when X denotes a GQ the truth value of John criticized X remains unchanged if we replace John with any other individual denoting DP and criticized with any other transitive verb provided the new individual bears the new binary relation to the same things that John bears *CRITICIZE* to. Below

[1] And any F from binary to unary relations satisfying the EC uniquely determines a GQ F' by setting $F'(K) = 1$ iff $a \in F(\{a\} \times K)$. Further, extending the EC to $n+1$-ary relations is little more than a matter of notation. Let $a = (a_1, ..., a_n)$ be an n-tuple of elements of E, and for R an $n+1$-ary relation write aR for $\{b \mid (a, b) \in R\}$. Note that (a, b) here is $(a_1, ..., a_n, b)$. Then we take GQ_E to be the set of functions F with domain the union of the $n+1$-ary relations over E, codomain the union of the n-ary relations over E, and which satisfy the EC: for each $n+1$-ary relation R, $F(R)$ is the set of n-tuples a such that $F(aR) = 1$. In this way the value of a GQ F at any relation is determined by its values at the unary relations (the subsets of E).

we exhibit some DPs whose denotations fail the EC. Among them lie the anaphoric functions.

Anaphor denotations DPs such as *himself* fail the EC, as they fail the test in (5). Suppose that John criticized just Sam, Bill, Rob, Sue and Maud, and that these are exactly the people Bill praised. Then the sentence *John criticized himself* is false, but *Bill praised himself* is true. So *himself* fails (5) and thus cannot denote a generalized quantifier. But the denotation of *himself* shares with GQs the property of mapping binary relations to unary ones. The $P1$ *criticized himself* denotes the set of objects x which stand in the *CRITICIZE* relation to themselves. We can interpret *himself* as the function *SELF* from binary relations to unary ones given in (6):

(6) $SELF(R) = \{a \mid (a,a) \in R\}$

Then *Every poet admires himself* would be compositionally represented as in (7), which is true iff each individual poet stands in the *ADMIRE* relation to himself, which is correct.

(7) Every poet admires himself
$EVERY \quad POET \quad ADMIRE \quad SELF$
$EVERY(POET) \quad SELF(ADMIRE)$
$EVERY(POET)(SELF(ADMIRE))$
$= True$ iff $POET \subseteq \{a \mid (a,a) \in ADMIRE\}$

This interpretation of *himself* is sufficient to support one interesting claim concerning the logical expressive power of English:

(8) There is no generalized quantifier F such that $F(R) = SELF(R)$, all binary relations R over E

Proof Let $a \neq b \in E$. (E has at least two elements recall.) Set $R = \{\langle a,a\rangle\}$ and $S = \{\langle b,a\rangle\}$. Then $aR = bS = \{a\}$. So for F any GQ, either $a \in F(R)$ and $b \in F(S)$ or $a \notin F(R)$ and $b \notin F(S)$. But $a \in SELF(R)$ and $b \notin SELF(S)$, so $SELF \neq F$, and since F was arbitrary, $SELF$ isn't a GQ. □

Comparable claims hold for more complex anaphors, such as only himelf, every/no poet but himself, as defined below:

(9) a. $a \in (ONLY\ HIMSELF)(R)$ iff $aR = \{a\}$
b. $a \in (EVERY\ POET\ BUT\ HIMSELF)(R)$ iff $POET - aR = \{a\}$
c. $a \in (NO\ POET\ BUT\ HIMSELF)(R)$ iff $POET \cap aR = \{a\}$

By (9a) *John criticized only himself* iff the set of people John criticized is $\{John\}$. *John criticized every poet but himself* is true iff he is the only poet who he didn't criticize. He criticized no poet but himself iff he's the only poet he did criticize.

The somewhat surprising semantic generalization in (8) tells us that even in a language with the full class of expressible generalized quantifiers (Keenan and Stavi 1986) (such as *most of John's cousins*, *not more students than teachers*, etc.) the addition of *himself* strictly increases expressive power. It cannot be paraphrased by any GQ denoting DP.

But, in satisfaction of [2], object anaphor denotations such as *SELF* or those in (9) share with such DPs the semantic property of mapping binary to unary relations, so we can interpret boolean compounds of them and referentially independent DPs in object position just using the mechanisms already needed for non-anaphoric DPs. The same mechanisms that guarantee the logical equivalence of (9a,b,c) work for (10a,b,c) as well since the arguments of *AND* (*OR*, *NOT*) are just functions from binary to unary relations.

(10) a. Sam criticized every teacher and some student
b. Sam criticized every teacher and criticized some student
c. Sam criticized every teacher and Sam criticized some student

(11) a. Sam criticized himself and some student
b. Sam criticized himself and criticized some student
c. Sam criticized himself and Sam criticized some student

Let us return to the deeper question of just what the denotations of anaphors depend on. Per the EC, if we know that the set K of individuals John criticized is the same as the set that Bill praised we can infer that *John criticized every doctor* and *Bill praised every doctor* have the same truth value. But we cannot under these conditions infer that *John criticized himself* and *Bill praised himself* have the same truth value. This case requires that we know something about the composition of K, specifically whether it contains just one of John and Bill or else both or neither. So varying the subject holding the VP constant may change truth value. But truth value is preserved under mere change of transitive verb. If John praised just the people he (John) admires then *John praised himself* and *John admires himself* must have the same truth value. Thus anaphors are functions from binary to unary relations that satisfy the Anaphor Condition:

Anaphor Condition (AC):
For an F mapping binary to unary relations and $a \in E$, if $aR = aS$ then $a \in F(R)$ iff $a \in F(S)$

To check that a DP X satisfies the AC check the following:

(12) If the individuals Bob praised are exactly those Bob admires then *Bob praised X* and *Bob admires X* have the same truth value.

Again, (12) is testable in elicitation and thus together with (5) it enables us to check whether an occurrence of an expression in a transitive S is anaphoric or not, in support of the claim in [1].

The denotations of object occurrences of *himself*, *only himself*, *everyone but himself*, *both himself and the two students* satisfy the AC: if John praised just the people he admires then he praised (only) himself iff he admires (only) himself, he praised everyone but himself iff he admires everyone but himself, etc. So the distinctive property of an object anaphor is that it is interpreted as a function satsfying the AC but failing the EC. More explicitly:

Definition 2
For each domain E,

1. A function from binary relations to unary relations over E is *anaphoric* iff it satisfies the AC and fails the EC.
2. An interpretation of a DP occurrence is *anaphoric* iff that interpretation is anaphoric as defined in **1**, above
3. An occurrence of a DP is an *essential anaphor* iff it has non-trivial interpretations and all of them are anaphoric (as defined in **1**, above).

A "trivial" interpretation of an object DP is one which denotes either 0 or 1, where 0 maps all binary R to $\emptyset$, the empty set, and 1 maps them all to E, the domain of the model. Both 0 and 1 satisfy the EC. For example an object occurrence of *fewer than zero boys* always denotes 0. Hence no occurrences of it are anaphoric or essential anaphors.

In support of the insightfulness of **Def 2**, the reader may verify that object occurrences of himself and the DPs whose denotations are in (9) are essential anaphors. But **Def 2** relativizes the notion of an anaphor to occurrences and this forces us to restate, and refine, the queries in [1], which makes them more accurate but also more cumbersome. For example we might refine "Do all languages present lexical anaphors?" to a weak and a strong version. The former is "Do all Ls have lexical items with anaphorically interpreted occurrences?" Here *Yes* a likely answer. The stronger version would be "Do all Ls have lexical items with occurrences which are essential anaphors?" And even here we really want more. We do not really want to claim that English has essential anaphors merely on the grounds that *his mind* must be interpreted anaphorically in *John lost his mind.* What we are generally interested in are **grammatically defined** occurrences, as when we said above that **object** occurrences of *himself* are essential anaphors. By "object" here we mean a grammatical object of a $P2$ (or a Pn, $n \geq 2$). See Keenan and Stabler (2003) for a grammar independent definition of grammatically definable.

To justify why we relativize the notion of *anaphor* to occurrences reconsider the Japanese (3a) where *zibun* occurs as a subject, *-ga* marked, and is interpreted deictically as Speaker. So that occurrence is not anaphoric and not an essential anaphor. So we cannot say simply that *zibun* in Japanese is an anaphor, but at most only that certain occurrences are. Surprisingly perhaps somewhat comparable cases occur in English. Keenan (1988) (citing J. McCloskey pc) notes that Irish English presents Ss like (13a,b):

(13) a. Watch it, himself is in a bad mood today (said to co-worker arriving late)

b. Ed, hurry up.
Wait a minute. Herself is getting herself ready.

In (13a) *himself* occurs as a subject and refers deictically, to the prominent male in context (the boss). In (13b) the first occurrence of *herself* refers to the prominent woman in context, say Ed's wife, and the second occurrence is most naturally understood anaphorically. So we cannot just say in this dialect (one shared by some communities in New England) that *himself/herself* are anaphors. Whether they are anaphorically interpreted depends on the occurrence. Less obvious examples occur in standard English. Here are two. First, compare:

(14) a. That cyclist collapsed

b. Several friends of a certain cyclist and his son testified against that cyclist at the hearing

In (14a) the demonstrative DP *that cyclist* is interpreted deictically, as some cyclist in the context of utterance. But in (14b) *that cyclist* is interpretable anaphorically, bound by *a certain cyclist* inside the subject DP. In distinction to standard English *himself* but like Japanese *zibun* and Irish English *himself*, *that cyclist* may also occur as a subject interpreted deictically. So some occurrences of *that+N* are anaphoric and some aren't. And since *that cyclist* could be interpreted deictically in (14b) its occurrence there is not an essential anaphor.

A second, subtler, case was pointed out to me by Jason Mattausch (pc). Compare:

(15) a. Neighboring countries should maintain clear boundaries (between them)

b. Most dictatorships fear attacks from neighboring countries

In (15b) the occurrence of *neighboring countries* has a natural anaphoric interpretation meaning "countries which neighbor them", with *them* bound by *Most dictatorships*. But as a subject in (15a) it is neither interpreted anaphorically nor deictically, but rather as "countries which neighbor each other". This interpretation is also available, if less so, in (15b). So again we have a DP some of whose occurrences can be interpreted anaphorically but none which must be, so no occurrences it seems are essential anaphors. A last example, with an overt reciprocal, is (16).

(16) a. Men who dislike each other are running for President

b. Rosa and Zelda date men who dislike each other

The object occurrence of *men who dislike each other* in (16b) has an anaphoric interpretation, on which (16b) means "Rosa dates a man who dislikes some man who Zelda dates, and vice versa". In (16b) we can also interpret the object DP as it is interpreted in (16a), in which case it implies that Rosa dates men who dislike each other and so does Zelda, so its object occurrence in (16b) has anaphoric interpretations but is not an essential anaphor.

In this way then we characterize anaphors in terms of their denotations. We have to be sure our characterization is limited to arguments of $P2$s but we take this context as the most basic one for anaphors. If a language has any anaphors in any context it has them in this one. Thus we feel that we have provided a way of detecting anaphors in an arbitrary language. And since such anaphoric occurrences constrain the relative interpretations of two arguments of a predicate we have a semantic basis for expecting [3] that such expressions either will not occur as intransitive subjects or will be interpreted non-anaphorically there (since an intransitive verb does not have two arguments which can be constrained as anaphors require). Let us end with some reflections logical expressive power, [4].

3. Logical Expressive Power of DP Anaphors

We have already noted that adding DPs denoting anaphors to a language allows us to denote more maps from binary relations to unary ones that can be done with "mere" generalized quantifier denoting DPs. We naturally wonder whether all maps from binary relations to unary ones can be denoted by functions satisfying AC (including ordinary GQs). The answer is negative. One systematic counter example is (17).

(17) John knows more students than Bill (does)

Here the $P1$ *knows more students than Bill does* denotes the set of objects x such that the number of students that x knows is greater than the number Bill knows. But (18a) does not imply (18b):

(18) a. The indviduals John admires and the individuals John knows are the same

b. John admires more poets than Bill does iff John knows more poets than Bill does

(18a) makes no claim whatever concerning the number of poets Bill admires or knows, those figures can be any we like. So imagine a situation in which John knows just 10 people, all of whom are poets he admires. But Bill knows 20 poets and admires just five of them. In such a situation (18a) is true and (18b) false, as the two Ss on either side of the *iff* have different truth values. Thus the function F_b mapping binary relations R to properties in (19) does not satisfy AC:

(19) For all $b \in E$, $F_b(R) = \{a \mid aR \cap POET| > |bR \cap POET|\}$

So we know that anaphors — *SELF*, etc. — satisfy the AC and functions like F_b above fail the AC. So adding anaphors to a language with GQs increases expressive power but not unlimitedly — not just anything (of relevance) can be said using anaphoric functions. Below we measure just what the increase in expressive power is. Our interest in that claim is twofold. First, even over small domains the set of anaphoric functions is vastly greater than the number of generalized quantifiers (that is, the functions which satisfy the AC and fail the EC vastly outnumber those which satisfy the EC). And second, our way of counting the strict anaphors suggests an alternate way of representing them, one which approaches that used in Jacobson (1999).

Our new representation is inspired by our remark at the beginning of this paper that the property understood to hold of the subject in *John criticized himself* is the one expressed by *criticized John*, whereas in *Bill criticized himself* it is *criticized Bill*. So once an entity b is given as an argument of the $P1$ we interpret *himself* as the GQ determined by that entity (**Def 3** below). For example in *John criticized everyone but himself* we interpret the object as *everyone but John*, in *John criticized both himself and the teacher* we interpret the object as *both John and the teacher*, etc. In this way we represent anaphor denotations as functions mapping entities to GQs. To see how this works in practice, we first define the GQs which correspond to entities in E:

Definition 3
For all $b \in E$, I_b, the *individual generated by b*, is that GQ given by:
$I_b(A) = 1$ iff $b \in A$

We remind the reader that classically we treat John as denoting an element $b \in E$ and stipulate that *John smiled* (ignoring tense and aspect) is true iff b ∈ SMILE, the set denoted by *smile*. Now we interpret *John smiled* as the truth value that I_b maps *SMILE* to, and that is 1 iff $b \in$ *SMILE*, as before. Consider also how the individuals I_b extend to binary relations:

(20) $I_b(R) = \{a \mid I_b(aR) = 1\} = \{a \mid b \in aR\} = \{a \mid (a, b) \in R\}$

So *criticize John* denotes the set of objects a such that (a, John) is in the *CRITICIZE* relation. We now could define anaphors as the functions f^* below:

Definition 4
For $f \in [E \to [P(E) \to \{0,1\}]]$ define f^* from binary to unary relations by

$$f^*(R) = \{a \mid f(a)(aR) = 1\}$$

Now consider that we can think of *himself* as denoting **self***, where **self** maps each b to I_b. Then the compositional interpretation of *John criticized himself* yields the correct results:

(21)
$$\begin{array}{lll}
\text{John criticize himself} & & \\
I_j \quad \text{CR} \quad \mathbf{self^*} & & \\
I_j \quad \mathbf{self^*}(CR) & & \\
I_j(\mathbf{self^*}(CR)) & & \\
= 1 \text{ iff } j \in \mathbf{self^*}(CR) & \text{Def 3} & \\
\quad\text{iff } \mathbf{self}(j)(jCR) = 1 & \text{Def } f^* & \\
\quad\text{iff } I_j(jCR) = 1 & \text{Def } \mathbf{self} & \\
\quad\text{iff } j \in \{a \mid (a, j) \in CR\} & \text{Def 3} & \\
\quad\text{iff } (j, j) \in CR & \text{Set Theory} & \square
\end{array}$$

Similarly we obtain correct results letting **only self** be that function mapping each b to $\mathbf{only}(I_b)$, **every student but self** maps each b to **every student but** I_b, etc.

There is more to be said about a systematic notation here, but let us rather measure our increase in expressive power. We are representing anaphoric functions as the f^*, for each function f from E into GQ_E. We observe first that each such f^* satisfies the AC:

(22) Let $aR = aS$. We must show that a $\in f^*(R)$ iff $a \in f^*(S)$. Now

$$\begin{array}{lll}
a \in f^*(R) \text{ iff } f(a)(aR) = 1 & \text{Def of } f^* & \\
\qquad\quad\;\; \text{iff } f(a)(aS) = 1 & \text{Assumption aR = aS} & \\
\qquad\quad\;\; \text{iff } a \in f^*(S) & \text{Def } f^* & \square
\end{array}$$

And (23) below computes the number of such f^*, which is easy since the map sending each f to f^* is one to one[1], so

[1] *Proof* let $f \neq g \in [E \to [P(E) \to 0,1]]$. Let $f(b)(K) \neq g(b)(K)$. Then, setting $R = \{b\} \times K$, we have that $b \in f^*(R)$ iff $f(b)(bR) = 1$, iff $f(b)(K) = 1$. Similarly $b \in g^*(R)$ iff $g(b)(K) = 1$, so by the assumption $f^*(R) \neq g^*(R)$. □

(23) $|\{f^* \mid f \in [E \to [P(E) \to \{0,1\}]]\}| = |[E \to [P(E) \to \{0,1\}]]| = 2^m$, for $m = |E| \cdot 2^{|E|}$.

Now (22) shows that each f^* satisfies the AC, so the number of AC functions is at least 2^m as above. In fact it is exactly 2^m since

(24) **Theorem** Each F satisfying AC is an f^* for some f.[1]

Thus the set of f^* is exactly the set of maps from binary to unary relations satisfying AC. It is worth noting how fast the number of possible anaphor extensions outstrips those of referentially independent DPs (the GQs).

(25)

$\|E\| = n$,	$\|GQ\| = 2^k$,	$k = 2^n$,	$\|AC\| = 2^m$,	$m = n \cdot 2^n$	$\|AC - GQ\|$
2	$2^4 = 16$	$m = 2 \cdot 2^2$	$2^m = 2^8 = 256$		240
3	$2^8 = 256$	$m = 3 \cdot 2^3$	$2^m = 2^{24} = 16,777,216$		$16,776,960$

So even in a minuscule world of just three individuals there are just 256 possible extensions for referentially independent DPs but almost 17 million ones for anaphors! This may seem crazy, but that number is itself minuscule compared to 8^{512}, the total number of maps from binary to unary relations over an E with 3 elements.

Keenan (1987) observes that lexical items in a category whose denotation set is relatively "small" (in terms of the size of the domain E) exhibit much freedom with regard to which elements in that set they may denote. For example there seem to be no general logical restrictions on the subsets of E that can be denoted by common nouns or $P1$s. But such restrictions do emerge as we move to categories whose denotation sets grow as a hyperexponential function of n. ('hyperexponential' here just means that the size of the set is $\geq 2^k$, where k itself is exponential in n, minimally 2^n). For the record, writing $Den_E(C)$ for the set in which expressions of category C find their extensions in a situation with domain E,

Lexical Freedom Law
As $|Den_E(C)|$ increases lexical denotational freedom decreases

[1] *Proof* Let F satisfy AC and define f_F from $E \to [P(E) \to \{0,1\}]$ by setting $f_F(a)(K) = 1$ iff $a \in F(\{a\} \times K)$
We show that $f_F{*} = F$. Let R be arbitrary. Then

$$\begin{array}{lll} a \in f_F^*(R) & \text{iff } a \text{ in } f_F(a)(R) & \text{Def } *, (17) \\ & \text{iff } f_F(a)(aR) = 1 & \text{per line (3) in (21)} \\ & \text{iff } a \in F(\{a\} \times aR) & \text{Def } f_F \\ & \text{iff } a \in F(R) & \text{For } S = \{a\} \times aR,\ aS = aR;\ F \text{ is AC} \end{array}$$

Thus F is f_F^*, completing the proof. □

For example the denotation set for the referential DPs is $[P(E) \to \{0,1\}]$ which is hyperexponential in n, having size 2^k, for $k = 2^n$. But the lexical DP denotations largely lie in the set of individuals, the I_b's, a set of size n in a one to one correspondence with E. Similarly common noun modifiers (adjectives), of type $((e,t),(e,t))$, lie in $[P(E) \to P(E)]$, which has size 2^n raised to the power 2^n. But extensional adjectives are not interpreted freely in this set, they must denote restricting functions: $F(A) \subseteq A$, all subsets A of E. And all but a few extensional adjectives meet a stronger condition: they are *intersective*: $F(A) = A \cap F(E)$. And the set of intersective functions has cardinality 2^n, the same as $|P(E)|$, and so is not hyperexponential in n. The denotation set for Dets, of type $((e,t),(e,t,t))$, has cardinality 2^k for $k = 4^n$. Conservativity reduces this to 2^k for $k = 3^n$, still hyperexponential in n. But most lexical Dets are either logical constants – *every*, *some*, *no*, *most*, etc. and so have no freedom of denotation, or they are deictic – *my*, *your*, *this*, etc. so their denotations are uniquely determined by context of utterance.

The Lexical Freedom Law is not an accident. Complex expressions have their denotations determined compositionally, but lexical items have to be learned by brute force. So the larger the set in which they can denote the harder it is to know which of the possible denotations is theirs. So the constraints on lexical freedom serve a useful learning theoretic function, one that applies to our remarks about anaphors. The set of maps satisfying AC is hyperexponential in n, and the lexical anaphors, like most lexical Dets, are logical constants denoting specific invariant elements in their denotation set: *SELF* is provably invariant.[1]

4. Conclusion

We have illustrated what it is like to interpret object anaphors directly, rather than "translate them away" with variable binding operators. This has enabled us to gain some insight into the way they increase logical expressive power. But our remarks here are far short of a complete theory of DP anaphora. For that we would have to treat anaphors in a greater diversity of syntactic positions, such as those in (26) and (27) below.

(26) a. Martha protected Billy from himself

b. Martha protected Billy from herself

c. Sam protected Billy from himself (antecedent ambiguous)

(27) a. Each student thought that no one but himself would get an A on the exam

b. Each student tackled a problem that only himself and the teacher could solve

c. No one likes to work with anyone smarter than himself (antecedent ambiguous)

[1] Invariant elements of a denotation set are those fixed (mapped to themselves) by all the automorphisms of the primitives E and $\{0,1\}$ of a model. Such an automorphism h is the identity function restricted to $\{0,1\}$ and any permutation of E. h extends to all denotation sets in a canonical way (see Keenan and Westerståhl 1997). For example its value at a subset A of E is $\{h(a) \mid a \in A\}$. It's value at a binary relation R over E is $\{(h(x), h(y)) \mid (x,y) \in R\}$, etc. Provably then the only invariant elements of $P(E)$ are E and $\emptyset$; the only invariant binary relations are $\emptyset$, $E \times E$, id, and $\neg id$. Now let h be a permutation of E. By the way h extends to elements of the type hierarchy built from E and $\{0,1\}$ we have that $h(SELF)$ is that map sending each $h(R)$ to that map sending each $h(a)$ to $h(SELF(R)(a))$. To show that $SELF$ is invariant we must show that $h(SELF) = SELF$. Let R and a be arbitrary. Then

$$\begin{aligned}
h(SELF)(hR)(ha) &= h(SELF(R)h(a)) && \text{Def extension of } h\\
&= h(SELF(R)(a)) && \text{Def extension of } h\\
&= SELF(R)(a) && h \text{ is the identity on truth values}\\
&= 1 \text{ iff } (a,a) \in R && \text{Def } SELF\\
&= 1 \text{ iff } h(a,a) \in hR && \text{Def extension of } h\\
&= 1 \text{ iff } (ha,ha) \in h(R) && \text{Def extension of } h\\
&= 1 \text{ iff } SELF(hR)(ha) = 1 && \text{Def } SELF
\end{aligned}$$

Thus $h(SELF)$ and $SELF$ take the same values at $h(R)$, and since every binary relation S is an $h(R)$ for some R, namely $R = h^{-1}(S)$, the proof is complete. □

Our "*" notation would have to be extended systematically and compared with more standard variable binding.[1] And the notion *antecedent of an anaphor* would have to be defined in denotational terms in order to respond to some of the queries in [1]. So, hardly surprising, much remains to be done from this perspective.

References

Büring, D.: 2005, *Binding Theory*, Cambridge Textbooks in Linguistics. Cambridge University Press.

Dukes, M.: 1996, 'On the Non-existence of Anaphors and Pronominals in Tongan'. Ph.D. thesis, UCLA Department of Linguistics.

Jacobson, P.: 1999. *Linguistics and Philosophy* **22**, 117.

Keenan, E., E. , and D. Westerståhl: 1997, 'Generalized quantifiers in Linguistics and Logic'. In: J. van Benthem and A. ter Meulen (eds.): *Handbook of Logic and Language*. Amsterdam: Elsevier, pp. 837–893.

Keenan, E.: 1987, 'Lexical Freedom and Large Categories'. In: J. Groenendijk, D. de Jongh, and M. Stokhof (eds.): *Studies in Discourse Representation Theory and the Theory of Generalized Quantifiers*. Dordrecht: Foris.

Keenan, E.: 1988, 'On Semantics and the Binding Theory'. In: J. Hawkins (ed.): *Explaining Language Universals*. Oxford: Basil Blackwell, pp. 105–144.

Keenan, E.: 1989, 'Semantic Case Theory'. In: B. et al (ed.): *Semantics and Contextual Expression*. Dordrecht: Foris, pp. 33–57.

Keenan, E.: 1993. In: U. Lahiri and Z. Wyner (eds.): *Proceedings of Semantics and Linguistic Theory III*. p. 117.

Keenan, E.: 1996, 'The semantics of determiners'. In: S. Lappin (ed.): *The Handbook of Contemporary Semantic Theory*. Oxford: Blackwell, pp. 41–63.

Keenan, E.: 2003, 'An historical explanation of some binding theoretic facts in English'. In: J. Moore and M. Polinsky (eds.): *The Nature of Explanation in Linguistic Theory*. Stanford: CSLI, pp. 153–189.

Keenan, E. and E. Stabler: 2003, *Bare Grammar: Lectures on Linguistic Invariants*. Stanford: CSLI.

Keenan, E. and J. Stavi: 1986. *Linguistics and Philosophy* **9**, 253.

[1] Worth noting here is that our treatment of **self*** as a map sending an entity b to the individual I_b it generates is not substantially different from Jacobson's intuition that unbound pronouns denote the identity function. Beyond this the two approaches differ, in that she treats *his mother* as denoting the function denoted by *the mother of*, locating the binding mechanisms in an operation (z-lifting) on predicates. We build it into the meaning of *his*.