Cellese - The Language of the Cell

(Or: How you pronounce your proteins)

Harry van der Hulst
Skidmore College

Lecture in B. Possidente's Genetics Class
November, 30th, 1999

0. Introduction

There have been numerous references to similarities between what is often called 'the genetic code' and (human) language. The genetic code *in the strict sense* (involving the 'translation' from mRNA into amino acids) is, however, only one step in the "language of the cell". The whole *derivation* leading from DNA strands (in the chromosomes) to proteins and protein complexes has been described in two cycles of 'letters' forming words. The DNA (or rather RNA) letters form words, called *codons*. The codons translate into 20 *amino acids*, which in turn can be taken as an alphabet of letters making up words, this time called *proteins*. The two cycles are sometimes called *different languages* because they are formed from different chemical building blocks.

(1)

DNA/RNA (letters)		
|		
Codons (words)	--translation→ (genetic code)	amino acids (letters)
		|
		Protein (word)

Thus, both DNA/RNA 'letters' and amino acids 'letters' form chains (*polymers*) that are referred to as 'words'.

In this paper, I will propose a more specific analogy between the architecture of language and the derivation that leads to proteins. This analogy is based on the components of language that we call *phonology* and *phonetics.* The phonology of a language characterizes the abstract mental 'code' of the pronunciation of words. The phonetic component handles the 'translation' of this code into an actual pronunciation. These two components, then, will be compared to the DNA/RNA representation and the translation into amino acids and proteins, respectively.

1. DNA > RNA > Amino Acid > protein

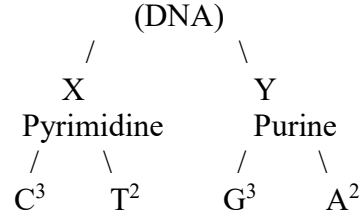
In this section I will sketch the relevant steps in the derivation of protein.

The human (somatic) cell has 23 pairs of homologue chromosomes. One member of each pair comes form the maternal egg cell, the other from the paternal sperm cell. In the process of fertilization, these two haploids (half sets), also called sex cells, or gametes) are combined to form a complete diploid set (a zygote), and from this a new organism develops.

Both members of each chromosome pair contain the same set of genes, possibly with different values (*allelles*). Simplistically, one gene could be about 'eye color', the values being the actual colors. If the values are conflicting, the *dominant value* determines the actual phenotypic manifestation of the relevant gene.

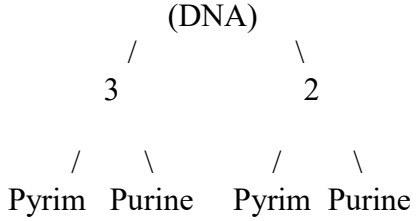
The most important material from which chromosomes are made is Di-oxy-ribo-Nuclear Acids (DNA). (Chromosomes contain roughly 50% DNA and 50% proteins for 'packing' the DNA.) Hence, we say that chromosomes are *strands* of DNA. A gene can be defined as *a string of DNA with a function*. Specific functions can range from determining aspects of the phenotype (such as eye color) to making proteins or enzymes that play a role in the 'household' of the cell. There are four such dioxyribonuclear acids:

(2) a.



Cytosine Thymine Guanine Adenine

b.



Cytosine Guanine Thymine Adenine

(2a) classifies the acids in terms of a grouping in pyrimidine and purine, whereas (2b) groups them in terms of their combinability. The latter property has been indicated by the superscripts 3 and 2; cf. below.

A DNA molecule contains a base, ribose (sugar) and phosphate. The differences between the four nucleic acids lies in the base. Purines have a double ring, pyrimidines have a single ring.

Any sequences of these four *DNA-letters* is a possible string of DNA. The DNA molecule contains two strands of DNA that go 'hand in hand' such that C always goes with G and T always with A:

(3)
CTTAGGTATGC
| | | | | | | | | | |
GAATCCATACG

A chromosome is thus formed by a long chain of DNA *pairs* (called *base pairs* or *nucleotides*). The correct pairing is obtained because both C and G have three

'connectors', while T and A have only two connectors. Hence a pairing of, e.g. C and T will fail because they do not fit. The connector-type has been indicated with superscripts in diagram (3).

This double strand (or *ladder*) is twisted in the form of the famous *double helix*. It should not escape our attention that the dual structure provides the basis for cell division (*mitosis*). In mitosis, the two strands are separated and each strand is provided with an appropriate new matching strand. This process, called *replication*, produces two exact copies (except for copying errors or *mutations*) of all chromosome-pairs within the nucleus of the cell. Subsequently, the cell splits, each new cell taking one copy of the replicated chromosome set.

(It is important to distinguish between the fact that each DNA strand has mirror-image copy *within* the chromosome, and the fact that each chromosome has a counterpart in the homologue pair.)

The production of sex cells or gametes (*meiosis*) also starts with replication, but first involves a type of cell division in which each new cell gets one member of each chromosome pair (including their copies). Thus we get cells with 23 *unpaired* chromosomes (and their copies). Then, as a second step, each resulting cell divides again, each new cell taking one copy of all the unpaired chromosomes. This gives what we called 'half sets' (haploids).

The sequences of DNA-letters form 'instructions' to regulate all that is going on in a cell, including mitosis and meiosis and the production of proteins and enzymes. Here we focus on the instructions that specify the production of *proteins*, but many instructions involve producing all sorts of enzymes and proteins that are necessary ingredients or helpers in this process. In addition, there are DNA-sequences that contain information on how, and when, to produce these enzymes and proteins. I will come back to that in more detail below.

The path from DNA to protein can be understood as a *derivation*, having an initial state (the DNA string) and a final state (the protein). Let us briefly review this derivation:

Transcription: a portion of the DNA string is copied into a RNA string. RNA is formed by a slightly different type of acids having one more oxide molecule. That's why we call them RiboNuclear Acids, whereas DNA, missing one oxygen atom at the 'second corner' of the ribose, is called di-oxide RiboNuclear acid. A further difference is that instead of the acid thymine, RNA uses uracil (or uradene); T and U differ in the base:

(4)

DNA	RNA
C	C
T	U
G	G
A	A

Various types of RNA-strings have to be distinguished.

When a string of DNA-letters is copied, we first have to *find* the relevant section of DNA and then we can form a string of RNA by copying the relevant section. The section of DNA that is actually copied is preceded by a section of DNA-letters that provide information *on* the copying process. This section is called the *promoter* (or *upstream region*); it contains regulatory DNA. Before copying takes place so-called transcription factors bind to the regulatory DNA strings in the promoter, and copying will occur only if the right sequence of transcription factors is present. The section of DNA that will actually be expressed into a protein is called the *structural gene*. Transcription factors can be enzymes/proteins that the cells have acquired from our food or other external sources, or it can be a protein that the cell itself has made.

The copy-machine is called the RNA-polymerase. This enzyme binds to the DNA-helix and copies a string of DNA; it binds at the so-called TATA-box, about 10 letters removed from the site where copying starts. The side of the double helix that is transcribed is called the sense strand. Transcription stops when a particular DNA-sequence is reached, like GGGUACCC. The C's will binds to the G's to form base pairs.

The string of RNA (still present in the nucleus) must now be *processed* into a string of mRNA (messenger RNA) that will make its way into the cytoplasm. RNA-processing involves three steps:

(5) a. *Prefixing* at the 5'- side a GG sequence (first G reversed)
b. *Suffixing* at the 3' - side a poly-A tail (50-300 A's)
c. *Splicing* (editing)

The RNA string contains instructions that include START (always AUG, methionine) and STOP signals (UAA, UAG, UGA). We could call these *boundary markers*. It also is also divided into 'chunks', i.e. parts that will or will not be relevant to the expression into a protein. Parts that will be expressed are called exons, parts that will not be expressed are called introns. One string of RNA can be spliced (by snurps or spliceosomes) into various strings of mRNA depending on which chunks will be expressed:

(6) DNA GGT AUG TCC ATT GCT GAC CGT UAG TAA

RNA **AUG TCC** *ATT* **CGT** *GAC* **CGT UAG**

mRNA AUGTCCCGTCGTUAG

Thus, one mRNA sequence can give rise to more than one protein by leaving out particular introns. In this way, we derive different varieties of a protein type, the choice between which is dependent on the local context of the protein (i.e. the locus in the organism where the protein is made and will be used).

Only when the mRNA string has been processes, it can leave the nucleus and go into the cytoplasm to be picked up by a Ribosome and be *translated* into an amino acid sequence forming a protein.

For the purpose of translation, the strings of letters are grouped in triplets, called *codons*. Each codon encodes an amino acid. There are 20 AAs. Since there are only four RNA letters, AAs can only be coded in terms of RNA-combinations. A combination of two NAs is not enough (4x4 = 16). Three letters gives 64 possibilities. That is too much. As a consequence most AA are coded by more than one triplet (sometimes 2, sometimes 4, once 6 and twice 1). Thus many codons are synonyms; I will go into that in section 3.

Translation: the mRNA codons are 'translated' into amino acids and groups of AAs form proteins.

(7)

```
DNA          GGT  AUG TCC ATT GCT GAC CGT UAG TAA
                  | | | | | | | | | | | | | | | | | | | | |        Transcription
RNA               AUG TCC ATT CGT GAC CGT UAG
                  \|/ \|/ \|/ \|/ \|/ \|/ \|/
                   c   c   c   c   c   c   c

mRNA           GG-AUG TCC ATT CGT GAC CGT UAG-AAAAA
                  \|/ \|/ \|/ \|/ \|/ \|/ \|/                     Processing
                   c   c   c   c       c   c
                   |   |       |       |                         Translation
AA                met thr     arg     ala                        (genetic code)
                    \   \       /       /
Protein                    PTN
```

The number of AA that define a protein varies from 5 to 1000. (7) is what I called a *derivation*. It contains 3 derivational stages: *transcription*, *processing* and *translation.*

The actual translation is done by enzymes called ribosomes, of which each cell contains 2 million. They take in the mRNA string on one end, each time reading two codons at the time. The first is called the P-site, the second is called the A-site (the workbench). A codon that is in the A-site is linked to a tRNA, which has two sides. The side that links to the mRNA codon is a complementary RNA string, called the anti-codon. The other side contains the amino acid. As the string of mRNA moves through the ribosome, the codon that is connected to a tRNA moves to the P-site. The codon that was in the P-site is pushed outside the ribosome and the amino acid which meanwhile has been bound to the previous amino acid forming the protein poylpeptide, is disconnected from the anticodon. (The anticodon then 'floats' into the cytoplasm).

The tRNAs are adaptor-molecules. They are produced by enzymes called tRNA-AA-synthetases, which is a protein that is itself formed on the basis of DNA. Thus the crucial information for translating DNA into protein, is itself encoded in the DNA-strands! If

these enzymes would mutate, we would effectively get a new genetic code, a new species.

I draw attention to the following important point: there is no natural relation a codon and the associated amino acid. *The relationship is arbitrary*.

A protein is a complex structure with at least three levels. The primary level is the sequence of AAs. The secondary level involves a coiling of the AA-string (e.g. like a telephone wire) due to the fact that AAs that are close to each other will attract each other. The tertiary level involves a coiling of the telephone wire, due to more attraction between more remote AAs. Thus, the information to get the secondary and tertiary structure comes from the AA sequences itself.

(8)

x x x x x x x x x x x x x x primary (AAs)

\\\\////// \\\\\////// \\\\\////// \\\\\\////// secondary

\\\\\\\\\////////// \\\\\\\\\////////// tertiary

AAs can be ranked in accordance with their degree of 'hydro-sympathy', i.e. there likelihood to dissolve in water. This property plays a role in determining the secondary structure. (8) does not exhaust the higher order structures. When several polypeptidea combine to form a protein, we get a quaternary structure. Also, proteins can form larger units: protein complexes.

Let me now first say a few words about 'how language works'. My characterization of language will be limited to the phonology and phonetics, and my wording will be such that the reader can anticipate the analogy for him or herself.

2. The structure of language

A language has a lexicon containing a 'code' for all the words. This code is called the *phonological representation*. Along with this code, we find the meaning of the word, and its syntactic properties. The latter comprise the word class (noun, verb, conjunction etc.) and its valency (i.e. how the word will or must combine with other words). In addition, we find information on the stylistic use of the word (formal, informal etc.). All these information types together make up, what we call, a *lexical entry* or *lexical item*.

The phonological code will eventually be 'translated' into another code that represents concrete sound. This code is called the *phonetic representation*. The mapping from phonological code to phonetic code is called *phonetic implementation*.

The phonological code has several layers of units:

(9) Elements

|
Phonemes
|
Morphemes

Elements are the building blocks for phonemes. A string of phonemes makes up a morpheme. A morpheme is defined as the smallest string of phonemes with a meaning or function. Thus *chair, crocodile, walk* are single morphemes. *Arm-chair, chair-s, friend-ship* consist of two morphemes.

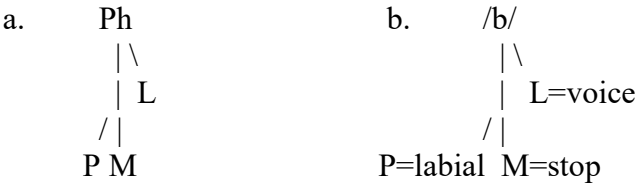
A phoneme consist of three unit slots for elements:

(10) a. **P**lace of articulation
b. **M**anner of articulation
c. **L**aryngeal specification

There are many theories about the nature and number of elements that may appear in each slot. In my own theory, there is exactly a four-way choice in each slot.

A phoneme, then, is a tripartite structure, just like the codon. The three slots in the phoneme structure are labeled L (laryngeal), M (manner) and P (place). The tripartite hierarchy that I propose for phonemes is not flat, but rather has two levels:

(11)



The phoneme /b/ is called a *voiced labial stop*.

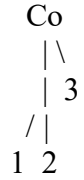
There is no linear order between P, M and L, as in (11). Rather, there is a hierarchical relationship. In the hierarchical structure, M is the primary position (called the HEAD), P the secondary position (called the COMPLEMENT) and L the tertiary position (called the SPECIFIER). Primary positions are dominated by a vertical line, non-primary positions by a slant line. The reasons for deciding which slot is primary and which are not have to do with the role that the elements in these slots play in determining the 'behavior' of the phonemes. Thus, M is the head because elements in M-position determine the place of segments in the syllabic organization of words. L is peripheral because the element in this slot is often redundant (i.e. not distinctive, without informational force).

Note that *linear order* and *hierarchical structure* are often substitutes for each other.

In section 5, I will argue that the linear order of the codon positions 123 can be linked to a hierarchical structure as well in which position 2 is the most important, and 3 the least

important position, position 1 being in between. This corresponds to the following head-dependency structure:

(12)



(This will be my contribution to genetics.)

I said that per phoneme slot there is a four-way choice. We can account for this by assuming that there are precisely four phonological elements. In my theory of phonological elements, I label them: Cc, Cv, Vc, Vc. Hence, the four choices are coded in terms of two letters 'C' and 'V'. The reason for coding the elements in this way is that they are ranked on a scale of *sonority.* The sonority-determining factors are C (low sonority) and V (high sonority).

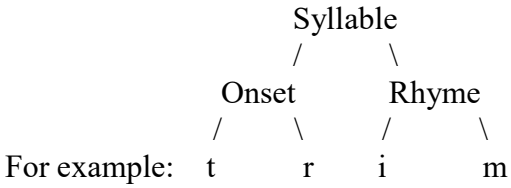
(13) a.

b. Cc Cv Vc Vv

The coding can be depicted hierarchically, as a template, which divides the phonetic spaces for place, manner and laryngeal into 4 subspaces, one for each element.

The precise phonetic implementation of each element is dependent on its position in the structure of the phoneme, and also on the position of the phoneme in the syllable structure. There are four positions in the syllable:

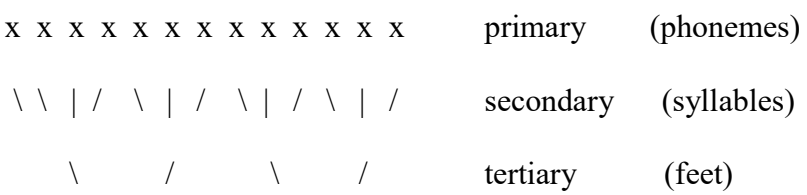
(14)



The context dependence of the interpretation of elements implies that real unit of interpretation is the phoneme (including its syllabic 'role') rather than the element. Elements cannot be interpreted in isolation given their 'polysemy'. Thus we can say that phonological elements have no independent phonetic interpretation (i.e. elements are not independently pronounceable). Thus phonemes are the smallest units that get 'translated' in terms of 'sounds'. When sounds are combined in the correct way, we get a *word*, a unit with a meaning or function.

According to phonological theory, words have a complex internal structure, which looks roughly as follows:

(15)



A foot is a rhythmic units consisting of two syllables. Above the level of the word here is further hierarchical structure: words combine into phrases and sentences.

3. The correct analogue

Most analogues with language are very simplistic. The DNA/RNA units are called 'letters' and the codons are called 'words'. Then, after translation into amino acids, the 20 AA are again called 'letters' and the proteins are called 'words':

(16)

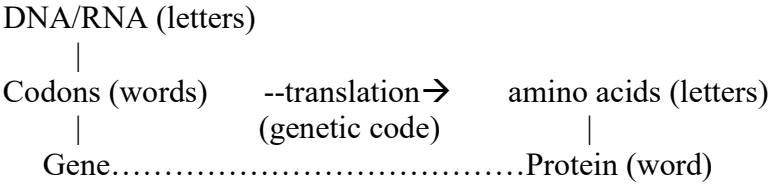
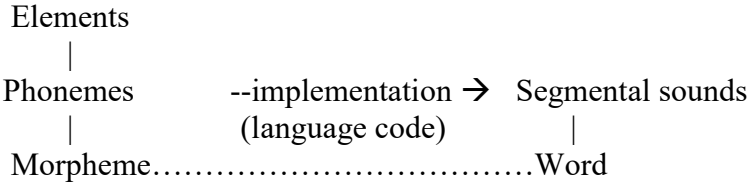


Diagram (16) is like (1), except I added the gene in the diagram (i.e. a sequence of codons).

I think that we can establish more interesting parallels. The basic idea is expressed in the following diagram:

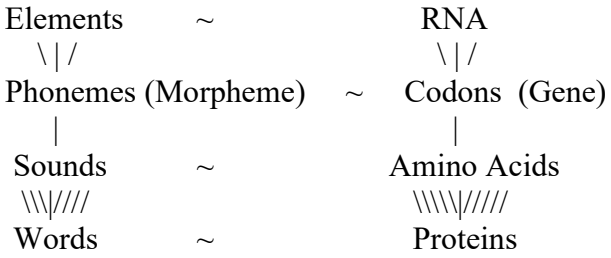
(17)



I use the term 'word' here as 'monomorphemic word'.

I regard NA-letters as analogues to elements, the building blocks of phonemes. Thus, codons are analogues to phonemes (and not 'words'). Translation is analogous to phonetic implementation, or pronunciation. Then, a sequences of amino acids (in themselves meaningless units, i.e. units without a function) make up a larger unit that does have a function or meaning, viz. words. (18) depicts the parallel in a different manner:

(18)



In the derivation of speech, we must 'blend' the sounds that make up the word into a smooth rhythmically structured, acoustic-articulatory stream. Thus, after translation we need a further system of rules to do that. These rules build the kind of structure depicted in (15). In Cellese too, we need further rules, viz. the rules that will construct the organization in (8).

What about transcription? If we follow the analogy in (17), the chromosomes that hold the DNA strands can be seen as a Lexicon, and transcription is like taking a word from the lexicon in order to pronounce it and use it in a phrase or sentence to express some meaning. Just like a lexicon in language contains a symbolic representation of the pronunciation of words, so does the DNA-lexicon (the collection of chromosomes) contain a symbolic representation of the expression of proteins. Transcription, then, is *lexical retrieval*: the symbolic representation of words is copied and then transmitted to the phonetic implementation module (the cytoplasm) for translation. The concept of copying for lexical retrieval is appropriate because, when we speak, we do not literally *remove* words from the lexical inventory when we use them in a sentence. If we did that, the lexicon would be quickly empty and we would be left speechless.

Note that this analogy presupposes that we view the relationship between the phonological and the phonetic as arbitrary. This is actually a controversial point, since in almost all theories of phonology & phonetics, the phonological primes are inherently defined as phonetic. Despite, this consensus, it must be said that this view cannot be correct and is actually hard to understand. Phonological representations are *cognitive* objects, whereas phonetic representations are *physical* objects. Cognitive and physical objects have a different ontological status. Even when we ultimately would reduce the cognitive object to a physical state of the brain, the physical brain structures would still be distinct from the acoustic phonetic structures that form the pronunciation.

Just like nucleic acids and amino acids are different chemically, mental representations and pronunciation are different physically. Thus there must be a mapping from the phonological objects to the phonetic objects. This mapping is a table, a code, just like the genetic code. We could call this code, the language code. One might ask whether the 'language code' is innate or must be acquired. I hold the view that it must be acquired. Thus, the genetic code *is* genetic, but the language code is 'memetic'. There is, of course, an innate aspect to language, viz. the ability to acquire the language code.

What about 'processing'? Is there 'processing' when we pronounce words? Yes, there is. Words can vary in their precise pronunciation depending on their local context, i.e. on the surrounding words in the polymorphemic words, phrase or sentence. This is called *allomorphy*. I compare this to the different proteins that can be derived from the same AA-sequences. A case of allomorphy that makes the analogy with introns and exons especially clear is the following.

Allomorphy in Tonkawa
An American Indian language once spoken in Texas (data Hoijer 1933, 1949), Tonkawa, shows the following instances of allomorphy:

(19)

picn-o?	he cuts it	**picna**-no?	he is cutting it
we-**pcen**-o?	he cuts them	we-**pcena**-no?	he is cutting them
ke-**pcen**-o?	he cuts me	ke-**pcena**-n-o?	he is cutting me
picen	steer		
notx-o?	he hoes it	**notxo**-n-o?	he is hoesing it
we-**ntox**-o?	he hoes them	we-**ntoxo**-n-o?	he is hoesing them
ke-**ntox**-o?	he hoes me	ke-**ntoxo**-n-o?	he is hoesing me
notox	hoe		
netl-o?	he licks it	**netlo**-n-o?	he is licking it
we-**ntal**-o?	he licks them	we-**ntale**-n-o?	he is licking them
ke-**ntalo**	he licks me	ke-**ntale**-n-o?	he is licking me
naxc-o?	he makes it a fire	**naxce**-n-o?	he is making it a fire
we-**nxac**-o?	he makes them a fire	we-**nxace**-n-o?	he is making them a fire
ke-**nxac**-o?	he makes me a fire	ke-**nxace**-n-o?	he is making me a fire

each of the stem morphemes has four different shapes:

(20)	picn	picna	pcen	pcena	=	picena
	notx	notxo	ntox	ntoxo	=	notoxo
	netl	netle	ntal	ntale	=	netale
	naxc	naxce	nxac	nxace	=	naxace

We need a couple of rules to derive each shape in the appropriate context. I will informally state these rules, but not discuss them here:

(21)
a. V -> 0 / - V (delete a vowel before a vowel)
b. V -> 0 / [CVC- (delete the second vowel in the word in verbs or in env. /CVC-CV if ordered after c)
c. V -> 0 / -]N (delete final vowel in noun)

Thus the morpheme *picena* is parallel to a particular mRNA unit that corresponds to different protein expressions depending on context (i.e. on the location of cell in the organism).

With respect to the secondary, and tertiary structure of proteins, we can see a parallel with syllable structure if we see the degree of sonority of a segment (as inherited from the elements) as analogous to the hydro-sympathy property of amino acids, and if hydro-sympathy (like sonority) determines the secondary structure. Structurally, the tertiary structure of proteins is like the foot structure of words, i.e. compare (8) to (15). Words, when combined, form phrases, and the restrictions on combining are called valency. Again, we might see a parallel here in the possibility for proteins to combine with other proteins.

There is actually a fine point that we must make here. If the sonority of a sound corresponds in a non-arbitrary way with the elements in the phonemic code, than, perhaps the relationship between the phoneme and the sound is not completely arbitrary. We might then ask whether, perhaps, there is also an trace of non-arbitrariness in the genetic code. In other words: is it perhaps the case that the hydro-sympathetic property of AA corresponds in some way to certain part of the codons. The answer seems to be affirmative: the second position in the codon is, in fact, the best predictor for the degree of 'hydro-sympathy'.

Human language has the property of dual patterning. One patterning involves the structure in the perceptible form of language, the other patterning being the morpho-syntactic structure. Does Cellese have dual patterning? The answer is no. Notice, though, that language only has one patterning above the lowest level of meaningful units, i.e. morphemes. Morphemes, in language do not have dual patterning; they only have the patterning in the form. From the view point of morpho-syntax, morphemes are ‘atoms’.The structure of Cellese concerns the “morphemic” organization, and there is thus only one patterning (analogous to the form patterning in language). The other patterning would become relevant at the level of protein combinations, but there would only be a reason for postulating a dual patterning if we could find a mismatch between the phonological and the morpho-syntactic structure.

4. The structure of codons

In this section, I use the proposed analogy as a heuristic device by raising the question whether the four nucleic acids can be ranked just like the four phonological elements. I propose to focus on the *informational strength* (or functional load) of the bases. (To follow this section one needs to take out a table containing the genetic code.)

The nucleic acid C in second position of the codon correlates with total redundancy of the third position. The letter A in second position correlates with allowing a distinction between purine and pyrimidine. This means that C has a 'the strongest informational'

load, whereas A is the weakest. G and U are in between having either full redundancy or a pu/py distinction. Thus:

(22) Codon position 2: C >> G, U >> A

A distinction between G and U (again in terms of informational load) can be made (with a few 'idealizations') as follows. With G and U in second position, we can look at the first position, and we then notice that G and C in first position correlate with (almost) total redundancy of the third letter. Thus G has more informational strength than U. This means:

(23) Codon position 1: G, C >> U, A

Adding both rankings (abstracting from position) we get the ranking:

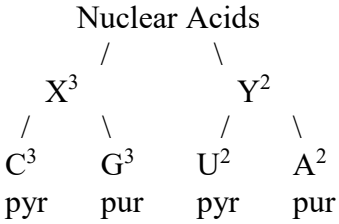
(24) Ranking I C >> G >> U >> A
pyr pur pyr pur

Since within each base pair pyr is 'stronger' than pur, we can also say that:

(25) Ranking II pyr >> pur

Ranking I correspond to a grouping of the bases in the following hierarchical structure:

(26)



The grouping in (26) classifies complementary bases together. If we call X ‘strong’ and Y ‘weak’, then within in each group the there is a stronger and a weaker acid. Thus C is a ‘prototypical’ X^3 base, just like A is a prototypical Y^2 base. Perhaps, 'prototypical' translates into 'most frequent' (e.g. it could be that amino acids with C, especially C in second position -for reasons given below- are the most frequent AAs; Is this so?). This line of reasoning also leads to the question as to whether there any reason for regarding U and G as somehow ‘intermediate’ bases? Is U somewhat purine like and is G somewhat pyrimidine like?

A second question that we could raise, again using the proposed analogy as a heuristic device, is whether there is any reason for proposing that the codon (just like the phoneme) has a non-flat structure, as in (27):

(27)

P1, as we have seen, has some predictive power with respect to the third position, so it is more important than P3, but less important than P2. The second position is the most important position, firstly, because it is the strongest predictor of what happens in the third position (cf. supra). Secondly, the second position is the best predictor for the degree of 'hydro-sympathy' and if that corresponds to syllable structure, than P2 (like Manner in phonemes) should indeed be the head. The third position, containing the least useful base is the most peripheral, i.e. the specifier position.

5. Different analogues

In this section I merely mention different ways of thinking about analogues between human language and the language of the cell. The ways I mention (except for the one in 5.4.) are not incompatible with the preceding analogue. Rather they derive from taking larger units as the point of departure.

5.1. Genes:alleles = features:values

Another parallel between DNA and phonology can be made. Let us compare the entire genotype with a phoneme. In the earlier section, a phoneme is characterized in terms of a set of elements, but another, more traditional view is that a phoneme is a set of distinctive binary features. ‘Binary’ means that each feature has a plus and a minus value. (In the ‘element’ view, this binary contrast is expressed in terms of presence of elements versus absence of element.)

We can regard the two values of features as *alleles*. A feature cannot have two values at the same time, and in most instances one value can be said to be dominant (or marked). The dominant value is the one that is pronounced in case there is a conflict. This is sometimes exemplified in vowel harmony processes. We can compare the entire genotype with a word. In vowel harmony languages, when the two values of one feature occur within the same word, the dominant value determines the ‘phenotype’ (i.e. pronunciation or manifestation) of the word.

5.2. Genes:organisms = phonemes:words = words:sentences = atoms:molecules

Abler (1989) compares language, chemistry and genetics, arguing that in all three cases we are dealing with self-diversifying systems, or ‘Humboldt-systems’: systems that make infinite use of finite means. We have a finite set of building blocks and a finite set of rules for combination. These two finite sets together allow for an infinite set of expressions.

In chemistry, the elements (or rather atoms) are the building blocks, and molecular structures are complex expressions. In language, we have dual patterning. Hence the property of self-diversification occurs twice. Phonemes are atoms of morphemes, and words are atoms of sentences. Applied to genetics, genes are the building blocks and organisms are the complex expressions.

Abler remarks that both chemistry and phonology have a periodic system, the periodic system of elements and phoneme charts, respectively. In genetics, there appears to be no periodic system, however. The 'lexicon of genes' has not been analyzed to begin with. It is therefore difficult to say at this point, whether the collection of genes forms a periodic system.

The calculus of combination is a *particulate* system, rather than a blending system. A particulate system is a system that combines the building blocks in complex structures in a compositional manner, i.e. the building blocks keep their own identity, attributing their properties to the complex structure. Interestingly, the calculus for combining sounds into words (after phonetic interpretation has taken place) may have the superficial appearance of a blending system.

5.3. Spreading of gene-values = spreading of language

In work by Cavalli-Sforza & Cavalli-Sforza (1995) a comparative study is made of the spreading of genetic variants over the earth to the spreading of language variants. This work reveals a parallel between the distribution of languages and species.

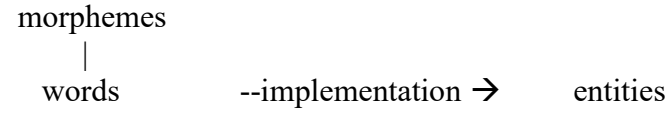
5.4. Form and meaning

There is another and completely different analogue that we could pursue. Calling to mind that the relationship between RNA and AA is arbitrary. We could compare that relation to the relationship between form and meaning. The relationship between form and meaning, after all, is arbitrary.

5.5. Proteins as truth values

Elsewhere, I have argued that phonology & phonetics are organized fully parallel to syntax & semantics. If that is so, we could also use the latter side of the grammar as a model for Cellese. That view can be depicted in a diagram that parallels the diagram in (17):

(28)



| |
sentence……truth values

Since language , in fact, has a dual articulation (which are fully parallel), and since Cellese has only one articulation (see supra), we, in fact, expect, that we can make a parallel between Cellese and either phonology or the syntax side of grammar.

6. Conclusion: what does all this mean?

A question is why the analogy as outlined in section 4 should exist? A possible answer is that we are bound to find the same kinds of structure everywhere, because the structures are properties of our observing and analyzing mind.

An alternative is that genetic and linguistic structures have a common ancestor. Thus, the phonology-phonetic interface would then be a 'copy' of the RNA-AA interface (i.e. the genetic code).

A third option is to say that systems that perform a certain function (i.e. passing on information) necessarily have certain properties.

I will leave the exploration of possible explanations for the parallelism between Cellese and (both sides of) human language for another occasion.