Natural Language Corpus Semantics: The Free Choice Controversy

Kjell Johan Sæbø
University of Oslo*

Abstract

This paper addresses the methodological issue of what data can safely be disregarded in the quest for a good analysis of a given semantic phenomenon. I argue that as there is no objective measure for this, making the temptation to adjust the terrain to the map hard to resist, it is strongly advisable to make use of text corpora in formal semantic research. I illustrate this argument by considering in some detail the recent history of research on Free Choice items.

1 Introduction: From Data to Facts in Semantics

It has not been customary to base formal semantic analyses on corpus studies. On the contrary, many studies in formal semantics have been based on a small number of select examples, simplified or constructed. To be sure, this has some good reasons. However, a formal semantic analysis is very sensitive to the constitution of the facts. Analyses, and ultimately theories, are dependent on the interpretation of the data. And the interpretation of the data in turn depends on the selection of the data. Of course, a conscientious selection may ensure a sound coverage. But since there is no standard of conscientiousness in this regard, there is a risk that accounts are influenced by a biased empirical basis.

I will first review some reasons that idealization is necessary in formal semantics. I will then consider some ways in which it is dangerous to rely on selective data, and argue that corpora can help us avoid such pitfalls. In Section 2, these points are illustrated through the controversy over the proper analysis of Free Choice items.

1.1 Simplification and Simplicity

Simplification is a necessary tactic in semantic research. Regardless of the topic, it is essential to disregard distracting factors and to abstract away from independent problems. As formulated by Lyons (1991: 3):

> The second reason [that semantics is not concerned with the totality of linguistic meaning] is that linguists by necessity idealize the phenomena selected and considered as data. They consider utterances in a methodologically and theoretically specific perspective. In fact, linguistics can be subdivided into several overlapping subdisciplines, both regarding the phenomena under consideration and regarding the methodological abstractions determining their scientific treatment.

*This paper is based on work in the project *Comparative Semantics for Nordic Languages*, NORDSEM, and in particular on my work with Sæbø (1999) and (2001). I owe a great debt to my fellow researchers in the project, in Oslo, Gothenburg, and Copenhagen, and to Lauri Carlson, Östen Dahl, and Cathrine Fabricius Hansen.

When doing formal semantics, we simplify at two levels – at the level of performance, concerning data, and at the level of competence, concerning facts. We simplify data to form a coherent set of facts and we simplify facts to fit a theory. Both are covered by the term **idealization** (cf. e.g. Partee 1979: 2f.).

There are some good reasons for idealizing facts and data in formal semantics. The theoretical ideal of formal elegance (Partee 1995: 348) constitutes a pressure on us to concentrate on "core" aspects of a phenomenon, hoping that the analysis will eventually extend to those aspects that remain ill-understood. It may be true that "the tools with which semantic structures can be individuated and described on the model-theoretic approach are many and varied, since there is no a priori limit on the metalanguage with which models are constructed or described" (Partee 1993: 16), but there is a pull towards recycling tools which have done good work before and thus to view some facts as better than other facts. It is all right for our analysis to not fit all the facts as long as it fits the right, the "core" facts.

To find the facts, it is moreover necessary to simplify the data. We abstract away from irrelevant detail and inessentials. The average length of an example sentence is of course well below that of a real sentence. We cannot solve all problems at once, and some problems are arguably pragmatic, not semantic, in nature. Furthermore, due to compositionality and what might be termed the "fragment heritage", the data we do introduce should be tractable. Phenomenona which are not very well understood, or where the proper treatment is controversial or very complex, tend to be avoided in combination with the phenomenon we are interested in. As DPs, for instance, proper names are overrepresented. In sum, we do not have to consider all data, as long as we do consider the right, the "core" data.

1.2 Deduction and Seduction

But what are the "right data", and what are the "right facts"? There is a tension between empirical coverage and theoretical quality. Of course, as formulated by Partee (1995: 348): "If we can find analyses with a high degree of both formal elegance and empirical generality, we can suspect that we are on the right track." But sometimes, we can only find analyses with a high degree of formal elegance, and the question is then what degree of empirical generality is sufficient to sustain our suspicion that we are on the right track.

The criteria for what counts as "core" data and "core" facts are varied, but one criterion seems to stand out: Core data are clear and plentiful data, data which are easy to interpret and which can easily be multiplied, and core facts are facts based on such data. Such data must be represented, and such facts must not be set aside without serious qualms.

However, it may be far from trivial to know what constitutes a set of core data, and to draw the line between the constitution of data and the constitution of facts. It may be difficult to decide which data belong together, and what data are clear. Selecting a set of simple sentences, you cannot be sure you do not miss out some factor with which the phenomenon under scrutiny interacts in an interesting way, or generally an interesting class of cases. Making choices on these issues may involve an interpretation of the data, influenced by hypotheses you already have in mind.

The conscientious semanticist will of course strive to establish a balanced set. Most progress made in semantic analysis has been based on data chosen with care. The semanticist usually knows the language and draws data from a range of sources, often impressionistically, but conscientiously, with an acute sense of truth to reality and awareness of potential problems. But since there is no universally acclaimed standard of conscientiousness, there is no way of knowing whether or not she has subconsciously been led to select amenable data.

Insofar, the situation in semantics is no different from the situation in syntax. However, data selection is an all the more critical point in semantics because here, the facts are less black and white, making the constitution of the facts a more complex affair. The question is not just whether a sentence is acceptable, but also whether it entails another sentence; indeed, in the last instance, what it means. Moreover, it can be difficult to distinguish between the contribution of an item and the contribution of the context where the item more or less regularly occurs and thus to know how much meaning to attribute to the item. Also, it may be difficult to draw the line between data and facts: Judgments about meaning, which are data but which are also themselves interpretations, easily fade into interpretations of themselves, i.e. constitution of facts.

For these reasons, it is crucial to take care that the examples are representative. Otherwise, we can have a "nocuous idealization": A covert simplification of facts, hidden behind selective data. Data selection can be a way of interpreting data and constituting facts without admitting it.

1.3 In Corpore Sano

According to Partee (1995: 322), compositional semantic analysis is typically a matter of working backward from intuitions about sentences' truth-conditions and reasoning our way among alternative hypotheses; referring to Cresswell (1982), Partee refers to intuitions about sentences' truth-conditions as the most concrete data we have. But if concreteness is a criterion, it might be added that the most concrete data one can have are intuitions about the truth-conditions of concrete sentences – such found in a corpus.

We will see in Section 2 that even clear and plentiful data can be disregarded when the semanticist is not required to consider a random sample and to recognize every substantial class of clear cases. Studying a corpus, chances are better that old hypotheses must be revised or new ones are inspired. Corpora offer an insurance against nocuous idealizations, where relevant aspects are disregarded and core facts are missed, by laying bare the relation between raw and interpreted data so it is open for inspection; and they offer a constructive means of assisting the imagination, guiding the researcher towards facts which would otherwise not be thought of.

The obvious way to ascertain representativity in examples is to consider a corpus, to base the semantic analysis on a corpus study. Of course, a corpus can never give us a piece of negative data, and it can never provide us with a minimal pair. Strictly, a corpus does not even offer data, for semantic data include judgments. But it can offer realistic points of departure for constructing minimal pairs and negative data, and it can supply circumstantial evidence of intuitions about truth conditions.

A corpus can be used in various ways, of course. The impressionistic method, where you look till you find what you like or want, can be contrasted with a "naturalistic", systematic method, which carries a commitment to count whatever emerges as core data. This latter method is what is meant by actually basing an analysis on a corpus study. Using this, you are not only insured against subconsciously or semiconsciously discounting core data, you can also hope to receive better impulses towards a fruitful hypothesis than from the examples you could invent – reality often surpasses the researcher's imagination.

Below, I present evidence of these points in a specific domain; a case study where corpora can be attested to make a definite difference. I review the treatment of one class of phenomena in formal semantics, showing how, indeed, analyses have been based on atypical data and fail to generalize to typical data and illustrating the need for basing the constitution of facts in a semantic domain on corpus studies. I will also discuss some methodological issues concerning the treatment of corpus data for eliciting semantic judgments – primarily, how to use substitutions.

2 The FC Controversy

Free Choice items (FCIs) are items like the English determiner *any* as it occurs in (1) or the English pronoun *anybody* as it occurs in (2).

(1) In the first year they were married, Ted had gone to work for a lithographer;
he'd instantly hated the job. "Ted would have hated *any* job,"
Marion told Eddie. (WIDOW:62)

(2) In short, she was not a woman who could be seduced by *anybody*;
yet Ted Cole wasn't just anybody, and he couldn't suppress
his unpredictable attraction to her. (WIDOW:125)

The qualification "as it occurs in ..." is necessary because English *any*(–) has another function as a Polarity Sensitive item (PSI), as in (2a). In (2b), *anybody* is ambiguous between the two functions.

(2) a. She was a woman who had not been seduced by *anybody*.
b. She was a woman who could not be seduced by *anybody*.

The formulation "items like ..." is very vague and reflects a very real vagueness. It is not evident that the semantic category of FCIs is cross-linguistically constant. Some items – in languages like Greek or Spanish – that have been labelled FCIs in the literature (Giannakidou 1997, Quer 1999) seem to have as much in common with the PSI *any*(–) or the item *some*(–).

Even when we concentrate on English, there is no consensus on the analysis. The following two basic questions about FCIs remain controversial:

- What are the constraints on their distribution?

- Do they have a universal quantificational force?

That there are constraints on the distribution is shown by (2c). The controversy is over where to draw the line between this and, say, (2d) and how to formulate the correct generalization.

(2) c. # She was a woman who had been seduced by *anybody*.
d. She was a woman who could be seduced by *anybody*.

In (1) and (2) above, the FCI would seem to express a universal quantification. However, it has been argued that this is only apparent (Kadmon and Landman 1993). Sometimes, the universality can instead be attributed to other factors, and it would be satisfying from a theoretical point of view to generalize this picture.

Below, I will review the use of data in the recent debate over these issues. The primary methodological interest in this lies in the fact that the theoretically preferable analysis just alluded to can be shown to be empirically in tune with no more than half the data. This I will not show for English directly, but for Scandinavian, more accurately, for the hypothesis as carrying over to Scandinavian.

In Scandinavian, there is a paradigm of FCIs which seems to correspond very closely to the English *any*(–) semantically, only that there is no PSI interpretation: The *wh*–, question word based paradigm is distinct from PSIs. Scandinavian FCIs furthermore constitute a good testing ground for a systematic use of corpora because they are extremely easily retrievable. I will concentrate on Mainland Scandinavian, on Norwegian and Swedish, when I confront the stances taken in the FC controversy with results from corpus studies underlying Sæbø (1999) and Sæbø (2001).

The first issue I consider is whether FCIs are more like universals or more like existentials, or indefinites – the Universalist versus the Existentialist hypothesis, and how corpus data can be taken to support one or the other hypothesis.

2.1 Universalism versus Existentialism

As shown by Horn (1999), there is a long ongoing debate over whether FCIs – first and foremost, English *any*(–) – is basically a universal or an existential (or indefinite), and both theses have been defended in various ways. As it seems, some data are more amenable to one view, others to another. Since most or all sentences where an FCI occurs have a universal interpretation, the "existentialist" has to show that this is not inherent in the item but comes from something in the context, or at least that it comes about indirectly, not through the item qua quantifier. First, I review the most influential recent existentialist contribution, Kadmon and Landman (1993), arguing that FCIs are parasitic on generics.

2.1.1 FCIs as Generic Parasites

The two uses of the determiner *any*, as a PSI and as an FCI, seem to be related, so from a semantic minimalist perspective, it is desirable that they be described as variants of one item with one meaning. This must basically be the meaning of the indefinite article, for the PS variant is clearly indefinite. According to Kadmon and Landman (1993), FC *any* is, just like PS *any*, an indefinite with additional semantic and pragmatic characteristics. The problem is that FC *any* seems to have a universal meaning, but this universality is attributed to a generic quantification outside *any*. More specifically, when the set denoted by the *any* argument seems to be subject to a universal quantification, this quantification does not stem from the determiner but from a generic operator or quantifier, which may be covert. Thus the *any* phrase gets a bound interpretation in the same way as (other) indefinites can get a bound interpretation. Davison (1980) had already proposed this strategy, but Kadmon and Landman worked it out more thoroughly.

To account for the difference between *a* and *any*, they propose that *any* – whether PS or FC – induces a widening of the interpretation of its argument and that this widening creates a stronger statement. This theoretical aspect of their theory I have assessed elsewhere (Sæbø 2001). Here I concentrate on the empirical prediction for FC *any*:

> NPs with FC *any* are allowed in the same kind of environment where generic indefinites are allowed. (Kadmon and Landman 1993: 357)

Kadmon and Landman give 7 examples of FC *any*, (3)–(9) (plus some variants, such as indicating accents: "Any PROFESSIONAL dancer ...").

(3) Take any apple.
(4) Any owl hunts mice.
(5) Any dog gives live birth.
(6) Any match I strike lights.
(7) I would dance with anybody.
(8) Any lawyer could tell you that.
(9) Any professional dancer would be able to do it.

In the introduction to the section on FC *any*, they write:

> The discussion ... will be limited to cases of FC *any* like [(4)], that is, to simple generic statements. [...] We think that the analysis extends to modal cases like [(8)], but the particular problems posed by the interaction between *any* and modals go beyond the scope of this paper. [...] Another type ... that we don't talk about ... is that of *any* in directives, like [(3)]. (Kadmon and Landman 1993: 405f.)

(As for the imperative case, this poses a problem for any theory of FCIs, and I will not pursue it here; just note (10) and (10a).

(10) Say something in Dutch to me, anything at all. (WIDOW:376)

(10) a. ? Say anything in Dutch to me.)

The empirical prediction quoted above – that NPs with FC *any* are allowed in the same kind of environment where generic indefinites are allowed – implies that

- it makes sense to substitute *any* for *a* whenever the latter is bound by a generic operator, and that
- substituting *a* for FC *any* makes sense and results in an interpretation where the indefinite is bound by a generic operator.

Both predictions can be falsified; in fact, the second can be falsified by utterances of sentences like (7)–(9) above, with an overt counterfactual operator, such as (1) above, or (11):

(11) Ruth resigned herself to the irony of reading a murder mystery; but, at the moment, Ruth would have read *anything* to escape her own imagination. (WIDOW:381)

(11) a. ? Ruth resigned herself to the irony of reading a murder mystery; but, at the moment, Ruth would have read a thing to escape her own imagination.

(11a) is odd no matter what is accented; *would*, *a*, or *thing*. Kadmon and Landman explicitly limit the discussion to cases like (4), "simple generic statements". Here, or in a sentence like (12), the substitution of the indefinite article does make sense and result in a bound interpretation.

(12) Ruth was still struggling to keep her memories of the past under control, as any widow must. (WIDOW:468)

(12) a. Ruth was still struggling to keep her memories of the past under control, as a widow must.

However, when we turn to Scandinavian, Norwegian and Swedish, and a sample of 1.000 cases from a general source corpus of either language, it turns out that this paradigmatic case is here rare. On the other hand, close to 50% of the occurrences of FCIs are modal cases like (8), or, more accurately, possibility modal contexts, most of which the generic analysis does not extend to.

In the below classification and quantification, the inaccurate estimates do not just reflect confidence intervals but also the intrinsic vagueness of the classes:

Context	**Proportion**
Possibility	40–50%
Lexical Modality	10–20%
Conditionals or Generics	10–20%
Comparative or Similative Constructions	10–20%
Negative Predicatives, Swedish Specialties	5–15%

Table 1: Distribution of Scandinavian *wh*(–) *som helst* FCIs

In the largest class, the FCI cooccurs with some possibility modal, most often *kan* ('can' or 'may'). The label "lexical modality" refers to cases where the FCI cooccurs with some word with a modal element, like *tåle* ('tolerate') or *beredd* ('ready'). "Conditionals or generics" are contexts with a conditional operator, like *skulle* or *ville* ('would'), or some predicate inducing a conditional or generic interpretation, often through a presupposition, like *adlyde* ('obey') or *nöjd* ('satisfied'). For the last two classes, see e.g. Sæbø (1999: 91ff.).

Now in a majority of the possibility contexts, the substitution of an indefinite fails to result in an interpretation where this is bound by a generic operator; often the sentence becomes uninterpretable. This is to some extent correlated with the variant of possibility, or, with reference to the theory of Kratzer (1991), the type of conversational background. We encounter the full spectre: Epistemic, normative, circumstantial, corresponding to epistemic possibility, permission, and ability, or disposition. Evidently, the indefinite substitution can make sense when the conversational background is normative or circumstantial, but often it does not:

(13) Där kan man använda vilken boll som helst, bara domaren godkänner den. (S)
there can one use which ball as rathest only referee-the licenses it
a. ? Där kan man använda en boll, bara domaren godkänner den.
there can one use a ball only referee-the licenses it

(14) Naturligtvis går det att spela Bach på vilka instrument som helst. (S)
naturally works it to play Bach on which instruments as rathest
a. ? Naturligtvis går det att spela Bach på instrument.
naturally works it to play Bach on instruments

(15) Sovjetunionen har raketter som kan nå et hvilket som helst mål på jorden. (N)
Soviet Union has missiles that can reach a which as rathest target on earth
a. ? Sovjetunionen har raketter som kan nå et mål på jorden.
Soviet Union has missiles that can reach a target on earth

And when the conversational background is epistemic, it seems that the indefinite substitution is bound to fail. Sentences like (16a) or (17a) do have an interpretation, but not a generic interpretation, and they are much weaker than the FCI originals.

(16) Et hvilket som helst menneske kan skjule seg bak dette kodenavnet. (N)
a which as rathest human may hide REFL *behind this codename*
a. Et menneske kan skjule seg bak dette kodenavnet.
a human can hide REFL *behind this codename*

(17) Drosjen kan ha vært fra et hvilket som helst selskap i Oslo. (N)
cab-the may have been from a which as rathest company in Oslo
a. Drosjen kan ha vært fra et selskap i Oslo.
cab-the may have been from a company in Oslo

In these cases, then, there is no independent evidence for a covert generic operator binding the FCI. One could argue that the FCI induces a genericity less available with the regular indefinite, but this would come close to saying that the FCI has an inherent universality after all, and to circularity.

Even outside the possibility contexts, FCIs are allowed in many environments where generic indefinites are not. Elliptic conditionals – simple sentences with a conditional structure – divide into two classes, one where the indefinite substitution makes sense and another where it does not. This seems to depend on whether the FCI is the only accented item: If it is, the indefinite version is uninformative.

(18) Den samme situasjonen ville ha oppstått under en hvilken som helst annen trener. (N)
the same situation would have developed under a which as rathest other coach
a. Den samme situasjonen ville ha oppstått under en annen trener.
the same situation would have developed under another coach
b. Den samme situasjonen ville ha oppstått under en hvilken som helst trener.
the same situation would have developed under a which as rathest coach
c. ? Den samme situasjonen ville ha oppstått under en trener.
the same situation would have developed under a coach

(19) Bente Skari ville ha hevdet seg i en hvilken som helst idrett. (N)
Bente Skari would have succeeded in a which as rathest sport
a. ? Bente Skari ville ha hevdet seg i en idrett.
Bente Skari would have succeeded in a sport

This seems to indicate that the counterfactual operator does not bind an indefinite the way a generic operator would and that the FCI does not get its universality from that operator.

We can conclude that Scandinavian FCIs are allowed in many – and "many" in a proportional sense – environments where generic indefinites are not, contrary to Kadmon's and Landman's postulate as applied to Scandinavian. If we now turn to the other side of this postulate, we discover that in many contexts where generic indefinites are allowed, Scandinavian FCIs are not. It is not true that it makes sense to substitute an FC phrase for an indefinite whenever the indefinite is bound by a generic operator. There are two stumbling stones for this substitution, and both seem to show that the FCI has a quantificational force of its own.

First, substitution in this direction makes us aware that while indefinites can have a bound interpretation both in simple and complex sentences, a reminiscent interpretation of the FCI only seems possible in simple sentences. An indefinite introduced in an *if* or a *when* clause is commonly considered to be bound by an overt or covert conditional or generic operator; but an FCI does not readily lend itself to such an interpretation. The fact that sentences like (20a) or (21a) tend to be odd suggests that the FCI is a quantifier, sensitive to scope islands.

(20) Hvis en plante reagerer på lyset, spiller temperaturen ingen rolle. (N)
if a plant reacts on light-the plays temperature-the no role
'If a plant reacts to light, the temperature plays no role.'
a. ? Hvis en hvilken som helst plante reagerer på lyset,
if a which as rathest plant reacts on light-the
spiller temperaturen ingen rolle.
plays temperature-the no role

(21) Vanligvis kommer det mange meldinger når en bjørn er i traktene. (N)
usually comes it many messages when a bear is in tracts-the
'Usually, there are many reports when a bear is in the area.'
a. ? Vanligvis kommer det mange meldinger når en
usually comes it many messages when a
hvilken som helst bjørn er i traktene.
which as rathest bear is in tracts-the

Second, an indefinite can, in a simple sentence, be bound by a covert or an overt generic or frequency operator, such as *normally* or *usually*; but, as has been pointed out by Dayal (1998), it seems that an FCI cannot be bound by an overt generic operator or an adverb of quantification. In (22a), in contrast to (22), the adverb does not quantify over letters but only over possible circumstances or courses of events; in (23a), where the predicate is individual-level, the FCI does not seem to be allowed, to make sense at all. This can be taken to indicate that the adverb is not a genuine generic operator in that it cannot quantify only over worlds.

(22) Et A-postbrev er normalt framme dagen etter. (N)
an A-post-letter is normally arrived day-the after
'A priority letter normally arrives the second day.'
a. ? Et hvilket som helst A-postbrev er normalt framme dagen etter.
a which as rathest A-post-letter is normally arrived day-the after

(23) En hovedstad er normalt en gammel by. (N)
a capital is normally an old city
'A capital is usually an old city.'
a. # En hvilken som helst hovedstad er normalt en gammel by.
a which as rathest capital is normally an old city

To sum up, it seems clear that the view that FCIs borrow their universality from a generic operator is, as applied to Scandinavian, empirically ill-founded.

2.1.2 Universalism: Parallel Evidence

Sidetracking for a moment from the use of Scandinavian data in the debate over whether FCIs are existentials or universals, we may note that a parallel corpus between a language like English and a language like German, where there are no clear FCIs, can provide circumstantial evidence in favor of the universalist hypothesis. Strikingly many occurrences of FC *any* are in fact translated by a regular universal determiner; indeed, specifying that *any* correspond to *jed-* or *all-* turns out to be an efficient means of retrieving FC cases without excluding too much.

This pattern cuts across all the major classes of contexts identified in 2.1.1; we encounter possibility, lexical modality, conditionals or generics, and similatives:

(24) Mechanically it is not much more complicated than a sewing machine,
and any reasonably competent blacksmith can repair it.
a. Die Mechanik ist kaum komplizierter als bei einer Nähmaschine, und
jeder Hufschmied, der sich auf sein Handwerk versteht,
kann es reparieren.

(25) There was a timid side to his character that made him tolerate any
ideology provided it left him in peace.
a. Ein furchtsamer Zug seines Wesens brachte ihn dazu, jede Ideologie
zu tolerieren, vorausgesetzt, dass sie ihn in Frieden liess.

(26) The more imaginative resorted to mixtures that would have been the envy
of any alchemist.
a. Die Phantasievolleren griffen auf Mixturen zurück, die der Neid eines
jeden Alchimisten gewesen wären.

(27) Like any other Cabbalist he believed that every event was already written
down in the Torah.
a. Wie jeder andere Kabbalist glaubte er, dass jedes Ereignis bereits in
der Tora niedergeschrieben sei.

Of course, the interpretation of such data poses special methodological problems. For one thing, one has to acknowledge that in a number of cases, the universal in the German version is supplemented by some modal item, like *beliebig* or *denkbar*, as in (28a) (where the universal phrase is further supplemented by an adjunct expressing indiscrimination) and (29a) (where the adjective is in turn modified by the particle *nur* 'only'); and even in (26a), the determiner is not just *jed-* but the complex indefinite-universal (not well-understood) *ein- jed-*.

(28) Rawlings could take just about any lock in manufacture.
a. Rawlings konnte jedes beliebige Schloss knacken, ganz gleich
um welches Fabrikat es sich handelte.

(29) And the men were bribed with cigarettes to do any favours one required.
a. Und die Männer wurden mit Zigaretten bestochen, einem jeden nur
denkbaren Gefallen zu tun.

And even the cases where the determiner translating *any* is a simple universal only show that it is possible to express (approximately) the same meaning with a regular universal as with *any* in such contexts where *any* is felicitous; not, strictly, that *any* has a universal meaning – it might be, say, as argued by Davison (1980), Kadmon and Landman (1993) and others, that the context expresses a universality and the FC element in FC *any* serves to widen or weaken the restriction for this universality, thus strengthening the statement. In the absence of such a widening or weakening item, it could be argued, German has to resort to a universal determiner even though it is not the most efficient means. However, it remains that the parallel data from a language which has to choose between existentials and universals do not provide evidence that FCIs are more closely related to the former than to the latter – rather the opposite.

2.2 Beyond Existentialism and Universalism

It might be possible to analyze FCIs neither as (special) existentials (indefinites) nor as (special) universals but as something different altogether, from which a universal interpretation could be derived in a more indirect way. In this section, I examine two such third roads, both consisting in analyzing the FCIs as a kind of definites: The scalar analysis proposed by Lee and Horn (1994), based on a paraphrase with a superlative, and the preferential analysis suggested, in particular, by the Scandinavian FC morphology, based on a paraphrase with a verb of intention. I show that ultimately, facts based on corpus data argue against both approaches.

2.2.1 The scalar hypothesis

Lee and Horn (1994) propose an analysis of FC *any* in terms of scalar implicature. Scalar implicature is a source of universal quantification, so if scalarity can prove useful in the analysis of FCIs it will account for the universality associated with them without actually treating them as universal quantifiers; regarding English *any* this is of course desirable. Many cases in the Scandinavian material lend themselves to a scalar interpretation. However, the hypothesis that FCIs are inherently scalar seems to meet too many counterexamples for a general analysis to be based on it.

The hypothesis that FC *any* is inherently scalar implies that a DP *any N* can be paraphrased by the DP *even the A-est N* for an adjective *A*. The choice of *A* will depend on the (intrasentential) context, primarily the verb in the sentence, in such a way that the superlative form of *A* will entail that the sentence frame is (not) true for every *N*.

To be sure, there are many cases of *(e-) (h)vilk- (...) som helst* conforming to this pattern in the Scandinavian corpora. Here are two Swedish examples:

(30) De lär kunna slå vilket lag som helst. (S)
they seem can beat which team as rathest
a. They seem to be able to beat even the best team.

(31) Jag är beredd att ta vilket straff som helst utom dödsstraff. (S)
I am prepared to take which penalty as rathest except death
a. I am prepared to take even the hardest penalty except death.

However, there are many cases where an appropriate adjective is hard to identify, because the relevant entities are not ranked on a scale, even when contextual information is taken into account. We may choose one and try to force a ranking along the corresponding scale, but a scalar implicature will not be generated or if it is generated it will fail to bring about a universal interpretation. Consider, first, (32):

(32) Man kan göra bordsdrycker av i princip vilken frukt som helst. (S)
one can make table drinks of in principle which fruit as rathest
a. One can make soft drinks from even the hardest fruit.

The adjective chosen in (32a) is as good a candidate as any, but we can easily imagine alternatives, like *sourest*. Regarding soft drink making, fruits are not ordered according to just one but to several scales, and in consequence, the superlative fails to generate a scalar implicature which covers all the cases; from (32a) we can conclude that we can make soft drinks from a soft fruit but not that we can make soft drinks from an *A* fruit for any other adjective *A*. Now (32) is a case where a scalar paraphrase has some plausibility; the larger context might supply sufficient information to narrow down the variation to one dimension. But in what seems a majority of cases, the choice of an adjective seems completely arbitrary and a paraphrase with a superlative gives a barely interpretable sentence, like (33a):

(33) Det går att spela Bach på vilket instrument som helst. (S)
it goes to play Bach on which instrument as rathest
a. ? Bach can be played on even the smallest instrument.

We could try to account for such cases by ranking the entities according to likelihood, a notion that has been used for a general analysis of *even*, choosing a superlative like *most unlikely*. This may yield reasonably good paraphrases, but the problem is that the notion of likelihood is so vague as to render the analysis rather vacuous: The paraphrase would be designed to ensure a universal interpretation. It may be added that the superlative is in itself not a primitive notion but a notion in need of analysis, and an ultimate analysis can reasonably be assumed to involve a universal quantification.

2.2.2 A Literal Interpretation: Preferential Paraphrases

Swedish has a particularly comprehensive paradigm of *wh- som helst* FC items. The items *hur som helst*, 'how as rathest', and *hur A (N) som helst*, where *A* is a gradable adjective and *N* is a noun, have no direct counterparts in Norwegian. We may ask what is used in this function in Norwegian. One answer is that we choose a paraphrase in terms of a verb of intention. This is part of a more general pattern, suggesting a way to analyze the FC phrases as a sort of definites from which a universality could be derived, much as on the scalar hypothesis discussed above.

It is often possible to rephrase *wh- som helst* by a preferential expression; in the case of *hur som helst*, by an equative construction 'as (A) (as)... V', where ... is a pronoun and V is a verb expressing an intention, typically 'want'. At the same time, *wh- som helst* is of course literally 'wh- as rathest', that is, the canonical FCI paradigm is based on a similative or relative conjunction and an adverb expressing preference. It would seem that paraphrases in terms of preference might point to a semantic regularity. (34a) and (35a) exemplify the pattern:

(34) Kvinnor kan klä sig hur som helst. (S)
women can dress REFL *how as rathest*
a. Kvinner kan kle seg som de vil.
(35) Vi kan fortsätta den här blockaden hur länge som helst. (S)
we can continue this embargo how long as rathest
a. Vi kan fortsette denne blokaden så lenge det skal være.

As mentioned, such paraphrases have a wider use than suppleting the Norwegian *wh- som helst* paradigm. More generally, *wh- som helst* can in Swedish or Norwegian often be replaced by 'th- (as)... want' or by 'wh- (as)... want' – that is, by an item composed of either the *wh* word or a corresponding demonstrative determiner, pronoun, or adverb, maybe a relative (or similative) conjunction, a DP (usually a pronoun coreferring with another nominal, usually the subject, in the sentence), and, finally, *vil* (N) or *vill* (S) or another verb expressing an intention. In (36) and (37), such a phrase alternates with *wh- som helst* without a change in meaning:

(36) De kan fritt si sin mening om hva de vil. (N)
they can freely say their opinion about what they want
a. De kan fritt si sin mening om hva som helst.
(37) Kvinner kan spille volleyball i det antrekket de vil. (S)
women can play volleyball in the outfit they want
a. Kvinner kan spille volleyball i et hvilket som helst antrekk.

There are limits to the use of preferential phrases in this sense. For one thing, they seem to require possibility contexts. Moreover, they seem to be sensitive to the conversational background for the possibility modal. Thus a *wh- som helst* item cannot be replaced by a preferential phrase if the modal is used in an epistemic sense. There are evidently still other, subtle restrictions concerning the connection between the conversational background and the intentions of the subject of the preferential phrase, responsible for the slight change in meaning if we substitute such a phrase for *wh- som helst* in (38) or (39):

(38) Hun kan spille hvor som helst i forsvarsfireren. (N)
she can play where as rathest in defence quartet
a. Hun kan spille hvor hun vil i forsvarsfireren.

(39) Han går med elektronisk sender så han kan spores hvor som helst. (N)
*he goes with electronic transmitter so he can trace-*PASS *where as rathest*
a. Han går med elektronisk sender så han kan spores hvor han vil.

Despite these restrictions, we might try to centre an analysis of FCIs in general around the cases where the substitution of a preferential phrase is possible, hoping that such an analysis would generalize to the rest of the cases. After all, there are preferential phrases in a wider sense that would work well in cases like (35), (38), and (39): The verb *vil* with a personal subject is replaced by *skal* with an impersonal subject, and the verb 'be' is added. This locution has a use outside possibility contexts in a narrow sense, cf. (40).

(38) b. Hun kan spille hvor det skal være i forsvarsfireren.
she can play where it shall be in defence quartet

(39) b. Han går med elektronisk sender så han kan spores hvor det skal være.
*he goes with electronic transmitter so he can trace-*PASS *where it shall be*

(40) Private bedrifter i helsesektoren betaler gjerne omtrent hva det skal være
private companies in health sector pay gladly about what it shall be
for å rekruttere denne type arbeidskraft. (N)
to to recruit this type workpower

We might say that while *vil* expresses an intention in the subject, *skal* expresses an intention in someone else; this someone could be the coach ((38)), the agent ((39)), and the personnel making wage demands ((40)). Thus we could maintain that in many cases, the FCI can be explicated by a phrase with a more transparent meaning. We could form a strategy of basing an analysis of *wh- som helst* on the literal interpretation of phrases like, for (37), 'the outfit they want to play volleyball in'; depending on the context, the paraphrase might vary, but the form of it as a definite description with an expression of choice would be constant.

There are, however, strong reasons not to follow such a strategy. First, the locution with *vil(l)* is not restricted to agentive or even animate nominals: Particularly in 'how' cases, we encounter a use of this preferential phrase that shows it to be strongly grammaticalized and unsuited for a literal analysis:

(41) En aksje kan være så attraktiv som den bare vil, det som… (N)
a share may be as attractive as it just wants, that which…

(42) Man kan ha vilka anlag för alkoholism man vill, utan alkohol… (S)
one can have which dispositions for alcoholism one wants, without alcohol…

(43) Sen får huset vara hur jordbävningssäkert det vill. (S)
then may house-the be how earthquakesafe it wants

Second, and even more critically, it can be shown (Sæbø 1999: 22) that a literal reading of the preferential phrase is definitely too weak to capture the meaning it has when it can supplant *wh- som helst* and function as a Free Choice expression. What we can conclude is that any preferential phrase that can replace an FCI has acquired a special meaning in the relevant contexts, less general and grammatical than wh som helst but more general and grammatical than what we would expect.

2.3 The Modal Context Generalization

So far, we have seen how theoretically attractive ideas, inspired by a limited set of selected or constructed examples, can be disproved by considering corpus data in some breadth. In a sense, this is a destructive, however necessary, use of corpus data in semantics. In the following, I will show how, conversely, it is possible to vindicate a theoretically attractive idea which has been criticized in the literature on the basis of artificial data, by considering a wider set of real examples in their contextual variety. This idea is the generalization, which can be traced back at least to Vendler (1967), that FCIs require a modal (intensional) context. Here, corpus data can be used to support a constitution of the facts which promises a simpler theory, attesting to a constructive or even creative use of such data in semantics.

2.3.1 Covert Conditions

Everybody agrees that FCIs have a limited distribution, but different scholars draw the lines differently. Thus Kadmon and Landman (1993), as we saw in 2.1.1, formulate a very strict condition, while Dayal (1998) is relatively liberal. The issue is often whether FCIs do require intensional contexts (Carlson 1981); the tendency, at any rate, is for them to occur in such contexts. One problem with determining the limits to the distribution of FCIs is that *any* functions both as a PS and as an FC item, and it may be difficult to discriminate between the two. This is one point where it is useful to consult a language where FCIs are lexically distinct from PSIs.

Carlson (1981) discusses the licensing environments of FC *any* and considers a characterization of them as intensional (modal) contexts. He notes that no overt modal need be present: Generic sentences, like (4), with no overt modal, sanctions *any*. "One could argue... that there is an unspoken modal... in such examples. There is a variety of other contexts, though, for which no such arguments can be made. Many stative verbs, such as *like*, many adjectives, and all predicate nominals allow *any*." (Carlson 1981: 10) Carlson brings the examples (44)–(46).

(44) Bob likes anyone.
(45) Any dog is reasonably intelligent.
(46) Any cat is a mammal.

Carlson is thus led to characterize the licensing contexts of FC *any* as either intensional ones or individual-level argument positions. In the light of later research on generics, it seems reasonable to group (45) and (46) together with (4) as generic sentences involving a covert modal. (44), however, does not seem to show the same genericity, so the only licensing feature here would seem to be the individual-level object position of the verb *like*.

However, a corpus search for FCIs like *anyone* in the object position of verbs in the simple present in sentences with no overt modal shows that it is very difficult to find verbs that are individual-level with respect to their object. This casts doubt on the assumption that the fact that *like* is individual-level with respect to its object is responsible for the fact that the FCI is licensed here as well. In fact, the verbs that are retrievable and that are similar in meaning to *like* are stage-level with respect to their object but have a habitual, conditional interpretation when the object is an FC phrase, with a covert habitual operator and an implicit conditional antecedent. This suggests a reinterpretation of examples like (44) along similar lines. Cases in point are the Norwegian predicates *ligge med* and *gå til sengs med* 'sleep with':

(47) Det er bare ett sted hvor... jentene ligger med hvem som helst: Ibiza! (N)
it is only one place where... girls-the lie with who as rathest: Ibiza
(48) Sharon går ikke til sengs med hvem som helst. (N)
Sharon goes not to bed with who as rathest

If we adopt a universalist analysis of the FCI, (47) could be paraphrased as follows: 'For every girl x and every person y, if y wants x to sleep with y, x sleeps with y.' Similarly, (44) could be paraphrased 'for every person x, if Bob meets x, he likes x'. Other verbs facilitate a conditional interpretation in carrying a presupposition which is read as an antecedent, e.g. *godta* 'accept' or *betale* 'pay':

(49) Det er slutt på at politiet godtar hva som helst. (N)
it is end on that police-the accept what as rathest
'The police have stopped accepting just anything.'

(50) Jeg betaler hva som helst for operasjonen. (N)
I pay what as rathest for operation-the
'I'll pay anything for the operation.'

Still other verbs, like *passe* 'match' or *tåle* 'endure', clearly have a modal element in their meaning, forming an intensional context for the FCI. The pattern that emerges is that what may seem an extensional context, in this case, a verb like *like*, turns out, on closer inspection, to fall into line with verbs involved in more complex, implicit intensional contexts. In this way, a more systematic study of data than has been customary can contribute to a reaffirmation of the generalization that FCIs require intensional contexts.

2.3.2 Subtrigging

This generalization has been argued to fail for another class of cases as well: Dayal (1998) emphasizes and in fact bases her analysis of FC *any* on the observation that this item is licensed in extensional contexts if only the NP is postmodified (this is known as "subtrigging"). Her examples include (51a–d).

(51) a. # John talked to any woman.
b. John talked to any woman at the party.
c. Yesterday John talked to any woman he saw.
d. John talked to any woman who came up to him.

There is no doubt that such a postmodification can be essential for the felicity of an FCI; on the basis of an extensional context where the FCI is infelicitous, it can create a context where the FCI is felicitous. The question is whether the postmodification saves the extensional context or whether it makes the otherwise extensional context intensional. Dayal gives the former answer: The context remains extensional but the statement is no longer "doomed to be false" (Dayal 1998: 453). – A search for the Norwegian equivalents of *anybody* or *anything* modified by a relative clause – search string *som helst som* (*as rathest that*) – yields cases like (52) and (53).

(52) Josva var en modig mann som utførte hva som helst som Herren påla ham. (N)
Joshua was a brave man that outcarried what as rathest that Lord onlay him
b. ? Josva var en modig mann som utførte hva som helst.

(53) Den brutale virkelighet var at hvem som helst som tok seg en skitur i
the brutal reality was that who as rathest that took REFL *a skitrip in*
avsidesliggende områder på Krokskogen og andre avsides områder 1944-45,
remotelying areas on Krokforest-the and other remote areas 1944-45
gjorde det med livet i hendene. (N) *did it with life-the in hands-the*
b. # ... at hvem som helst gjorde det med livet i hendene.

These cases are interesting because they suggest that the relative clause facilitates a modal interpretation by providing material to form a conditional antecedent from. To the extent that (52b) is interpretable, it has a conditional interpretation where the presupposition of the verb 'carry out' is accommodated into the antecedent.

But this interpretation is more available in the original sentence, and the reason seems to be that the relative clause ‘that the Lord enjoined him to’ adds descriptive content to the antecedent by satisfying the presupposition.

As for (53b), the expression ‘did it’ is too poor in descriptive content for the presupposition to be accommodated. In the original sentence, the presupposition is satisfied by the relative clause, and the resulting interpretation is arguably a conditional interpretation where the relative clause expresses the antecedent.

In short, authentic cases where subtrigging plays a role in apparently nonmodal contexts seem to point in the direction that the effect can be traced to the need for a modal context, insofar as the postmodification can supply the content material for an otherwise too implicit restrictor of a covert binary modal, conditional or generic, operator. Thus the modal context generalization is again vindicated.

3 Conclusions

The review of the recent history of research on Free Choice items has attested some reasons for using data from natural language corpora in natural language semantics, and some ways of going about it.

In this area, influential hypotheses, like the Existentialist hypothesis, have been based on small handfuls of constructed examples. This might be innocuous if only the examples were representative and did not abstract away from essentials; then the hypotheses would in fact be based on a broader range of data, of which, however, only an essence would be represented. This does not turn out to be the case, though.

Any independent evidence for a hypothesis like the Existential hypothesis must be based on a substitution argument: It must be possible to substitute a “regular” existential (indefinite) for the FCI and retain a universal interpretation – not the same interpretation, of course, but at least a reminiscent interpretation. But this fails in several classes of contexts. The proportional weight of such contexts makes it difficult to overlook this failure once corpus data are considered systematically. A similar criticism is valid for related hypotheses, such as the Scalar hypothesis.

The Existentialist hypothesis is a theoretically attractive hypothesis, promising a weak theory, so it is understandable that corpus data are not considered at once. But it is important to note that corpus data do not necessarily have a falsificationary function. Concerning the other central controversy in the area, the question whether the items require intensional contexts or not, corpus studies can be seen to support a generalization which has been rejected on the basis of a few constructed examples, and thus indirectly to render a strong theory redundant. Here, the study of corpora serves a constructive purpose, suggesting how counterexamples can be subsumed under the general constraint.

As I have argued elsewhere (Sæbø 2001), it is possible to unite a “Universalist hypothesis” (FCIs have a quantificational force) and an “Intensionalist hypothesis” (FCIs require an intensional context) in a coherent theory. This theory may have its problems, but it does show how a corpus based constitution of facts can feed a formal semantic analysis and how the corpus basis can make a difference.

Sources

- Swedish data from Språkbanken, University of Gothenburg
- Norwegian data from Norsk Tekstarkiv, University of Bergen
- English data from John Irving, *A widow for one year*, Random House
- English–German data from The English–Norwegian Parallel Corpus, Oslo

References

Carlson, Greg (1981) "Distribution of Free-Choice *Any*", in Papers from the 17th Regional Meeting of the CLS, 8–23.

Cresswell, Max (1982) "The Autonomy of Semantics", in Peters and Saarinen (eds.) *Processes, beliefs, and questions*, Dordrecht: Reidel, 69–86.

Davison, Alice (1980) "*Any* as universal or existential?", in van der Auwera (ed.) *The Semantics of Determiners*, London, 11–40.

Dayal, Veneeta (1998) "*Any* as Inherently Modal", in Linguistics and Philosophy 21, 433–476.

Giannakidou, Anastasia (2001) "Linking sensitivity to limited distribution", in Proceedings of the 11th Amsterdam Colloquium, 139–144.

Giannakidou, Anastasia (2001) "The Meaning of Free Choice", to appear in Linguistics and Philosophy.

Horn, Laurence (1999) "*any* and *(-)ever*: Free choice and free relatives", to appear in Wyner (ed.) Proceedings of the Israeli Association for Theoretical Linguistics 15.

Kadmon, Nirit & Fred Landman (1993) "*Any*", in Linguistics and Philosophy 16, 353–422.

Kratzer, Angelika (1991) "Modality", in Stechow and Wunderlich (eds.) *Semantics: An Interdisciplinary Handbook of Contemporary Research*, Berlin: de Gruyter, 639–650.

Lee, Young-Suk & Laurence Horn (1994) "*Any* as an indefinite plus *even*". Ms., Yale University.

Lyons, John (1991) "Theories of Meaning", in Stechow and Wunderlich (eds.) *Semantics: An Interdisciplinary Handbook of Contemporary Research*, Berlin: de Gruyter, 1–24.

Partee, Barbara (1979) "Semantics – Mathematics or Psychology?", in Bäuerle, Rainer, Urs Egli and Arnim von Stechow (eds.) *Semantics from different points of view*, Berlin: Springer, 1–14.

Partee, Barbara (1993) "Semantic Structures and Semantic Properties", in Reuland, Eric and Werner Abraham (eds.) *Knowledge and Language.* Vol. 2: *Lexical and Conceptual Structure.* Dordrecht: Kluwer. 7–29.

Partee, Barbara (1995) "Lexical Semantics and Compositionality", in Osherson, Daniel (ed.) *An Invitation to Cognitive Science.* 2nd ed. Cambridge, Mass. Vol. 1: *Language*, ed. Lila Gleitman and Mark Liberman. 311–360.

Quer, Josep (1999) "The quantificational force of free choice items", paper presented at the Colloque de Syntaxe et Semantique de Paris.

Sæbø, Kjell Johan (1999) *Free Choice Items in Scandinavian*, NORDSEM Report 4, University of Gothenburg.

Sæbø, Kjell Johan (2001) "The Semantics of Scandinavian Free Choice Items", to appear in Linguistics and Philosophy.

Vendler, Zeno (1967) "Each and Every, Any and All", in Vendler, Zeno: *Linguistics in Philosophy*, Ithaca, 70–96.