Building a Corpus of Indefinite uses Annotated with fine-grained Semantic Functions

{M.D.Aloni, A.W.vanCranenburgh, Raquel.Fernandez}@uva.nl
Marta.Sznajder@lrz.uni-muenchen.de
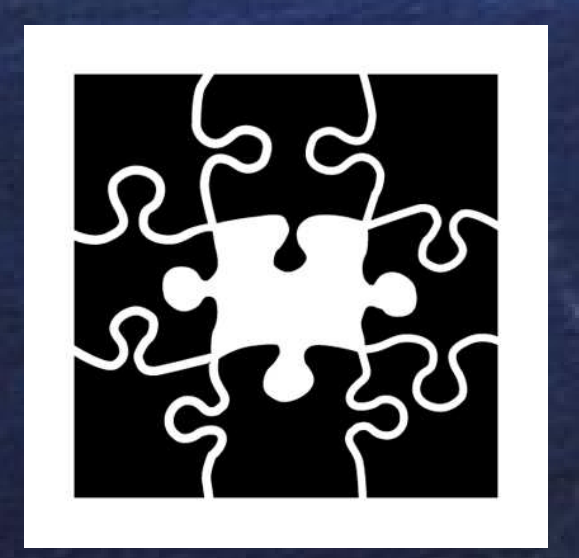
Institute for Logic, Language and Computation, University of Amsterdam

Functions of indefinites
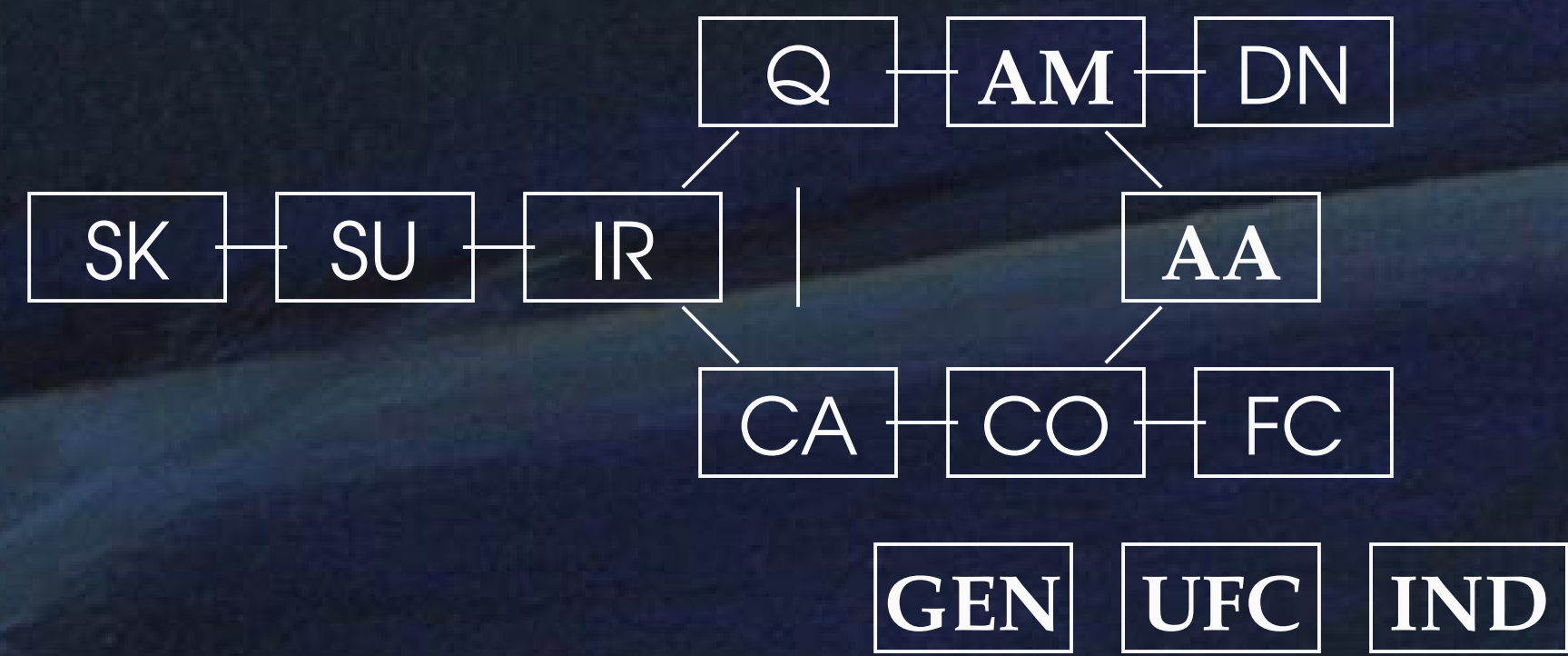
Extended version of Haspelmath's (1997) map

abbr.	function	Example
SK	specific known	*Somebody* called. Guess who?
SU	specific unknown	I heard *something*, but I couldn't tell what it was.
IR	irrealis	You must try *somewhere* else.
Q	question	Did *anybody* tell you anything about it?
CA	conditional antecedent	If you see *anybody*, tell me immediately.
CO	comparative	John is taller than *anybody* else.
DN	direct negation	John didn't see *anybody*.
AM	anti-morphic	I don't think that *anybody* knows the answer.
AA	anti-additive	The bank avoided taking *any* decision.
FC	free choice	You may kiss *any* woman.
UFC	universal free choice	John kissed *any* woman with red hair.
GEN	generic	*Any* dog has four legs.
IND	indiscriminative	I don't want to sleep with just *anybody* anymore.

Example

"You can pick any example".

Decision tree proposed by Aguilar-Guevara et al. (2011)

Guidelines

[a] *specificity* (S+/−): does the given indefinite refer to a specific individual in the actual world? NO ⇒ S−
[c] *universal meaning* (∀+/−): Does the given sentence $\mathbf{Op}(\ldots \text{indefinite} \ldots)$ imply the universal sentence $\forall x\ \mathbf{Op}(\ldots x \ldots)$? YES ⇒ ∀+
[e] *anti-additivity* (AA+/−): Does it hold that $\mathbf{Op}(a \vee b) \Rightarrow \mathbf{Op}(a) \wedge \mathbf{Op}(b)$? YES ⇒ AA+
[g] *negative meaning* (Neg+/−): Is $\mathbf{Op}(a \vee \neg a)$ inconsistent? NO ⇒ Neg-
[j] *free choice* (FC+/−): Is there a free choice between certain (maybe opposite) alternatives? YES ⇒ $FC+$

Method

- 5 annotators
- 80 instances of *any*, 20 instances of *some*
- Guidelines included detailed descriptions of tests.
- Additional label "unclear".

Disagreements

- No annotator violated the functional contiguity hypothesis.
- Most of disagreements about *any* concerned the functions added to the original map.

Per label scores

Q	85.0 %	FC	48.15 %
IND	83.33 %	IR	39.58 %
DN	73.26 %	UFC	37.5 %
CO	67.5 %	GEN	28.57 %
AA	66.67 %	AM	20.45 %
CA	64.0 %	SU	8.82 %
SK	61.90 %		

Accuracy for individual functions, obtained by accumulating agreement scores among all possible pairs of annotators.
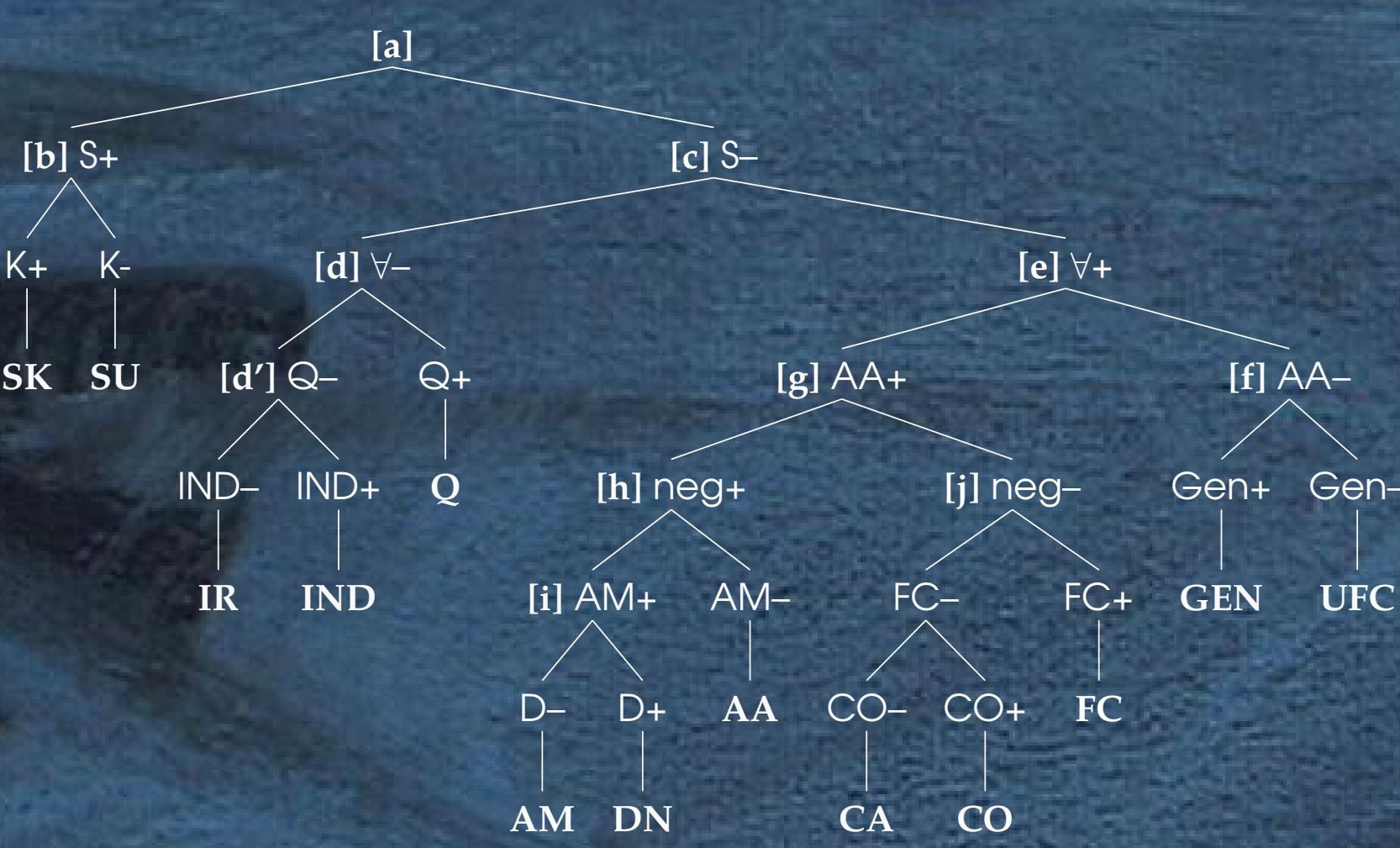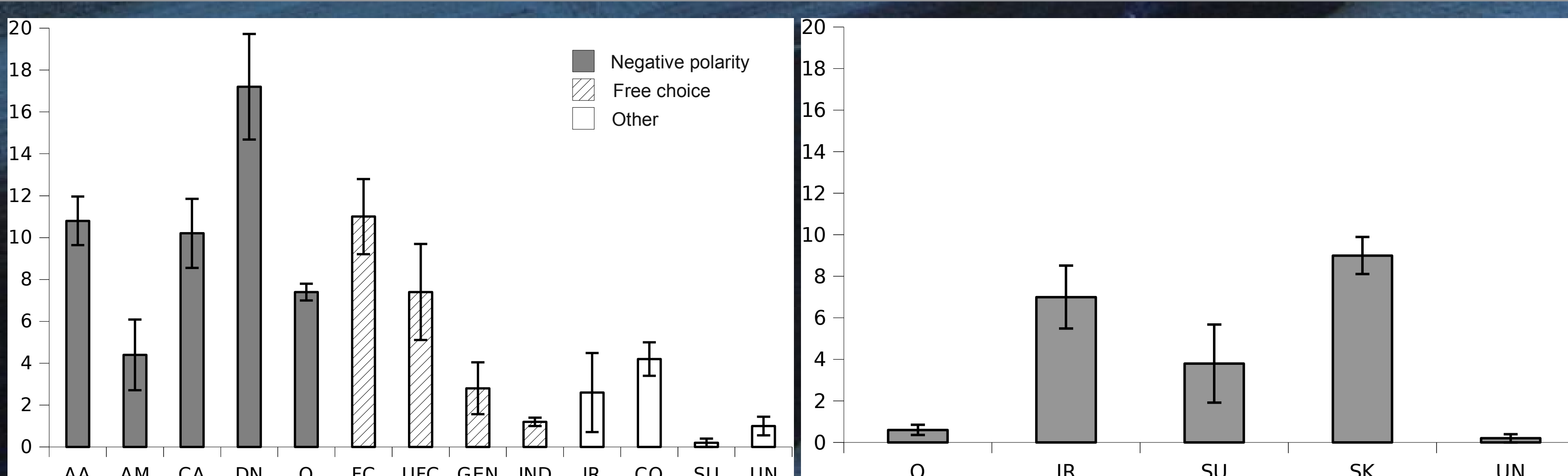Formula:

$$f(x) = \frac{\sum_{a,b \in C \wedge a \neq b}^{a,b} \mid D_x^a \cap D_x^b \mid}{\sum_{a,b \in C \wedge a \neq b}^{a,b} \mid D_x^a \mid}$$

where C is the set of coders, and D_x^a gives the set of items annotated with category x by coder a.

Corpus

The current English corpus of 100 items of indefinite uses with the multi-coder annotation is available at `http://staff.science.uva.nl/~maloni/Indefinites/corpus.html`.
We plan to extend it with annotated uses in other languages.

Results

Average distribution of functions for *any* (left) and *some* (right) over 5 annotators. The error bars show the standard error.

Inter-annotator agreement (Artstein and Poesio, 2008):

- Overall *kappa* score was 0.52 (sd = 0.07).
- When SK, SU are collapsed, *kappa* score is 0.56 (sd = 0.07)
- *Kappa* of 0.62 (sd = 0.05) results when we consider the original Haspelmath map.

References

A. Aguilar-Guevara et al. 2011. Semantics and pragmatics of indefinites: methodology for a synchronic and diachronic corpus study. In *Proc. of the DGfS Workshop "Beyond Semantics: corpus-based investigations of pragmatic and discourse phenomena"*.

R. Artstein and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

M. Haspelmath. 1997. *Indefinite pronouns*. OUP, Oxford.

Painting: Christine Bittremieux (2007), Untitled. 70 × 100 cm. Detail. Oil on canvas. www.bittremieux.nl