Language as a Natural Encryption System

Jan Odijk[*] J.ODIJK@UU.NL

[*] *UiL-OTS, Utrecht, Netherlands*

Abstract

1. Introduction

This paper makes two parts: in the the first part I claim that language has excellent properties as an encryption system (and thus it makes communication difficult). In the second part, which is more tentative, I discuss the relation of these excellent properties of language as an encryption system to evolutionary strategies.

With *language* in this context, I mean *I-language* in the sense of (Chomsky 1986). It consist of 3 major components: syntax, phonology and semantics. Syntax includes the structure-building parts of what is traditionally called morphology. Phonology includes the parts of morphology mapping abstract feature matrixes to sequences of phonemes and relates syntax and the Sensory-Motor system. Semantics relates syntax and the Conceptual-Intentional system. *Language* includes rules and principles for discourse (i.e. combining sentences into coherent text) but I assume here that these belong in part to syntax and in part to semantics.

The conceptual system itself (possibly including a "Language of Thought") does *not* belong to language. nether do systems for knowledge of the world, for knowledge of the current situation, for knowledge of or beliefs about the intentions of other speakers, etc.

We speak of *communication* if *information* at a *sender* is intentionally *encoded* in a *signal* (thereby becoming a *message*) and transferred over a *channel* to a *receiver* who attempts to *decode* the *signal*. *Successful communication* is communication in which the information from the sender (the *message*) is decoded by the receiver as the original information at the sender.

It is a desirable property that a signal is decodable for the intended receiver but not for unintended receivers. This has also been suggested by (Baker 2003, 351) as a property of natural language:

(1) *'Suppose that the language faculty has a concealing function as well as a revealing function. Our language faculty could have the purpose of communicating complex propositional information to members of our group while concealing it from members of other groups.'*

In this paper we will elaborate on this suggestion and extend it also to properties of language within one language community.

2. Properties of language that make communication difficult

2.1 Language Diversity

(2) Large language diversity, see Table 1, making communication impossible (or at least difficult) with most other people (and here we even ignore mutually incomprehensible dialects within languages)

(3) a. no *logical* reason for this. A system in which
   i. each atomic element of the "Language of Thought" is associated with a fixed phoneme sequence (innate, species-specific)
   ii. there is a single species-specific innate set of purely syntactic formatives
   iii. each purely syntactic formative is associated with a fixed phoneme sequence
   iv. there is a single species-specific set of I-language rules

   b. yields a unique language with all the flexibility of natural language (one can make indefinitely many new concepts, the words for these new concepts follow automatically)

   c. (probably with the exception of proper names)

(4) no *biological* reason for this: many animal signaling and communication systems have fixed, innate associations between e.g. sounds and their meanings (but they are, in general, expressively very limited)

situation	**# lgs**	**population**	**nocom**
world now	6000	6000 million	99.98%
world 2030?	4000	8000 million	99.98%

Table 1: Percentage of people that one cannot communicate with, on average, using natural language (nocom), given a certain population size (population) and number of mutually incomprehensible languages (#lgs)

2.2 Critical period

(5) Critical Period for language acquisition
   a. there is a critical language acquisition period
   b. it starts very early (at or even before birth)

c. and it decays very rapidly: it ends somewhere between the age of 5 and 10 years
d. language acquisition starting in the critical period is effortless, automatic and yields native-level command of the language
e. language acquisition starting after the critical period requires a lot of effort, is difficult and almost never results in native-level command of the language

(6) Why?
a. not *logically* or *biologically* necessary (cf. puberty, between 10 and 18)
b. the results reported in Table 1 are worthless if they block communication with the 'right' people and allow communication with the 'wrong' people
c. kin selection (cf. (Fitch 2010, 425))
i. the 'right people' = the ones who carry your genes
ii. proxy for that: the young ones that are raised by you and your (extended) family
iii. but others that come in later (do not carry your genes) should not be able to acquire the same language to the same level (so they remain recognizable as 'others')
iv. in this light we can understand the properties of the critical period in (5)

2.3 Volatility and Locality

(7) a. the natural modalities for language are speech and signing
b. these are limited to the *here* and *now*: every language utterance is a local and *self-destructing message*
c. this is not *logically* necessary: after all, humans invented writing and recording of sound
d. this is not *biologically* necessary: many animals communicate (signal) with longer lasting means and/or over long distances:
i. scent marking with urine, faeces, or from special scent glands, e.g. to demarcate territory
ii. lion's roar extends over multiple square kilometers
iii. wolf's howl can stretch over 130 square kilometers (claimed on Wikipedia citing (Feldhamer et al. 2003, 496))
iv. elephants make infrasonic calls to one another at distances as far as ten kilometers

2.4 Ambiguity

Phoneme Ambiguity

(8) a speech signal must be converted into a sequence of phonemes

(9) in one set-up (Wermter et al. 1996, 114), average phoneme ambiguity was 2.3; average utterance length 31 phonemes, so 2.3^{31} possible phoneme sequences

Token ambiguity

(10) average ambiguity of occurrences of inflected word forms (= tokens) → lexical entry
   a. lexical entry identified by lemma and basic part of speech (N, V, A, etc.)
   b. e.g. occurrence of inflected word form = *graven*
      i. lemma=*graf*, pos=N, 'grave'
      ii. lemma=*graaf*, pos= N, 'count'
      iii. lemma=*graven*, pos=V, 'to dig'
   c. Average ambiguity per token = 3.99 (based on LASSY-LARGE, TWNC, file WORD-AMB.freq: 370 million tokens)

Lexical ambiguity

(11) a. *Van Dale Hedendaags Nederlands* (Sterkenburg and Pijnenburg 1984), digital version.
   b. all lexical entries: 1.5 meanings per lexical entry on average
   c. 8000 frequent lexical entries: 2.05 meanings per lexical entry on average
   d. See Table 2 (ignoring the token ambiguity described in (10)

Syntactic ambiguity

(12) a. some lexical ambiguities are resolved by syntax
   b. but many new ones are created (structural ambiguities)
   c. see Table 3

Semantic Ambiguity

(13) a. some lexical and syntactic ambiguities are resolved by semantics, e.g. by a system of semantic selection restrictions[1]
   b. but many new ones are created: quantifier scope ambiguities, specific v. non-specific interpretation, collective v. distributive interpretation of plurals, *de dicto* v. *de re* interpretations, interpretation of anaphoric and other referring expressions, interpretation of ellipted phrases, sloppy v. strict readings of ellipted phrases, intersective v. subsective readings of (adjectival) modifiers, interpretation of the relation between possessive

1. Since the boundary between semantics and world knowledge is often difficult to draw, it is not easy to assess to what extent semantics contributes to disambiguation, but I assume it plays some role.

#frq words	#infrq words	length	average ambiguity
2	3	5	14
3	2	5	19
2	5	7	32
5	2	7	81
3	7	10	147
7	3	10	514
5	10	15	2,088
8	7	15	5,329
10	5	15	9,954
5	15	20	15,854
10	10	20	75,588
15	5	20	360,383

Table 2: Average lexical ambiguity of a sentence as a function of sentence length for certain distributions between frequent and infrequent words (based on Van Dale Hedendaags Nederlands).

and head noun in an NP, interpretation of relation between non-head and head in a N-N compound, result v. event readings of nominalisations, ...

c. some of these might also be syntactic or have a syntactic reflex in some theories/implementations, but that just introduces the ambiguities already in syntax

d. (no concrete figures to support this)

(14) Summary: natural language shows massive ambiguity at each level of representation, and though next levels of representation resolve some of the ambiguities, they do not resolve all ambiguities and in fact introduce many more themselves. This is hopeless for successful communication, but very good for hiding what you actually intend.

2.5 Redundancy and Variation

(15) Redundancy is (in general) good for an encoding used for communication (more robust against noise on the channel)

(16) but natural language has *too much redundancy*!

(17) new (less-redundant) variants are created, often ad-hoc, very often in specific niches (jargon)

(18) initials, abbreviations, acronyms, pseudo-acronyms, initialisms, portmanteaus, nicknames / short names, etc.

length	average ambiguity
1	1
2	1.3
3	1.8
4	2.6
5	3.2
6	7.1
7	13.3
8	16.8
9	27.5
10	41.2
11	66.5
12	90.3
13	207.7
14	254.9
15	397.2
16	698.9
17	947.2
18	2498.3
19	2835.6
20	4652.0

Table 3: Average syntactic ambiguity as a function of sentence length, based on the Alpino corpus (7100 sentences cdbl part of the Eindhoven corpus). Table kindly provided by Gertjan van Noord (p.c.)

(19) makes communication more difficult: try to read the CLARIN Annual Report 2014 without looking in the 26 page long acronym explanation table!

2.6 Say something else than what you actually mean

(20) a. Use expression A with semantics A' to express B' (A'$\neq$ B') (where there are certain parallels between A' and B'): Metaphors

b. Use expression A with semantics A' to express $\neg$A' or the opposite of A': Irony / sarcasm

c. Use expression A with semantics A' to express B' where A' is (in some sense) weaker than B': understatements

d. Use expression A with semantics A' to suggest meaning B' in the hearer (implicatures, indirect speech acts)

3. Why do humans use natural language for communication at all?

(21) humans are excellent in natural language

a. Human beings have a natural ability for using natural language.

b. Human beings have an extremely good (unconscious) knowledge of their own native natural language.

c. They have a natural tendency to use natural language. It is often very difficult for human beings not to talk or to listen (in natural language).

d. Natural language has extremely rich expressive possibilities. It allows a wide range of topics to talk about.

(22) humans are bad in artificial languages

a. Human beings are generally very bad at working with artificial languages

b. even though artificial languages are much simpler than natural language in many respects and counted by many measures (this usually turns out to be a kind of simplicity that is irrelevant for human beings)

c. Their expressivity is often very restricted.

(23) a. In many circumstances exact transmission of the message is not crucial. Communication in natural language often fails (or succeeds only partially), and may fail without the participants being aware of it as long as exact transmission of the message is not crucial

b. human beings have extremely good systems to deal with knowledge of the world, for knowledge of the actual situation, for knowledge of or beliefs about the intentions of other speakers, and this very often compensates and makes successful communication possible *despite* natural language.

c. the message need not be transmitted correctly immediately in all cases: it can often be done by starting a dialogue to get a better transmission of the original message in a number of steps

If successful communication is crucial...

(24) People turn to artificial languages and/or artificial subsets of natural language:

a. logicians, mathematicians, etc: artificial languages (e.g. for various logics, arithmetics / calculus, set theory, etc.)

b. programmers: (artificial) programming languages,"pseudo-code" (i.e. *programming language* for the crucial aspects and *natural language* for the non-crucial aspects)

c. lawyers: highly fixed phrases and clauses from natural language with a limited number of open slots. The meanings of these clauses have been fixed by legislation and jurisdiction

(25) People turn to pictures etc. ((language independent and partially iconic, though certainly not unambiguous)
   a. pictograms in traffic / travel
   b. pictures in furniture construction manuals

(26) people turn to unique, unambiguous, non-redundant, language-independent expressions for concepts:
   a. language names (ISO 639), country names (ISO 3166), currencies (ISO 4217), date/time (ISO 8601, EDTF), names of scientific journals (ISO 4), books (ISBN), journals (ISSN), language resources (ISLRN), authors / researchers (DAI, ORCID, etc.), units (SI), ...
   b. even URLs based on natural language: they have too much redundancy and are language-dependent (tinyURL, Goo.gl URL shortener)

4. Relation to Evolutionary Strategies?

(27) How can cooperation be understood from an evolutionary perspective?
   a. *'indiscriminate sharing with others is not an evolutionary stable strategy (ESS)'* (Fitch 2010, 415)
   b. this holds both for sharing in the context of *'physical rewards (food, nest sites)'* and for sharing *'truthful information'* (Fitch 2010, 415)
   c. *'The cooperative sharing of information thus remains a central puzzle in language evolution'* (Fitch 2010, 417)

(28) Alternative accounts that, according to biologists, **can** account for cooperative sharing of information (Fitch 2010, 416):
   a. "cynical" theory by (Dawkins and Krebs 1978): *organisms signal, and attend to signals, when it is in their own best interest to do so*
   b. kin selection
   c. reciprocal altruism

(29) (Fitch 2010, 417) claims that neither of these alternatives can explain cooperation in modern humans since 'humans cooperate with a wide variety of unknown, unrelated individuals, even when there is little chance of reciprocation'.

(30) Account by Fitch: a two-stage model
   a. initially a proto-language by kin selection ('a stage of kin communication'), followed by
   b. 'the implementation of regulated information exchange among adults'

(31) Is human language a stable phenomenon?
   a. it exists only for some 100,000 years (which is very short in evolutionary terms)

b. its properties are unique in nature

c. so maybe it is just an "evolutionary error" (Fitch 2010, 417), which will be eliminated over time

(32) Let's assume it is stable. Is it an instance of a system that enables 'indiscriminate cooperative sharing of information'?

a. hardly, if the claims above are correct

b. it enables some information sharing, but this is mainly due to non-linguistic intelligence of humans and only with a select group.

c. humans have the option to share information, but they are not obliged do so (in contrast to many instances of signaling by animals). Is this relevant to evolutionary stability?

d. the properties of natural language described here can be understood in terms of *kin selection*

(33) Cooperation: perhaps there is still a problem in understanding *cooperation* in modern human beings from an evolutionary perspective

(34) but not with natural language: it is more an obstacle to cooperation 'with a wide variety of unknown, unrelated individuals' than a facilitator

5. Conclusions

(35) Encryption:

a. natural language is very good as an encryption system

b. natural language is not particularly suited for successful communication

c. communication via natural language is nevertheless (sometimes) possible, despite natural language, thanks to non-linguistic intellectual abilities of human beings

(36) Origin and existence of natural language:

a. it is not obvious that natural language is a stable phenomenon in nature

b. natural language is NOT an instance of a device that enables 'indiscriminate sharing with others'

c. its origin and existence need not be understood in terms of the (dubious) evolutionary strategy of 'cooperative sharing of information'.

d. The properties of natural language described here can be understood in terms of kin selection

References

Baker, Mark C. (2003), Linguistic differences and language design, *Trends in Cognitive Sciences* **7** (8), pp. 349–353.

Chomsky, Noam (1986), *Knowledge of language. Its Nature, Origin and Use*, Convergence, Praeger, New York.

Dawkins, R. and J. R. Krebs (1978), Animal signals: Information or manipulation?, *in* Krebs, J. R. and N. B. Davies, editors, *Behavioural Ecology*, Blackwell Scientific Publications, Oxford, pp. 282–309.

Feldhamer, G.A., B.C. Thompson, and J.A. Chapman (2003), *Wild Mammals of North America: Biology, Management, and Conservation*, JHU Press.

Fitch, W.T. (2010), *The Evolution of Language*, Cambridge University Press.

Sterkenburg, P.G.J. and W.J.J. Pijnenburg, editors (1984), *Groot Woordenboek van Hedendaags Nederlands*, Van Dale Lexicografie, Utrecht.

Wermter, Stefan, Ellen Riloff, and Gabriele Scheler, editors (1996), *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, Vol. 1040 of *Lecture Notes in Artificial Intelligence*, Springer, Berlin / Heideberg. DOI: 10.1007/3-540-60925-3.