Lectures on linguistic complexity

Östen Dahl
Course description

- The general topic of the course is the notion of complexity and its application to linguistics.
- Specific topics to be included are:
  - different ways of understanding complexity both in linguistics and in other disciplines;
  - the notions of maturity and maturation in language change;
  - possible differences among languages in complexity and external conditions that may give rise to them.
Lecture 1: What is complexity?

- This lecture will treat different ways of understanding the notion of complexity, starting with a general perspective, grounded in information theory, and zooming in on its application to linguistics.
- The delimitation of complexity against other similar notions such as difficulty will be discussed.
- The notion of non-linearity as a specific type of complexity in linguistics will be introduced.
Lecture 2: Maturation processes in linguistics

- This lecture will discuss how linguistic forms and constructions evolve over time in processes known as grammaticalization and lexicalization.
- A mature linguistic phenomenon is one that by necessity has a non-trivial prehistory.
- Mature phenomena tend to involve non-linearity as defined in the first lecture; their development will be outlined.
Lecture 3. Are all languages equally complex?

- It is often stated in textbooks and other publications in linguistics that there are no differences in complexity between languages, but this thesis has recently been contested by different scholars.

- In the lecture, the different points of view in this discussion will be outlined.

- The possible causes of differences in complexity between languages, in particular the role of language contact, will be discussed.
Two different notions of linguistic complexity

- In speaking of linguistic complexity, people tend to have two rather different notions in mind:
  - “objective” complexity
  - “agent-related” complexity
Objective complexity

- Objective complexity is the notion employed in information theory and the theory of complex systems.
- It involves the idea that complexity is an objective property of an object or a system.
- It is notoriously difficult to give a rigid definition of complexity in this sense.
Objective complexity

- Intuitively the complexity of an object is to be measured in terms of
  - the amount of information needed to re-create it or alternatively,
  - the length of the shortest possible complete description of it.
Complexity as the inverse of compressibility

- $3 \times \text{ha}$: haha
  - 6 symbols compressed to 4

- $2 \times \text{bye}$: byebye
  - 6 symbols compressed to 5

- $1 \times \text{pardon}$: pardon
  - 6 symbols – no compression
Kolmogorov complexity

- As applied to strings, this notion of complexity, which is sometimes called “Kolmogorov complexity” or “algorithmic information content”, comes out as an inverse of compressibility:
- The most complex string is one which cannot be compressed at all.

Andrej Nikolaevič Kolmogorov
1903-1987
Effective complexity

- If we implement this idea in the most straightforward way, we obtain what Murray Gell-Mann calls “effective complexity”,
- which differs from Kolmogorov complexity in that it does not measure the length of the description of an object as a whole,
- but rather the length of the description of the “set of regularities” or structured patterns that it contains.
Patterns

- The term “pattern” is very versatile...
- It can refer to a regularity that we observe in nature, e.g. a comet that comes back at regular intervals:
  - **Halley’s comet:**

  1531  1607  1682  1759  1835  1910  1986  2061

- ...or other natural phenomena:
Patterns guiding behaviour

- ... but a pattern can also be a rule or a principle that we apply in our behaviour or when creating an artifact:
Patterns in language

- ...but it may also be a regularity or a construction in language:

Долой + NP_{acc}
Patterns generalized

• A pattern is something that enhances the compressibility of an object...
• ...that is, makes it possible to obtain a shorter description of it...
• ...and thus reduce Kolmogorov complexity
• It is, basically, something non-random
Back to complexity

- A random string of characters, such as w509mf0wr6435217ro0l71734 will have maximal Kolmogorov complexity (the string is its own shortest description), but no effective complexity since it contains no structured patterns.

- This corresponds better to an intuitive understanding of the notion of complexity.

- We also come close to a notion which may feel more familiar to linguists: the set of patterns that an object contains can be said to equal its structure, so the complexity of an object is really a measure of the complexity of its structure.
System complexity vs. structural complexity

- In linguistics, such a complexity measure could apply to different things. Most importantly, it could apply on the one hand to
  - a language seen as a system (system complexity)
  - to the structure of utterances and expressions (structural complexity)
Comparison of system complexity

- The (written) English definite article is less complex than the English indefinite article and than the French definite article.

**definite article - the**

**indefinite article - an before vowels; a elsewhere**

**definite article - l’ before vowels; les before plural nouns; else la before feminine nouns; le elsewhere**
System complexity

- System complexity could be seen as a measure of the content that language learners have to master in order to be proficient in a language.
- It does not as such tell us anything about the difficulty they have in learning, producing and understanding the language --
- -- that would take us to the other notion of complexity, viz. agent-related complexity.
Structural complexity

- These sentences are generated by the same grammar but differ in structural complexity (B>C>A)
Agent-related complexity?

- Although agent-related complexity is perhaps the most popular way of understanding complexity in linguistics, I would in fact prefer to reserve the term “complexity” for objective complexity and use other terms such as “cost”, “difficulty”, and “demandingness” to denote ‘complexity for a user’. 
Cost

- **Cost** is essentially the amount of resources – in terms of energy, money or anything else – that an agent spends in order to achieve some goal.
- What we can call cost-benefit considerations are certainly of central importance in explaining many aspects of communicative behaviour.
Difficulty

- Difficulty is a notion that primarily applies to tasks and is always relative to an agent: it is easy or difficult for someone.
- Difficulty can of course be understood and measured in several different ways.
- One measure of the difficulty of a task is in terms of “risk of failure”: if a large proportion of all agents fail or one agent fails more often than he or she succeeds, the task is difficult for that agent or group of agents.
The base line – the fridge magnet language

- In the fridge magnet language, there is no indication of how the words hang together
- Each word is an independent unit
The fridge magnet language IRL: Riau Indonesian according to David Gil

- Riau Indonesian: a variety of Indonesian spoken in the province of Riau
- *Ayam makan* (lit. 'chicken eat') can mean, in context, anything from 'the chicken is eating', to 'I ate some chicken', 'the chicken that is eating' and 'when we were eating chicken'
Introducing order

fish three dozen
three fish dozen
three dozen fish
fish dozen three
dozent fish three
dozent three fish
Introducing grammatical marking

two pounds of sausage

trista gramm(ov) kolbasy
Introducing non-linear marking

`en liten man`

Swedish

'one little man'

`två små män`

Swedish

'two little men'

suppletion

umlaut
Non-linearity

- Non-linearity in a wide sense: a lack of a one-to-one correspondence between conceptual entities and segmental elements of the expression "little".
Verbosity

- Verbosity is a term I use to indicate that a construction in a language contains more elements/is longer than is necessary for communication

papa bilong mipela

vår far

‘our father’
Maturational processes

Östen Dahl
Grammaticalization

- Narrowly conceived, grammaticalization is the process by which lexical items develop into grammatical markers.
- In a broader perspective, grammaticalization covers the processes by which grammatical systems arise and develop over time.
Grammaticalization

- The term *grammaticalisation* was introduced in 1912 by Antoine Meillet:
  
  « l'attribution du caractère grammatical à un mot jadis autonome »

the attribution of a grammatical character to an erstwhile autonomous word
Maybe not Meillet after all?

- ...but it is unclear if Meillet thought of it as a term:

...and the idea goes as far back as to the French 18th century philosopher Condillac

...and was known in the 19th century under the name “Agglutinationstheorie”
A classical example of grammaticalization

Latin

facere habeo

do.INF have.PRS.1SG
'I have to do = I must do'

faciam

do.FUT.1SG
'I will do'

French

(je) ferai

(l) do.FUT.1SG
'I will do'

Italian

farò

do.FUT.1SG
'I will do'

Spanish

haré

do.FUT.1SG
'I will do'

niens. C'est par ce moyen que les langues romanes se sont donné un futur quand le futur du latin ancien est devenu trop faible, trop inexpressif: facere habeo, qui est l'original de (je) ferai, signifie « j'ai à faire », c'est-à-dire « je dois faire ». L'infinitif et
What happened?

● In this case, it is easy to see that the following things have happened:

  – Change of meaning:
    • obligation $\rightarrow$ future

  – Phonological reduction:
    • -re habeo $\rightarrow$ [re]

  – "Univerbation"
    • word $\rightarrow$ affix
Other things that happen in grammaticalization

- **Obligatorification**
  - Items become obligatory in certain contexts
  - The absence of an item becomes meaningful

- **Featurization**
  - Items are understood as expressing inflectional features which form paradigms

- **Non-linear marking**
  - Grammatical marking is done by other means than by adding separable morphemes
Non-linear marking may arise through “cheshirization”

- A phonetic change by which a sound disappears may leave a trace behind
- like the Cheshire cat who left his grin behind when disappearing
- Example:
Constructions, not words

- In the spirit of Meillet, grammaticalization is often conceived of as involving words:
  - In *linguistics*, **grammaticalization** (also known as **grammatization** or **grammaticization**) is a process by which words representing objects and actions (i.e., **nouns** and **verbs**) transform through **sound change** and **language migration** to become grammatical objects (**affixes** and **prepositions**, etc). (Wikipedia)

- ...but it should be clear that we are dealing with constructions rather than with words

- ...and it is disputable if the source is always lexical
Some more or less grammaticalized progressive constructions

English: he is a-hunting < he is at hunting
German: er ist am Schreiben
French: il est en train d’écrire une lettre
Swedish: han håller på att skriva ett brev
Spanish: está escribiendo una carta

- In fact, a word cannot be grammaticalized by itself but only as part of a grammatical construction or pattern
Semantic changes connected with grammaticalization

- Metaphor
  - Estonian *peal* 'on X’s head’ > ’on (top of)’

- Metonymy
  - Often: conventionalisation of implicatures
  - English *will* ’want to’ > intentional future (desire+prediction)

- Semantic bleaching
  - Semantic generalization, loss of semantic components
  - English *will* ’intentional future (desire+prediction)’ > non-intentional future
Other changes in grammaticalization

- **Phonology/phonetics:**
  - reduction
  - loss of prosodic autonomy or salience

- **Recategorization**

- **Loss of word status**

- **Fusion of elements**

- **Obligatorification**
  - Items become obligatory in certain contexts
  - The absence of an item becomes meaningful

Examples:
- at hunting > a-hunting
- demonstratives > articles
- adverbs > adpositions
- is going to > gonna
What does “obligatory” mean?

- In morphology, “obligatory” can mean that a certain morpheme is always present (in Latin, every noun has a case ending)
- However, the existence of zero markings risks making obligatoriness empty
- With a different notion of obligatoriness, a marking is obligatory if it has to be present given a certain condition
The notion of maturation

• Informal definition:
  – a mature linguistic phenomenon is one which necessarily has a non-trivial prehistory

• Examples:
  – ablaut (apophony), as in Germanic strong verbs: drink – drank – drunk
  – grammatical gender, as in French
    un beau jour – une belle journée

• Maturation tends to increase complexity in the form of non-linearity

• It also decreases phonetic redundancy
Some more mature phenomena

- phonology: nasal vowels, tones
- morphology: fused expression, prosodic marking
- syntax: agreement, non-hierarchical order
The connection between grammaticalization and maturation

- Grammaticalization processes tend to form shorter or longer chains
- The results of the grammaticalization chains usually cannot come about without those chains
- In other words, the degree of maturation of these results depends on the length of the preceding grammaticalization chains
- Highly grammaticalized items thus are usually highly mature.
Stages in the development of morphological structure

isolating structures
("periphrastic constructions")

Old Norse
*kalla sik*

agglutination
(affixing)

Swedish
*kalla-s*

ProtoGmc.*mann-iz*

English
*men*

Old Norse.
gifa / gaf

Swedish
*[je:]/*[ga:v]*

fusion

suppletion

maturity increases
A fuller story: Finnic comitative

- It is not so common for a complete grammaticalization chain to be attested
- This example from Finnic illustrates the whole path from a full lexical item to an affix:

Proto-Germanic *hansō 'group'

Finnish kansa 'people'

Finnish kanssa 'with'<kansassa 'in the group of'

Estonian -ga 'with'

German Hanse
Components of pattern evolution

- **pattern spread**: a pattern comes to be used in situations where it was not used before

- **pattern regulation**: the choice between two patterns sharing the same niche is constrained in one way or the other

- **pattern adaptation**: a pattern undergoes changes which make it better suited to its new uses, changes such as reduction and condensation (tightening)

*all in you all* starts being used even when totality is not emphasized

*you all* becomes obligatory when more than one person is referred to

*you all* is reduced to *y’all*
Featurization and non-linear marking

jump-ed

VERB STEM + PAST

ran

Cf. French [e] = first person singular present indicative of ‘have’
< Latin habeo

PAST is a global property of the word form rather than a segmentable morpheme
Preservation of complexity in language change

- Cf. Italian (conservative) vs. French (innovative):
  
  \[\begin{align*}
  \text{bello} & : \text{bella} \\
  \text{caldo} & : \text{calda} \\
  \text{vecchio} & : \text{vecchia}
  \end{align*}\]

- The French forms are shorter (less phonetically redundant) but preserve the masculine-feminine distinction.

- More information is packed into lesser space as the structure becomes less linear.
Stable complex systems

- Highly mature systems can be very stable, once they have arisen.
- Germanic strong verbs have existed in the same form for at least 2,000 years and go back on Indo-European ablaut which is at least 6,000 years old.
- Similarly, the vowel alternations of the Semitic languages may go back to Proto-Afro-Asiatic, which was spoken perhaps 10,000 years ago.
Are all languages equally complex?

Östen Dahl
Complexity is natural

- Complexity in the form of mature features is by definition a result of historical development.
- Without this development, all languages would look more or less like creole languages, claimed by John McWhorter to be maximally simple among languages.
Complexity is natural

- If we assume that all languages at some earlier point lacked mature features, and complification and simplification were equally probable, we would predict a curve as in the left-hand diagram. However, what we get is the curve to the right (Nichols, McWhorter).

- It thus appears that complification at least up to a certain point is more frequent than simplification
The distribution of (morphological) complexity

Complexity distribution according to Nichols 1992
Stable complex systems

- Highly mature systems can be very stable, once they have arisen.
- Germanic strong verbs have existed in the same form for at least 2,000 years and go back on Indo-European ablaut which is at least 6,000 years old.
- Similarly, the vowel alternations of the Semitic languages may go back to Proto-Afro-Asiatic, which was spoken perhaps 10,000 years ago.
The thesis of equal complexity

● It is often said that all languages are equally complex
● Complexity in one area is supposedly always compensated by lack of complexity in another and vice versa
● However...
No clear empirical evidence for the equality thesis

- Nobody has presented any statistics that shows that the equality thesis holds
- If the equality thesis is taken literally, it implies that a loss of complexity in a language will always be followed by a gain of complexity somewhere else in that language...
- ...which has never been shown to be the case
No clear way of comparing the complexity of different components of language

- Nobody has proposed a working way of weighing for instance syntactic complexity against morphological complexity
- In fact, it is far from clear how to measure syntactic complexity in the first place
Word order vs. inflection

- It is often said or implied that lack of inflection is compensated by rigid word order rules...
- ...but this is a tendency rather than a strict correlation
- Germanic languages with case systems (German, Elfdalian) do not show clear differences in the rigidity of word order relative to other Germanic languages (English, Swedish)
- Again, it is not clear how to compare word order and inflections with respect to complexity
- If anything, strict word order does not seem to imply any high degree of system complexity
Languages may not be equally complex

- It has been suggested that high-contact languages tend to be less complex than low-contact languages, in particular with respect to morphological complexity.
- This may have to do with the difficulties that second-language learners meet in trying to acquire the kind of complexity that is characteristic of low-contact languages.
- This is “L2-difficulty”, not system complexity as such, but the claim is that there is a correlation between L2-difficulty and particular kinds of system complexity.
Relationship between complexity and degree of contact

- Certain types of complexity – notably mature phenomena – seem to be inversely correlated with the degree to which the language in question is in contact with other languages.

- Small and geographically isolated languages tend to have more complex and “mature” morphology than large ones.

- This is at least partly explained if we assume that mature phenomena tend to be among those aspects of language that are particularly difficult to acquire for adults.
Language as a secondary genome
Vertical and horizontal transmission

- **First language acquisition:**
  - Vertical
  - Early
  - Near perfect

- **Second language acquisition:**
  - Horizontal
  - Late
  - Error prone
The filtered version becomes an L1

The filtered version may then be transmitted to the next generation as L1
The effect of second-language learning

- If you learn a language as an adult, you are likely to learn certain things imperfectly – in particular mature and complex phenomena.
- Second-language learners may transfer their imperfectly acquired second language to their children.
- This leads to a decrease in complexity.
- Creole languages can be seen as extreme examples.
What is a pidgin?

- According to Suzanne Romaine, the traditional view is that a pidgin is a contact variety
  - “restricted in form and function”
  - “native to no one”
  - “formed by members of at least two (and usually more) groups of different linguistic background”
What is a creole?

- According to the same view, a creole is “a nativized pidgin, expanded in form and function to meet the communicative needs of a community of native speakers”
- Thus, on this view, a creole by definition originates in a pidgin
Maps

The Atlas of Pidgin and Creole Language Structures (APiCS)

Link to big map
Pidgin and creoles – orphan languages

- Pidgins and creoles are usually seen as languages without a parent.
- In other words, they are thought of as being created anew rather than being derived genealogically from any other language.
- The language that they get their lexicon from is called a **lexifier** language.
- A lexifier can also be thought of as a superstrate.
- ... and the original native languages of the speakers as substrates.
Creoles

- Most of the creoles discussed in the literature are “Atlantic creoles” which arose in connection with the Atlantic slave trade.
McWhorter’s “creole prototype”

- According to John McWhorter, creoles exhibit some features that may be used to distinguish them from other languages without referring to the socio-historical dimension:
  - to use grammatical inflection via affixing,
  - to develop productive, nontransparent derivational affixes, or
  - to use tone to either mark lexical differences or as grammatical markers
Maturation and borrowability

- Highly mature phenomena are seldom borrowed from another language to another
- We do not expect other languages to borrow the strong verbs of Germanic
- Less mature grammatical phenomena are more readily borrowed, however
- The grammaticalization scale can also be seen as a scale of borrowability
Esoteric vs. exoteric

- Languages spoken in the *esoteric niche*: smaller population, smaller area, fewer linguistic neighbors
- Languages spoken in the *exoteric niche*: larger population, larger area, more linguistic neighbors

“Esoteric” languages have more of:

- case markings
- ergative alignment
- grammatical categories marked on the verb
- person marking on adpositions
- noun/verb agreement
- inflectional evidentiality
- affixal negation
- morphological future tense
- remoteness distinctions in the past tense
- alienability/inalienability distinctions
- optative mood marking
- distance distinctions in demonstratives
- morphological marking of pronominal subjects
- separate associative plurals

"Exoteric" languages have more of:

- obligatory plural marking
- dedicated question particles
- syncretism in case and agreement
- ... but they also have bigger vowel inventories and more often tonal distinctions
- ... and they also tend to have a basic word order where the verb is not in final position (VSO, SVO)
The esoteric index

Highly esoteric areas
Northern and western Europe
West Africa
South East Asia

Highly esoteric areas
Northeast Siberia
North and South America
Australia