Simulating language change in an artificial-intelligence computational model

Sander Lestrade
Centre for Language Studies, Radboud University Nijmegen
s.lestrade@let.ru.nl

Grammatical argument marking is shown to emerge spontaneously, from general cognitive and communicative principles.

General principles

Activation: influence of frequency and recency:
(Balota and Chumbley 1985)

(1) DODGE COLT: CAR ≺ DODGE COLT

Check: make sure you're understood
(Grice 1975, Levelt 1983; Aristar 1997, de Swart 2011, Lestrade 2010)

{COMPUTER, BOOK, **PEN**, MUG, ...}
↕
{***pen***, *ballpoint, it, fountain pen*, ...}

(2) Apatani (after Abraham 1985: 38-40)
a. *mó sɨhɨnɨ pabɨne*
3SG cow killed
'He killed the cow.'
b. *sɨhɨnɨ mó **mi** alitubɨne*
cow 3SG **DAT** kicked
'The cow kicked him.'

Proximity: stand together, belong together:
(Givón 1995)

(3) *book yellow and man big*
'yellow book and big man' (not other way around)

Reduction: shortening of automatized forms:
(Jurasky et al. 2001, Heine and Kuteva 2007)

(4) *eigenlijk* → *eik* 'actually'

Erosion: actual storage of shortened forms & ***Bleaching***: loss of meaning specificity:
(Nettle 1999; Bybee 2010, Heine and Kuteva 2007)

(5) *ēwa-haft* 'century-like' → *echt* 'real'
(van der Sijs 2010)

Fusion: attachment of attenuated forms:
(Bybee 1985)

(6) *ik loop* → *'k=loop* 'I walk'

Recruitment: use of supporting expression:
(Ariel 1999)

(7) French (Heine & Kuteva, 2002, 234)
a. *La jeune [...]. Elle est danseuse.*
'The girl [...]. She is a dancer.'
b. ***Ma femme** il=est venu.*
my:F wife AGR=is come
'My wife has come.'

FirstInFirstOut: incremental production:
(van Bergen 2011)

(8) {BALL, TAKE, I}: *I* ≺ *ball* ≺ *take*

Abbreviations

123 person; AGR agreement; *D1:9* meaning dimensions, *Ext1:9* ∼ external role, *Int1:9* ∼ internal role; DAT dative; SG singular; *tgt* target event; *V1:V9* referential properties of actions, *A1:A9* ∼ actors, *U1:U9* ∼ undergoers.

Decomposing meaning

CAT:
entity=1, animate=1, ..., covering=fur, legs=4.

HIT:
action=1, participants=2, ...;
role 1: cause=1, volitional=1, ..., means=hands;
role 2: affected=1, control=0, ..., result=pain.
(Guiraud 1968, Wierzbicka 1996, Gärdenfors 2000)

Modeling meaning

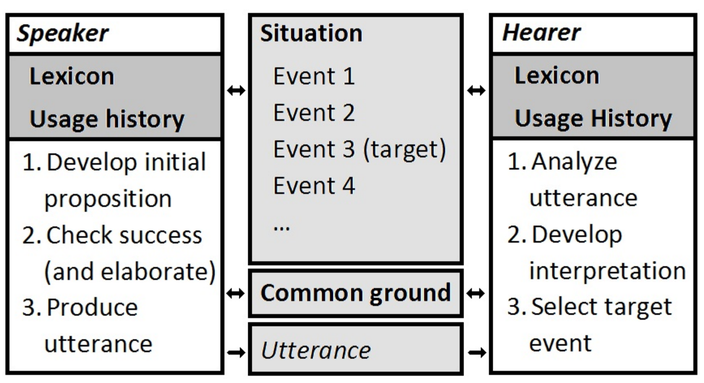
D1	D2	D3	D4	D5	D6	D7	D8	D9	form
1.00	0.00	1.00	1.00	0.00	0.75	0.25	1.00	1.00	*atadoso*
1.00	1.00	0.00	1.00	1.00	0.38	0.38	0.62	0.88	*nimator*
1.00	1.00	0.00	0.00	0.00	0.62	0.50	0.25	0.62	*umimota*
1.00	0.00	1.00	0.00	1.00	1.00	0.12	1.00	0.62	*isomera*
0.00	0.00	1.00	1.00	1.00	0.00	0.25	0.75	0.00	*enolate*
1.00	1.00	1.00	0.00	0.00	0.88	0.75	0.75	0.12	*romutil*

Table 1: *First entries of noun lexicon (abbreviated)*

D1	...	D9	Ext1	...	Int1	...	Int9	type	form
1.00	...	0.50	1.00	...	0.00	...	0.00	twoPlace	*rirunes*
1.00	...	0.50	1.00	...	0.00	...	1.00	twoPlace	*amumali*
1.00	...	0.75	0.00	...	1.00	...	0.62	twoPlace	*emimano*
0.00	...	0.75	0.00	...				onePlace	*litaril*
1.00	...	1.00	1.00	...	0.00	...	0.25	twoPlace	*adasumu*
0.00	...	0.75	1.00	...	1.00	...	0.12	twoPlace	*edesito*

Table 2: *First entries of verb lexicon (abbreviated)*

Communication procedure



(Grice 70s, Levelt 80s, Steels 90s)

Situation

V1	...	V9	A1	...	A9	123	U1	...	U9	123	tgt
1	...	0.625	0	...	0.375	3					0
1	...	0.000	1	...	0.875	3	1	...	0.375	3	0
0	...	0.625	1	...	0.500	3	1	...	0.500	3	0
0	...	0.625	1	...	0.375	3	1	...	0.250	3	0
0	**...**	**0.875**	**1**	**...**	**0.000**	**3**	**1**	**...**	**0.750**	**1**	**1**
1	...	0.375	1	...	0.250	3	0	...	0.375	3	0

Table 3: *First six events of a situation (abbreviated)*

Step 1: initial proposition

U:

D1	D2	D3	D4	D5	D6	D7	D8	D9	person	ID	form	**freq**
1	1	1	1	1		0.25	1	1	1	697	*nosonen*	**444**

arg	NM	VM	**recency**	prodEffort	weight	**match**	**coll**	typing
444	0	0	**0**	24	0.889	**0.598**	**0**	0.622

V:

D1	...	D9	Ext1	...	Ext9	Int1	...	Int9	type	ID	form
0	...	1	1	...	1	1	...	0	twoPlace	335	*naronol*

freq	**recency**	prodEffort	weight	**match**	**coll**
1	**781**	18	1	**0.986**	**0**

A:

D1	D2	D3	D4	D5	D6	D7	D8	D9	person	ID	form	**freq**
1	1	0	1	0	0.25	0.625	0.5	0	3	575	*otesere*	**0**

arg	NM	VM	**recency**	prodEffort	weight	**match**	**coll**	typing
0	0	0	**1013**	24	1	**0.986**	**0**	0.664

Step 2: check and elaborate

	Ext	Int
A	.664	.642
U	**.939**	.622

Table 4: *Argument–role qualifications ("typing")*

Best marker available to disambiguate:

(9) ID form
69 *emomene*
'naronoler'

Step 3: produce utterance

First generations: random word order, lexical ad-hoc marking, no verb agreement:

(10) *nosonen naronol otesere **emomene***
1 naronol.V otesere **naronoler**
'Otesere naronols me.'

Results

After 150 generations...

development of grammatical **word order** (SOV) and morphological **case marking**:

(11) *nelalor eremeso-**ol** osenoname*
lelalor eremeso-U osenoname.V
'Lelalor osenonames eremeso.'

development of **verb agreement** and recruitment of **local pronouns**:

(12) ***momel** arelom-ol sosolal-**ma***
2 arelomol-U sosolal-.V-2
'You sosolalm arelomol.'

(13) ***etosan** tolosam-ol somenom-**na***
1 tolosam-U somenom.V-1A
'I somenom tolosam.'

(14) erorese **etosan** asatason-**ol**
erorese 1 asatason.V-1U
'Erorese asatasons me.'

References

P.T. **Abraham** (1985), *Apatani grammar*. M. **Ariel** (1999), "The development of person agreement markers: From pronouns to higher accessibility markers". In *Usage based Models of Language*. A.R. **Aristar** (1997), "Marking and hierarchy. Types and the grammaticalization of case markers". *Studies in Language* 21 (2). D.A. **Balota** and J.I. **Chumbley** (1985), "The locus of word-frequency in the pronunciation task: Lexical access and/or production?" *Journal of memory and languages* 24. G. van **Bergen** (2011), *Who's first and what's next*, Ph.D. dissertation, Radboud University Nijmegen. J. **Bybee** (2010), *Language, Usage and Cognition*. J. **Bybee** (1985), *Morphology. A study of the relation between meaning and form*. P. **Gärdenfors** (2000), *Conceptual Spaces: The Geometry of Thought*. T. **Givón** (1995), *Functionalism and grammar*. H.P. **Grice** (1975), "Logic and conversation". In *Syntax and semantics: Speech acts* 3. P. **Guiraud** (1968), "The semic matrices of meaning". *Social Science Information* 7 (2). B. **Heine** and T. **Kuteva** (2007), *The genesis of grammar. A reconstruction*. B. **Heine** and T. **Kuteva** (2002), *World lexicon of grammaticalization*. D. **Jurafsky** et al. (2001), "Probabilistic relations between words: Evidence from reduction in lexical production". In *Frequency and the emergence of linguistic structure*. S. **Lestrade** (2010), *The space of case*, Ph.D. dissertation, Radboud University Nijmegen. W. **Levelt** (1983), "Monitoring and self-repair in speech". *Cognition* 14. D. **Nettle** (1999), *Linguistic Diversity*. **R Core Team** (2014), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing. N. van der **Sijs** (ed) (2010), *Etymologiebank*. L. **Steels** (1997), "Constructing and sharing perceptual distinctions," *Machine Learning*, ECML-97. P. de **Swart** (2011), "Sense and simplicity: Bidirectionality in differential case marking". In *Bidirectional Optimality Theory*. A. **Wierzbicka** (1996), *Semantics: Primes and Universals*.

Radboud University