On the duality of patterning

Jordi Fortuny

Centre de Lingüística Teòrica
Universitat Autònoma

Foundations and structure of language
2011
Table of contents

1 Description
2 Hockett’s objectives
3 Chomsky’s Strongest Minimalist Thesis
4 Hockett’s approach
   - Hockett’s list
   - Hockett’s general observations
5 The duality of patterning
   - Information theory
   - Data compression
     - The concepts of ‘coding algorithm’, ‘code’ and ‘codeword’
     - Code length
   - Data transmission
     - Shannon’s Fundamental Theorem for a Discrete Channel with noise
   - Double purpose of the duality of patterning
   - Structural complexity and the size of the morphological storage
   - Conclusion
We will begin by briefly presenting, as a preliminary step, Hockett’s 1960 seminal approach to the thirteen design features he listed.

Subsequently, we will focus on how certain basic considerations mainly related to information theory and structural complexity favor the emergence of Hockett’s thirteenth feature, the duality of patterning (Fortuny 2010).
Hockett’s objectives

- What are the essential properties of language? How does language differ from non-human animal communication systems?
- Hockett’s goal is to compare language to animal communication systems in order to find out the process by which language appeared in the evolution.
- Hockett’s objective: to set up a short list of basic “design features” that were all present in all languages although no animal communication system displayed them all.
Hockett’s objectives

- Hockett’s comparative method: “Although the comparative method of linguistics, as has been known, throws no light on the origin of language, the investigation may be guided by a comparative method modeled on that of the zoologist” (p.5).
- After arriving at a particular list of thirteen features, he suggested that such a list could be used “to work out the sequence by which the ancestral system became language as the hominids -the men-apes and ancient men- became man” (p.6).
Chomsky’s Strongest Minimalist Thesis

**The Strongest Minimalist Thesis** (Chomsky 2000)
Language is an optimal solution to legibility conditions

Given an alleged linguistic property $P$, we consider whether or not $P$ is a genuine linguistic property and if it is so, whether it can be derived from independent principles.
Chomsky’s Strongest Minimalist Thesis

The Strongest Minimalist Thesis (Chomsky 2000)
Language is an optimal solution to legibility conditions

Given an alleged linguistic property $P$, we consider whether or not $P$ is a genuine linguistic property and if it is so, whether it can be derived from independent principles.
The Strongest Minimalist Thesis (Chomsky 2000)

Language is an optimal solution to legibility conditions

Given an alleged linguistic property \( P \), we consider whether or not \( P \) is a genuine linguistic property and if it is so, whether it can be derived from independent principles.

- We may investigate whether or not certain design features can be derived from independent principles or rather constitute linguistic primitives.
Hockett’s list

1. Vocal-auditory channel. Linguistic expressions are mainly externalized through the larynx and the supralaryngeal vocal tract and perceived through the ear.

2. Rapid fading. Linguistic signs are transitory.

3. Broadcast transmission and directional reception. “A linguistic sign can be perceived by an auditory system within earshot and the source can normally be localized by binaural direction finding” (Hockett 1960, 6)
Hockett’s list

1. Vocal-auditory channel. Linguistic expressions are mainly externalized through the larynx and the supralaryngeal vocal tract and perceived through the ear.

2. Rapid fading. Linguistic signs are transitory.

3. Broadcast transmission and directional reception. “A linguistic sign can be perceived by an auditory system within earshot and the source can normally be localized by binaural direction finding” (Hockett 1960, 6)

4. Total feedback. “The speaker of a language hears everything of linguistic relevance in what he himself says” (p.6)

5. Interchangeability. “A speaker of a language can reproduce any linguistic sign he can understand” (p.6)
Hockett’s list

6 Specialization. “The bodily effort and spreading sound waves of speech serve no function except as signals” (p.6)

7 Semanticity. “A message triggers the particular result it does because there are relatively fixed associations between elements in messages and recurrent features or situations of the world around us” (p.6)

8 Arbitrariness. “The ties between meaningful message-elements are arbitrary (...); (this) has the great advantage that there is no limit to what can be communicated about”.

9 Displacement. Linguistic expressions can refer to objects or events that are not present in the speech act.

10 Productivity. “The capacity to say things that have never been said or heard before and yet to be understood by other speakers of the language”.
Hockett’s list

Traditional transmission. “Human genes carry the capacity to acquire a language, and probably also a strong drive toward such acquisition, but the detailed conventions of any one language are transmitted extragenetically by learning and teaching”.
Hockett’s list

11 Traditional transmission. “Human genes carry the capacity to acquire a language, and probably also a strong drive toward such acquisition, but the detailed conventions of any one language are transmitted extragenetically by learning and teaching”.

12 Discreteness. The minimal units that compound the expressions of a language do not constitute a continuum.

13 Duality of patterning. Linguistic expressions can be decomposed into minimal meaningful discrete units commonly called morphemes that in turn can be decomposed into minimal distinctive though meaningless discrete units called phonemes (cf. Roselló (2006)).
Hockett’s general observations

- “It is probably safe to assume that nine of the 13 features were already present in the vocal-auditory communication of the protohominoids - just the nine that are securely attested for the gibbons and humans of today” (Hockett 1960: 8).

- “The problem of the origin of human speech, then, is that of trying to determine how such a system could have developed the four additional properties of displacement, productivity and full-blown traditional transmission” (Hockett 1960: 8) [sic].

- Important qualm: does arriving at a particular list help us to understand how language appeared in the evolution?
Information theory

Information theory is an attempt to construct a mathematical model for each of the components of a communication system (Ash 1965).

- The source of messages is the person or machine that produces the information to be communicated.
- The encoder associates with each message an object that is suitable for transmission over the channel.
- the channel is the medium over which the codeword (the coded message) is transmitted.
Information theory

- The *decoder* operates on the output of the channel and attempts to extract the original message.
- *Noise* is a general term for anything that tends to produce errors in transmission thereby reducing the reliability of the decoder in reconstructing the original messages.
- The *object* (*sign*, or *codeword*) is usually a string of symbols in digital computer applications or a continuous wave in radio communication.
Data compression

Data compression or noiseless coding theory is one of the components of information theory (Ash 1965, Cover & Thomas 2006).

Let $\Omega = \{\omega_1, ..., \omega_n\}$ be a set of messages to be coded by a coding algorithm $C$.

For instance, $\Omega = \{\spadesuit, \heartsuit, \clubsuit, \diamondsuit\}$.

The messages to be coded by $C$ are generated by a discrete random variable $X$ ranging over $\Omega$ with associate probability function $p(w_i)$, for $\forall (\omega_i) (\omega_i \in \Omega)$.
Data compression

For $\Omega = \{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}$,

$$X_\Omega(\omega) = \begin{cases} 
2/4 & \text{if } \omega = \clubsuit, \\
1/4 & \text{if } \omega = \diamondsuit, \\
1/8 & \text{if } \omega = \heartsuit, \\
1/8 & \text{if } \omega = \spadesuit. 
\end{cases}$$

$C$ needs an alphabet, i.e., a finite set of basic symbols $A = \{a_1, \ldots, a_k\}$. Commonly, $A = \{0, 1\}$.

$A^*$ will refer to the set of finite length strings of symbols of $A$; for $A = \{a_i, \ldots, a_k\}$, the four strings $a_1, a_1a_2, a_2a_1, a_k\ldots a_1$ would be elements of $A^*$.
The concepts of 'coding algorithm', 'code' and 'codeword'

Intuitively, a coding algorithm $C$ proceeds by assigning to each element of $\Omega$ an element of $A^*$. 

Coding algorithm

Let $X$ be a random variable ranging over $\Omega$ and $A^*$ a set of finite strings of $A$; a coding algorithm $C$ for $X$ is a mapping from $\Omega$ to $A^*$.

A codeword is an element of $A^*$ that is assigned by a coding algorithm to an element of $\Omega$; $C(\omega_3)$ is the codeword associated to $\omega_3$, and the set of codewords (a subset of $A^*$) is called a code.
Code length

- Objective of data compression: to code as much data as possible using the smallest number of digits.
Objective of data compression: to code as much data as possible using the smallest number of digits.

This is achieved by inversely correlating the length of the codewords with the frequency of the corresponding outcomes belonging to $\Omega$ according to a given probability function, in such a way that the most frequent outcomes are assigned short codewords and the less frequent outcomes are assigned longer codewords.
In sum, data compression attempts to minimize the expected length $L(C)$ of a coding algorithm $C$.

$L(C)$ is naturally defined as

$$L(C) = \sum_{\omega_i \in \Omega} p(\omega_i) l(\omega_i),$$

where $l(\omega_i)$ refers to the length of any codeword $C(\omega_i)$. 

where $l(\omega_i)$ refers to the length of any codeword $C(\omega_i)$. 

where $l(\omega_i)$ refers to the length of any codeword $C(\omega_i)$. 

where $l(\omega_i)$ refers to the length of any codeword $C(\omega_i)$.
Code length

Illustration of three coding algorithms \(C_1, C_2, C_3\), each characterized as a mapping from \(\Omega\) to \(A^*\). \(p\) is not uniform provided that it does not assign the same value to each element of \(\Omega\). \(C_1\), but not \(C_2\) or \(C_3\), minimizes its expected length by inversely correlating \(l(\omega_i)\) to \(p(\omega_i)\) for \((\forall \omega_i)(\omega_i \in \Omega)\). 
\(C_3\) directly correlates \(l(\omega_i)\) and \(p(\omega_i)\) and \(C_2\) does not correlate the two magnitudes whatsoever.

\[\Omega = \{\spadesuit, \diamondsuit, \heartsuit, \clubsuit\}\]
\[A = \{0, 1\}\]
\[A^* = \{x : x \text{ is a finite string of symbols of } A\}\]

<table>
<thead>
<tr>
<th>(X) with probability distribution (p)</th>
<th>(C_1)</th>
<th>(C_2)</th>
<th>(C_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(p(\spadesuit) = 2/4)</td>
<td>(C_1: \spadesuit \rightarrow 1)</td>
<td>(C_2: \spadesuit \rightarrow 1)</td>
<td>(C_3: \spadesuit \rightarrow 100)</td>
</tr>
<tr>
<td>(p(\diamondsuit) = 1/4)</td>
<td>(C_1: \diamondsuit \rightarrow 00)</td>
<td>(C_2: \diamondsuit \rightarrow 101)</td>
<td>(C_3: \diamondsuit \rightarrow 11)</td>
</tr>
<tr>
<td>(p(\heartsuit) = 1/8)</td>
<td>(C_1: \heartsuit \rightarrow 10)</td>
<td>(C_2: \heartsuit \rightarrow 0)</td>
<td>(C_3: \heartsuit \rightarrow 1)</td>
</tr>
<tr>
<td>(p(\clubsuit) = 1/8)</td>
<td>(C_1: \clubsuit \rightarrow 100)</td>
<td>(C_2: \clubsuit \rightarrow 00111)</td>
<td>(C_3: \clubsuit \rightarrow 0)</td>
</tr>
</tbody>
</table>
Code length

\[ L(C_1) = \left(\frac{2}{4}\right)1 + \left(\frac{1}{4}\right)2 + \left(\frac{1}{8}\right)3 + \left(\frac{1}{8}\right)3 = \frac{2}{4} + \frac{2}{4} + \frac{3}{8} + \frac{3}{8} = 1.75 \text{ bits} \]

\[ L(C_2) = \left(\frac{2}{4}\right)1 + \left(\frac{1}{4}\right)3 + \left(\frac{1}{8}\right)1 + \left(\frac{1}{8}\right)5 = \frac{2}{4} + \frac{3}{4} + \frac{1}{8} + \frac{5}{8} = 2 \text{ bits} \]

\[ L(C_3) = \left(\frac{2}{4}\right)3 + \left(\frac{1}{4}\right)2 + \left(\frac{1}{8}\right)1 + \left(\frac{1}{8}\right)1 = \frac{6}{4} + \frac{2}{4} + \frac{1}{8} + \frac{1}{8} = 2.25 \text{ bits} \]

The notion of coding algorithm is based on a general type of duality. If there were no distinction between \(A\) and \(A^*\), all codewords would have the same length regardless the frequency of their respective outcomes; therefore, the average codeword length could not be minimized when the frequency of the elements of \(\Omega\) was not uniform.
The duality of patterning seems a particular case of the general duality of patterning assumed in coding theory: the set of phonemes of a natural language corresponds to $A$ and the set of morphemes approximately corresponds to the notion of code.

By virtue of the duality of patterning it is possible to assign different lengths to words depending on how often they are used, thereby minimizing the average length of words. Indeed, it is a well-grounded observation that “the length of words tends to bear an inverse relationship to its relative frequency”, although such a relationship may not be proportionate (Zipf 1936: 38).
“There is excellent reason to believe that duality of patterning was the last property to be developed, because one can find little if any reason why a communicative system should have this property unless it is highly complicated. If a vocal-auditory system comes to have a larger and larger number of distinct meaningful elements, these elements inevitably come to be more and more similar to one another in sound. There is a practical limit, for any species or any machine, to the number of distinct stimuli that can be discriminated, especially when the discriminations typically have to be made in noisy conditions” (p.12).
Shannon’s concern: increase the reliability of a communication system by designing codes that did not exceed the channel capacity.
Shannon’s Fundamental Theorem for a Discrete Channel with noise

- Shannon’s concern: increase the reliability of a communication system by designing codes that did not exceed the channel capacity.
- Channel capacity: maximum rate at which data can be transmitted through a channel with arbitrary small equivocation.
Shannon’s Fundamental Theorem for a Discrete Channel with noise

- Shannon’s concern: increase the reliability of a communication system by designing codes that did not exceed the channel capacity.
- Channel capacity: maximum rate at which data can be transmitted through a channel with arbitrary small equivocation.
- Shannon’s Fundamental Theorem provides a sequence of codes of exponentially decreasing error probability with linearly increasing codeword lengths and exponentially increasing number of codewords.
Shannon’s Fundamental Theorem for a Discrete Channel with noise

- Shannon’s concern: increase the reliability of a communication system by designing codes that did not exceed the channel capacity.
- Channel capacity: maximum rate at which data can be transmitted through a channel with arbitrary small equivocation.
- Shannon’s Fundamental Theorem provides a sequence of codes of exponentially decreasing error probability with linearly increasing codeword lengths and exponentially increasing number of codewords.
- In Shannon’s Theorem, the existence of codewords is a prerequisite for the existence of codes that can overcome the transmission error as well as for increasing the number of objects that can be encoded (Plotkin and Nowak 2000).
Double purpose of the duality of patterning

- The emergence of codewords is motivate by elementary considerations related both to data compression and data transmission. It allows to minimize the average codeword length and to increase the length of words in order to avoid transmission errors.
Double purpose of the duality of patterning

- The emergence of codewords is motivate by elementary considerations related both to data compression and data transmission. It allows to minimize the average codeword length and to increase the length of words in order to avoid transmission errors.

On Hockett’s reasoning:

1. Duality is a prerequisite for data compression, which is independent of how large the number of elementary symbols is or how close the number of stimuli is to the maximum number of stimuli that can be correctly discriminated. Thus, it is not necessary to resort to the channel capacity (“practical limit”) to understand why duality emerged.

2. Hockett’s reasoning does not consider the conceivable case where a language neither displays the duality of patterning nor increases the number of distinct meaningful elements up to exceeding “the practical limit (...) to the number of distinct stimuli that can be discriminated”.

Jordi Fortuny (Centre de Lingüística Teòrica)
Structural complexity and the size of the morphological storage

- Consider the hypothetical case of a language that displayed a system of around 40 sound units each of which corresponded to a particular semantic atom or monad.
Consider the hypothetical case of a language that displayed a system of around 40 sound units each of which corresponded to a particular semantic atom or monad.

In such a language, a few sound units corresponding to monads would be combined to express non-atomic expressions that conveyed compounded meanings.
Structural complexity and the size of the morphological storage

- Consider the hypothetical case of a language that displayed a system of around 40 sound units each of which corresponded to a particular semantic atom or monad.

- In such a language, a few sound units corresponding to monads would be combined to express non-atomic expressions that conveyed compounded meanings.

- This hypothetical language, which would not display the duality of patterning, would not differ from natural languages as to the degree of articulatory resolution and perceptual discriminability, although the structural complexity of the expressions of the former would be remarkably higher than those of the latter.
More precisely, we may say that a language that did not display the duality of patterning and had the expressive power of natural languages would need to resort to the unbounded syntactic procedure (syntax) responsible for combining minimal meaningful units to generate complex expressions much more extensively than natural languages.

Thus, the duality of patterning constrains resorting to the compositional components of language, and thus reduces the complexity of grammatical compositions. Note, though, that resorting to the compositional component is also necessary to reduce the size of the morphological storage.
More precisely, we may say that a language that did not display the duality of patterning and had the expressive power of natural languages would need to resort to the unbounded syntactic procedure (syntax) responsible for combining minimal meaningful units to generate complex expressions much more extensively than natural languages.

Thus, the duality of patterning constrains resorting to the compositional components of language, and thus reduces the complexity of grammatical compositions. Note, though, that resorting to the compositional component is also necessary to reduce the size of the morphological storage.
Conclusion

Advantages for the appearance of the duality of patterning:

1. it minimizes the number of distinct meaningful units, as noted by Hock-ett
2. it allows to overcome transmission errors
3. it restricts the resort to the unbounded syntactic procedure
4. it minimizes the average length of words
5. it allows to avoid that articulatory restrictions relative to how sounds are combined to form syllables constrained the set of possible compound meanings that can be expressed.


