Building Bridges

A Cognitive Science Approach
to a
Human Dolphin Dialog Protocol

Thesis for the
Cand. Philol. degree
in
Language Logic and Information
University of Oslo
2002

Preben Wik

Building Bridges

A Cognitive Science Approach
to a
Human Dolphin Dialog Protocol

Building Bridges

Bridges are impressive engineering accomplishments, standing like symbols of mankind's many great achievements. But bridges are also something more than physical objects, connecting two pieces of land. A bridge is a metaphor. You can bridge cultural differences, racial differences, misunderstandings, species barriers and so forth. In general, you use bridges to cross gaps.

In this thesis I will be discussing the bridging of several gaps. Language gaps, cognitive gaps, gaps of belief and presupposition, and gaps between species, to mention some. I will also be discussing methods to cross gaps, and problems inherent in certain kinds of gaps. The main subject is a discussion on an inter-species communications bridge. More specifically, I will be discussing a human dolphin communications bridge.

This is not the first time human dolphin communication has been up for discussion. For centuries there has been an idea of dolphins having their own language. Aristotle wrote in Historia Animalium [1] about dolphin communication. During the cold war, the US and USSR militaries spent millions of dollars on dolphin communication studies. Books have been written and movies have been made on the subject. Is there any basis for this interest in dolphin communication? There is little, or no doubt that dolphins communicate vocally. Could this vocal communication be used as a bridge into our way of communication?

Different species have evolved specific adaptations that make them more or less well-equipped for specific tasks. Claws for climbing trees, fins for swimming, and wings for flying are adaptations some animals have, that humans have not. This has been put forward as an argument for other animals perhaps being mal-adapted for linguistic communication [2]. So if claws are needed to climb trees, and wings to fly, what is needed for linguistic communication? Dolphins are chosen because they have evolved several features that make them very well-adapted for linguistic communication, and hence good candidates for such an endeavor for reasons that I will come back to later. Is it possible to build this bridge? How should it be done?

Bridge building is an engineering job, but when bridges of a new kind are to be constructed, theoretical work has to precede the actual construction work. From a theoretical point of view, one might say that two objects that belong to different domains could be bridged if some requirements are fulfilled. Typically, if you want to build a bridge of any kind, the domains to be bridged can not be too far away from each other. Distance, that is, how different they are, is one of the hinders. But as with real bridges, new understanding and technology enables us to build bridges where it was previously impossible to do so. Any bridge will have capacity limits. A small suspension bridge can carry people across, but perhaps not cars or trucks. What kinds of units that can be transported across, is an important aspect of the bridge. With human language we are able to express many things, but certainly, there is a vast area

outside the realm of language. The bandwidth, i.e. how many units that can be transferred in a certain amount of time, is another aspect. In other words: is it a one-lane or a six-lane bridge?

In remote areas such as the Andes and the Himalayas, there is a certain kind of bridge called a "flying fox". It is a small cage hooked up to a wire across a ravine, large enough for a person, an animal, or some cargo to fit. A consequence of the flying foxes is that villages that were previously totally isolated, get in contact with each other and can exchange knowledge and ideas and learn from each other. An isolated idea on one side combined with an isolated idea on the other side can have a catalytic effect on each other and new knowledge will arise. This is one of the effects of bridges. Domain A and domain B connected with a bridge can become more than A+B, because by linking them together, new properties may emerge. This is what a network does, and a network is in this respect a system of bridges. Linking many individual units together creates a new entity that is more than the sum of its parts. A person is more than the sum of its cells, a society is more than the sum of its people, and the Internet is more than a bunch of computers connected together.

Inter-Disciplinary Bridges

Building bridges between different scientific disciplines has also led to new knowledge analogous to the flying fox example. Bridging between biology and chemistry has created one of the most dynamic and expanding branches of science today: biochemistry. Cognitive science is also a relatively new science that has come out of the bridging between philosophy, psychology, artificial intelligence, neuroscience, linguistics, and anthropology. Cognitive science is the inter-disciplinary study of cognition, mind and intelligence. It has so far not been too concerned with cognition in an inter-species perspective.

The bridge metaphor will stay with us throughout this thesis. Language will be seen as a process of crossing gaps at many levels. At its top level, language is a bridge between two or more individuals, enabling an exchange of intentions and ideas that, by many people, is considered the most important innovation in the history of mankind.

By using the bridge metaphor we can look at language and communication issues in a more detached way without getting caught up in the particularities of human language.

A language is a metaphorical bridge, enabling information of a certain kind to be transferred between two human beings.

A protocol is also a bridge, enabling information of a certain kind to be transferred between two computers.

The 'Human Dolphin Dialog Protocol' which I am going to outline in this thesis, will be a bridge, enabling information of a certain kind to be transferred between humans and dolphins.

Inter-species communication, one of the biggest gaps around, is not likely to be bridged with a 6-lane-freeway right away. Part of the thesis will be an attempt to narrow the gap intellectually, by pointing out structural similarities that will simplify the construction. The rest of the thesis will be looking at the remaining distance, and what kind of methods can be utilized to over come it.

New materials such as steel and concrete combined with new building techniques, enabled us to build bridges over previously unbridgeable gaps. The tremendous advances in computer power and new algorithms for processing data is today making it possible to create a new kind of communications bridge, something that was not possible even few years ago. A new kind of bridge will be presented in this thesis. If dolphins are capable of, and interested in utilizing this bridge, we would be able to achieve some form of inter-species communication. This thesis then, is kind of an engineering job, and the purpose of the project is not to prove that dolphins can acquire language, but to make that investigation possible by providing a bridge.

Reading guide

Different readers will have different backgrounds, and motives for reading this thesis, and some are perhaps only interested in bits and pieces of it. Although it is written in with the intention that it should be read starting with chapter one and finishing with chapter seven, I will make an overview, based on the various topics that are being presented, to guide the impatient reader into relevant sections.

A new software is being presented in chapter 7, which is the core of the thesis. The other chapters are building up an understanding of what it is doing, and why it is needed. It is written in the form of a user manual, and people with a good knowledge in DPS and Neural network classification strategies, can skip chapter 6 which is explaining these issues. People with a background in linguistics, may skip first part of chapter five, as this is an introduction into central concepts in linguistics. Chapter three is a brief introduction into Douglas Hofstadter's analogy-making ideas, plus a short story of a personal experience with dolphin analogy-making, and not crucial to understand the rest. Language in general is discussed in chapter one, four and five, brains and cognition in chapter one, two and three, and programming and engineering aspects in chapter six and seven.

Chapter 1: **Language as a Bridge**

- is an abstract description of the landscape in which this bridge belongs. - It will form the base perspective on how language is treated throughout the thesis. Emphasis will be put on the division between the internal processing of language on the one hand and the external system that binds the users together on the other hand. Frege, Saussure, memetics, universal Darwinism, complexity theory, classification, and culture will be linked together in an inter-disciplinary perspective of language.

Chapter 2: **Brains and Neural Networks: Pattern-Acquisition Devices**

- is an overview of the fundamental mechanisms used in language processing. Emphasis is put on the plasticity of brains and how experience shapes the brain. The brain is seen as a general pattern-matching machine and neural networks are described as functionally similar to brains in this respect. With the same general architecture, neural networks can learn to recognize a vast variety of things, including aspects of language.

Chapter 3: **Creative Analogies**

- is discussing Douglas Hofstadter's ideas that analogy-making lies at the heart of intelligence. A personal experience with a dolphin in Hawaii is given as a possible example of analogy-making dolphin intelligence.

Chapter 4: **Further Requirements for Language Acquisition**

- is discussing the importance of an interface. Voluntary motor-control of speech organs, or other organs that can be used for symbolic reference, are emphasized as a requirement for communication. The chapter also looks at the communicative abilities of different animals, and discusses the terms communication vs. language.

Chapter 5: **Linguistic Aspects of the Bridge**

- discusses some considerations and choices one must take in the design of a new language and goes through some linguistic concepts such as: phonetics, morphology, syntax, semantics, and pragmatics.

Chapter 6: **Computational Aspects of the Bridge**

- is focusing on how to bridge the barrier of incompatible interfaces. It is also a map of the landscape in which this bridge belongs, but from more of an engineering point of view. Digitizing of audio, feature extraction, classifiers, neural network architectures, Viterbi searches, self-organizing feature maps, learning-vector quantization, and dynamic time-warping are being described.

Chapter 7: **General Audio Pattern Recognizer - GAPR**

- is a description of the software that was made for this thesis. The General Audio Pattern Recognizer: GAPR (pronounced ‘gapper’ as in someone who crosses a gap) is a tool for recognizing audio patterns and translating them into symbolic representations. It will take care of the interface aspects of the bridge.

Acknowledgments

I would first of all like to thank my family Li-Hui and Ronya for putting up with me and supporting me during this time.

Big thanks to my friend Kim Sørenssen who has been a constant source of inspiration and assistance on all levels - language, structure and ideas - in this project. IOU

Thanks to: Ken Marten, Sea Life Park Hawaii, and EarthTrust for help, and access to their wealth of dolphin recordings.

Thanks to: Frode Wik, Simen Gan Schweder, and Jan Dyre Bjerknes: friends who have read blue prints of the thesis, for invaluable criticism and comments.

Thanks to: my supervisor Herman Ruge Jervell for the guts and integrity of allowing me to investigate such a controversial subject, for believing in me, and for giving me good guidelines along the way.

Last but not least, thanks to my mother and father(s) for a good life.

External code

The software written for this thesis is, apart from my own code, using code from two external sources.

1. CSLU speech toolkit from Oregon Health and Science University (OGI) (http://cslu.cse.ogi.edu/toolkit/) CSLU offers, apart from many other things, some packages which extends Tcl with dedicated c code for processing audio. The packages used in GAPR are: Lyre, Wave, Rtcl, Audio, Sdet, Mx, Prep, TTS, and Features. Furthermore, one of the modules in GAPR, The AudioTab, is a reworked version of CSLU's "Speech view" which handles opening, viewing, recording and playback of audio files.

2. LVQPak from Helsinki University of Technology (HUT) (http://www.cis.hut.fi/research/lvq_pak) Home of Tuevo Kohonen, creator of the Kohonen networks and the LVQ algorithm. HUT offers both the SOMPak and the LVQPak, two packages of c code dedicated to statistical classification processing by the SOM and LVQ algorithm respectively. An earlier version of GAPR had them both incorporated, but the current version is focusing solely on the LVQPak.

Table of Contents

– 1 –
Language as a Bridge

The endeavors of cognitive science are exiting for a number of reasons. Partly because it is exploring something that is such an intimate part of ourselves – our own minds – and yet is something we know so little about. We don't really know what is going on in our heads when we reason. We know even less about what is going on in the heads of other animals, mainly because we can't ask them. If we are to attempt a bridge-building project, maybe it's a good idea to start by finding out how wide the gap is? How different are we?

I will start by investigating some fundamental aspects of cognition and show that they are basically the same in all animals of a certain complexity. I will then continue to see whether language is based on those same fundamental aspects.

1.1 Classifications, Abstractions, Tokens and Types

Part of being human is to classify things all the time. Sorting out sensory input and classifying it, is one of the basic jobs of brains. Any sensory input must be classified somehow, in order to make sense. In fact, the survival of any mammal will be dependent on its ability to classify things well. For a rabbit to be able to tell the difference between a carrot and a fox, a classification mechanism must take place. Being able to run fast is an important quality, but even more important is knowing *when* to run. Classification is a kind of abstraction. Making an abstraction can be described as a mental crossing of the gap from token to type, a basic cognitive bridge. Let me expand on this for a moment and at the same time show how language is closely related to the same cognitive bridge – our ability to make abstractions.

For example, sensory data of a bowl of fruit is sent through our eyes, and is being processed somehow in our brains, so that we perceive the data as a bunch of separate objects grouped together in a cluster, not just as one big blob. Each of the objects will have shape, texture and color. Each object can be viewed as a token of some kind of thing, perhaps a kind that we have seen before. A process will take place where the

shape, texture and color properties of each token will be matched up against some internal concepts in a search for a type that the object can be classified as. If a match is found, we recognize the object as being a token of a certain type that we already have a concept for in our minds. If, on the other hand there is an item in the bowl of fruit that we have never seen before, no such match can be made, and a new category must be created for objects that have this novel shape and color.

So far, the example is something most animals could have done, analogous with the carrot/fox example above. But as a human being, we can pick up the item and ask someone next to us: "What is this?" Perhaps the answer comes as: "It's a wing-ding". A remarkable feat known as symbolic reference has taken place. We know that the sounds the person next to us made is not the same as the object in the fruit bowl. It is a name – a label – that we can use to associate with the newly created concept in our minds, and the next time someone makes the same sound, we can bring up an image of that concept in our minds and there's a good chance there will be a correspondence with the concept in the other person's mind. The general function of symbolic reference is – as implied by the name – to refer to a concept by the use of a symbol. Sounds can work as a symbolic reference, and is our standard choice, but we could just as well have used hand signals, dots of ink on paper, flashing lights, or anything else. That is because there is no natural relation between the label and the concept: i.e., *the choice of label is arbitrary*. An important differentiation is established between three aspects of a word: its label (the symbol), its sense (concept), and its reference (an item in the external world). I will come back to the concept of symbolic reference many times in this thesis, as it is the fundamental abstraction on which language is built.

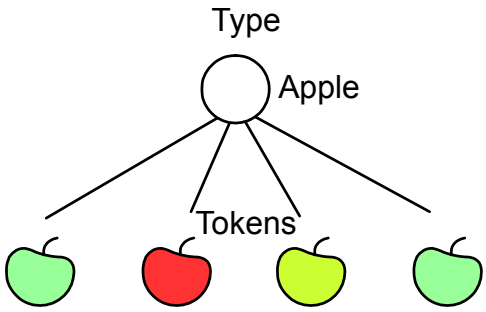


Figure 1. Four tokens of the abstract type 'apple'.

We don't usually think of apples and oranges as being abstract. They are in a way the opposite of abstract – very real. The point is that the apple we hold in our hand is not abstract, but by making a symbolic representation of it, we must have a generalized idea of which items can be called apples, and this idea is abstract. The act of finding the correct symbolic representation for a thing is an abstraction. Abstractions are layered, and where a token can become a type on one level, that type can in turn function as a token on another level. The type 'apple' is a token of 'fruit', etc.

When children are exposed to new categories, their interpretation is initially very general – any animal or cute toy may for a while be a 'doggie' – but with new categories being introduced to the child, the doggie will soon meet clear boundaries towards cats, teddies and so on. Thus the child's interpretation of the category is increasingly specialized. In this fashion we build up elaborate hierarchies of concepts in our minds, linked to each other in networks called conceptual structures. We know that in every human mind there is a conceptual structure of some kind. We also know that it differs from person to person, merely by the fact that we all have different backgrounds that has shaped our conceptual structures in unique ways. We don't share the same vocabulary, or the same insights. The first words we learn will serve as pegs for other words to hook onto, words that would otherwise be impossible to understand. A child could for example not be explained the concept of interest rate without first having the concept of money well grounded.

Animals too, have conceptual structures although not necessarily linked up with labels. The simplest form would be dividing the world into 'friend' and 'foe', and a continuous division into further categories may continue from there. The basic cognitive bridge of abstraction is a general quality of all brains known as pattern-matching. We will return to this concept later. Higher cognitive traits may result from this basic quality. Planning, decision-making, attention, thinking, evaluating, insight, creativity, choice, purpose, seeking, planning, generalization, judgment, introspection, interest, preference, discrimination, learning, habituation, memory, recognition, retention, and knowledge, are all brain functions neuro-scientists claim are expressible by the circuitry of the brain of any mammal. [3] The differences between human mentality and that of other mammals can be seen as quantitative rather than qualitative. A difference of degree and not of natural kind.

What about language? I will attempt to show that the required processing abilities that a brain must have in order to generate and understand language, could be aspects of this same pattern-matching that our brains are performing, and hence shared by all mammals as well as many other animals. Could there perhaps be other details such as the size of short-term memory, the presence of an efficient sound-producing organ, or cultural effects that has prevented the natural emergence of language in other species?

Before we can look at the actual bridge it is necessary to get a clearer picture of these issues. We must understand the landscape first, before we can decide on what kind of structure to choose. Structural similarities and differences on the two sides of the bridge (sender and receiver) will affect the construction. Right now we don't even know how wide the gap is.

1.2 Bridging Internal Gaps

The German mathematician and logician Gottlob Frege was interested in the computational aspects of language. He recognized that there are certain gaps that must be bridged in the process of communicating a message from one individual to another. The following section will look at language in a Fregian manner, using the labeling of Herman Ruge Jervell [4]. Schematically it looks something like this:

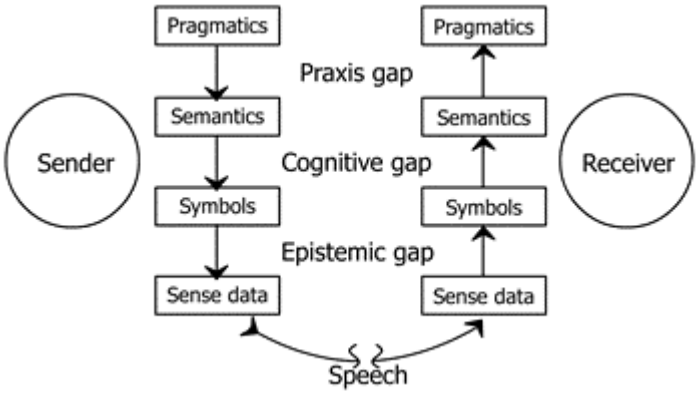


Figure 2. The gaps that must be bridged for communication to take place, according to Frege.

From a receiver's point of view, a message might begin with a stream of audio signals (sensory data) that the ear will pick up, and that will be processed and categorized into a string of sound symbols. This process we will call 'bridging the epistemic gap'. Audio tokens of some kind will be transformed into types of some kind. Please note that, although an abstraction has been made, and a transformation from token to type has been achieved, there is no meaning to the message yet. You might hear a sentence in a language completely unknown to you, and yet manage to write down the sounds you hear.

Let's say for example that you have just arrived at an airport in China, and a person is insisting on telling you something apparently important, but totally unintelligible to you. So you listen and faithfully write down the sounds that you hear uttered by the person in front of you. Perhaps you write: "Jeli shr nü shen". Then, you bring the piece of paper to a friend of yours who you believe is able to make sense of the letters you jotted down, and asks him if he knows what this might mean. The combination of letters is now again viewed as a token of some kind, but linked to concepts, with reference to the real world. Only a person familiar with that particular language will be able to transform this token into a type, by bridging 'the cognitive gap'. Your friend might go: Ah, "Jeli shr nü shen" is Chinese and means 'Here is woman'. Your friend might add: "But who told you this, and why?" and suddenly you realize you (being a man) had incorrectly entered the women's rest room. You have bridged 'the praxis gap' something your friend could not, because he didn't know the context in which to put the sentence.

In Jervell's classification, 'the praxis gap' is where the judgment of the sentence is made, and a truth-value will be assigned. A sentence like: "The ball is red" will be true if the ball mentioned is indeed red, otherwise it will be false. In the pragmatics of the real world however, quite a different judgment will be placed on the sentence (or perhaps on the person who uttered the statement). Say I walk up to a stranger on a bus stop, hold up a red ball and say: "The ball is red". The stranger's first thought is probably that I must be stark, raving mad. After a second, seeing that I look quite sane, he might consider the possibility that I'm speaking in code, and confusing him for a Russian spy, or proposing something indecent. He is not going to be concerned with the truth-value of my statement. He will be wondering *why* I said what I said. What my intentions are. That is because ultimately, we communicate in order to convey intentions of some kind, and we use language as a means to do it. When we bridge the praxis gap, we have understood the intentions of the sender.

Similarly, a sender will begin a communicative act with an intention and then traverse the gaps, assemble what is needed along the way, and finally move his larynx, tongue, and lips in order to produce some sounds that will work as an appropriate representation for his initial intention.

With this view of language, any system, regardless of design, that manages to transfer the pragmatic intentions between a sender and a receiver successfully, will qualify as a language. Whether the language scheme is using this or that medium (e.g. sound, images, sign language, Braille, flashing lights) for transferring the data, or whether it is utilizing this or that style of syntax, is irrelevant. What is relevant however, is how much, and what kind of messages can be transferred with this scheme, i.e., the expressive power of the language. Street signs, body language, and simple call systems could be considered a transfer of intentions, and therefore qualify as a language on these terms. But they fail to enable the transfer of certain kinds of information: for example past events, so we must ask ourselves: "What kind of intentions is this language able to convey?" When designing a language, we must also ask ourselves what kinds of intentions are desirable to communicate.

Engineering Communication Bridges

Apart from the internal bridges that must be traversed in the heads of the sender A and the receiver B, there is a critical point of transfer between A and B. In a normal face-to-face conversation the air is working as a bridge, carrying the sound waves from A's mouth to B's ear. This is however, also a domain of possible engineering. A telephone could be inserted in the gap, enabling the conversation to take place even if A and B are located at opposite sides of the earth. A camera/TV, a radio, a piano, and Internet are other examples of engineered bridges that can be inserted here. (This is also the domain for the General Audio Recognizer, GAPR, the software that will be presented in chapter 7 of this thesis).

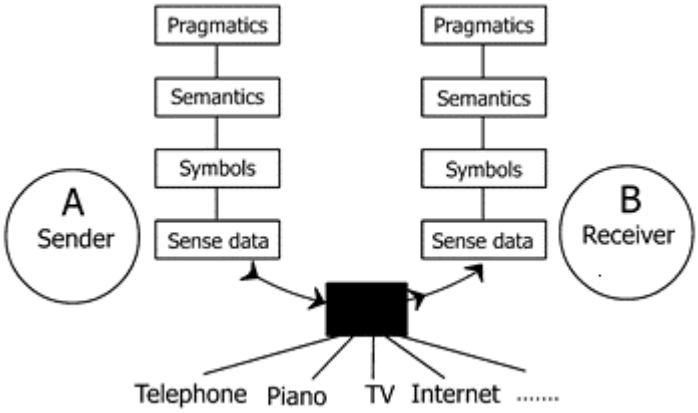


Figure 3. The gap between sender and receiver: a domain of possible engineering.

1.3 The External Bridge: Saussure's Perspective.

In order for language to work in the manner described above, there is yet a crucial component that we - haven't discussed. That is the correspondence between how A and B interpret the sounds that are being sent back and forth. Where do the rules needed to interpret the messages reside? Language may demand large and complex brains in order to exist, but equally important, it also demands a common system for the users to use. In order to read the following passages correctly, a shift of perspective is needed. Focus is no longer on the individual language user, but on the whole population, and the system that all language users are a part of. One of the first linguists to attempt a structure of language, was the Swiss linguist Ferdinand de Saussure (1857-1913). Although his main academic focus was on Indo-European languages, he is mostly remembered today for his lectures on general linguistic theory. Saussure pointed out that we need to make a distinction between language in use, *le parole*, and language as a system, *le langue*. *Le parole* is the stuff residing in people's heads. The system is not something in people, but *between* people, much in the same sense as culture. This system, although not created by any one in particular, dictates the rules of how the language should be spoken, and gives meaning to the units that

the language consists of. The system transcends the choices of the individual users, in that no individual has the power to change the meaning of a word, even if he wanted to.

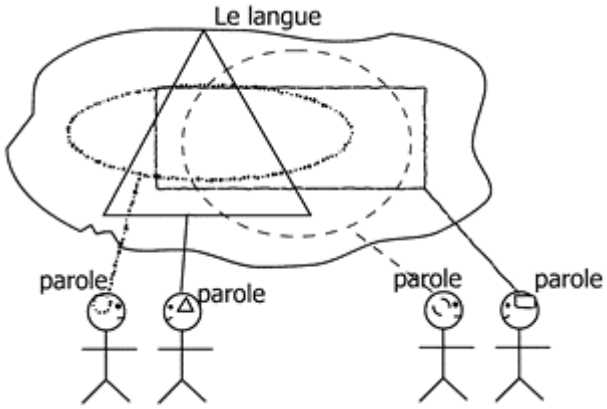


Figure 4. Le langue *is not something in people, but between people.*

Even if this system does not contain any physical objects, only information, it is possible to view the system as analogous to an evolving ecosystem. The insights of Charles Darwin changed the way we look at the origin of man and how species evolve. What is called the Darwinian algorithm, has in recent years also been applied to other systems that at a certain level of abstraction have similar properties. This inter-disciplinary exercise can be fruitful in several ways. Some of the tools of analysis from biology can be directly mapped on to other systems. We can say that languages evolve, since both grammar and the meaning of words change over time. Which elements are in the system at any given time (i.e., which parts will survive and which parts will die out) is dictated by reproductive success, just like in its biological counterpart. If for example, a feature of a language is too difficult to learn for the next generation of language users, this feature will simply die out. The users of the system dictate the survival of the elements, but at the same time the system itself dictates the playing field in which this survival game takes place. This kind of interplay between a system and certain elements of the system (in this case the users), is on a more generalized level the focus of a scientific discipline called complexity theory. The Norwegian social anthropologist Kim Sørenssen puts it this way:

> "One of the dramatic implications of complexity theory is that it adds a more general level of evolutionary theory over Darwinism: complex adaptive systems are capable – through spontaneous autopoiesis (self-organization) – of evolving into even higher degrees of complexity. This does not run counter to, but arches over, the Darwinian algorithm of competition for limited resources. 'The survival of the fittest' becomes a special case of the general tendency of complex patterns to self-organize, 'the survival of the most autopoietic patterns'. Real-life examples of complex adaptive systems include ecosystems, financial markets, bird flocks during flight (forming patterns), the Internet, large corporations, and language." [5]

Forgetting is an ingredient of the environmental pressure that shapes language. To illustrate this: irregular verbs are much fewer than regular verbs in number, but in frequency of usage, the irregular verbs are the most common. By their frequent usage, the population of users are constantly reminded of how to use these verbs correctly. Typically, irregular conjugation patterns in less frequent verbs will gradually be forgotten and replaced by a regular conjugation [6].An extreme example of this phenomenon is in the various Chinese languages. These languages have undergone an evolution of unparalleled continuity. As a result the conjugation of the verbs has disappeared altogether! Instead of using the tense of the verb to indicate when something happened, context (or an additional clause if needed) determines how to interpret the utterance.

Saussure distinguishes between a synchronic and a diachronic study of language. A synchronic study will take a snapshot, and look at the system at a given point in time. Here we find the study of present day grammar for example. A diachronic study will look at changes and the dynamics of the system, or how a certain state of the system historically came into being. Language, from this perspective, is a complex adaptive system, self-organizing in ways that are not possible to predict for any significant length of time. Just like we are not able to say what the weather will be like on a given day two hundred years from now, we are not able to say what the English language will be like a few hundred years from now.

Environmental pressure favors the evolution of general, pattern-detecting organs (brains) that are capable of structuring and classifying sensory data. Saussure's model of language shows us that the environment that these brains are exposed to, will determine which patterns these brains will detect and lock on to. That is, exposure to *le langue* triggers the brain to conduct a mapping process of the linguistic patterns it detects.

1.4 Memes Across the Bridge

The analogy between evolution in *le langue* and in ecosystems is not the only such analogy with relevance to linguistics. We need to consider yet another complex adaptive system, which likewise evolves through the dynamic interplay between the shared system and its individual users: culture.

Culture is in this context seen as the system of information that is distributed amongst a community of users. Pattern-detecting brains lock on to cultural patterns – just as they do with the patterns of *le langue* – and conduct a continuous mapping of these.

To illustrate this: let's think of some such communities of users that one could come across in the city of London. There are socially determined communities such as the

upper class, ethnically determined communities such as West Indians, or context-dependent communities such as those riding the underground. Notice that all these are geographically independent: all examples are from the same city. All individuals belong to several communities, some only for a fleeting moment, others lifelong. Whatever community you interact with, there are social codes to pay attention to. Everyone riding the underground, even the first time tourist, will detect and mostly respect the shared etiquette for tube travel. On the other hand, the intricacies of upper-class social etiquette are likely to expose virtually any nouveau-riche impostor. In this case the pattern-detecting brains of the insiders will detect irregularities in the behavior of the impostor.

In cultural systems, just as we saw in Saussure's concept of *le langue*, feedback from the collective modifications of the users, and the varying reproductive success of the many cultural elements, determine changes over time in the overall system.

There's an inter-disciplinary branch of science dedicated to the study of culture as analogous to biological ecosystems: memetics. The British geneticist Richard Dawkins coined the term 'meme'. In the words of Dawkins:

> "DNA replicators built survival machines for themselves – the bodies of living organisms including ourselves. As part of their equipment, bodies evolved on-board computers – brains. Brains evolved their capacity to communicate with other brains by means of language and cultural traditions. But the new milieu of cultural tradition opens up new possibilities for self-replicating entities. The new replicators are not DNA and they are not clay crystals. They are patterns of information that can thrive only in brains or artificially manufactured products of brains – books, computers, and so on. But given that brains, books and computers exist, these new replicators, which I called memes to distinguish them from genes, can propagate themselves from brain to brain....As they propagate they can change – mutate. It is manifested in the phenomena that we call cultural evolution. Cultural evolution is many orders of magnitude faster than DNA based evolution." [7]

Any user (e.g. members of a cultural community) will pass memes on to other users to varying degrees, depending on each meme's adaptation to its total environment. Any such evolving, self-organizing system – whether ecosystems, languages or cultures – depend on so-called replicators, or elements that can be reproduced with varying success depending on its degree of adaptation to the total system.

Language is intimately and dynamically interrelated with memetic systems. Continuing with our bridge metaphor, memes are what's being transferred across the bridge. Meme transfer is the purpose of language.

Suppose you, while reading this thesis, were asked what it is all about. If you answered: "On an abstract level, it's about building bridges", I would have successfully transferred my 'bridge metaphor meme' to you. And if it wasn't clear yet, I feel rather

confident that you, by reading this passage now, have received the meme that my idea of making a human-dolphin dialog protocol can be likened to designing a bridge.

The Power of the System

Language, and the spreading of memes are immensely important aspects of our knowledge. If we didn't have these, every human being would have to start from scratch and gain new knowledge without the help of previous insights.

Language then, may be the most important innovation in the history of mankind. In the words of Daniel Dennet:

> "Our brains are in effect joined together into a single cognitive system that dwarfs all others. They are joined by one of the innovations that has invaded our brains and no others: language. I am not making the foolish claim that all our brains are knit together by language into one gigantic mind, thinking its transnational thoughts, but rather that each individual human brain, thanks to its communicative links, is the beneficiary of the cognitive labors of the others in a way that gives it unprecedented powers. Naked animal brains are no match at all for the heavily armed and outfitted brains we carry in our heads." [8]

I find the last sentence very significant. Dennet is not referring to the brains themselves, but to the arms or concepts that animal brains do not carry. We have shifted focus from the apparatus as a carrier to the things being carried as the important thing. The system, *le langue* is part of the dress code. What does it take to be a carrier? My intention is to explore whether these communicative links could be extended to incorporate non-humans.

Genetically, there are no differences between a Modern man and a Stone Age man. The differences are found on another level. Memetically, we have evolved much more, because as Dawkins puts it: cultural evolution is many orders of magnitude faster than DNA-based evolution.

Suppose an unknown tribe were discovered, living in total isolation with no concept of the rest of the world and with a technological level equivalent to that of the Stone Age. If a new-born baby from this tribe were adopted and brought to New York City, it would grow up knowing just as much about computers and French fries, hip hop and fashion, as any other kid in the neighborhood. The child would have performed a quantum leap, bypassing endless steps of cultural evolution, just by entering into another cultural environment. Had the child grown up in its native tribe however, her brain would have been outfitted with Stone Age thoughts and knowledge instead. There is a kind of hardware – software issue here. Two brains that are genetically equivalent will perform radically different if they are linked to different networks.

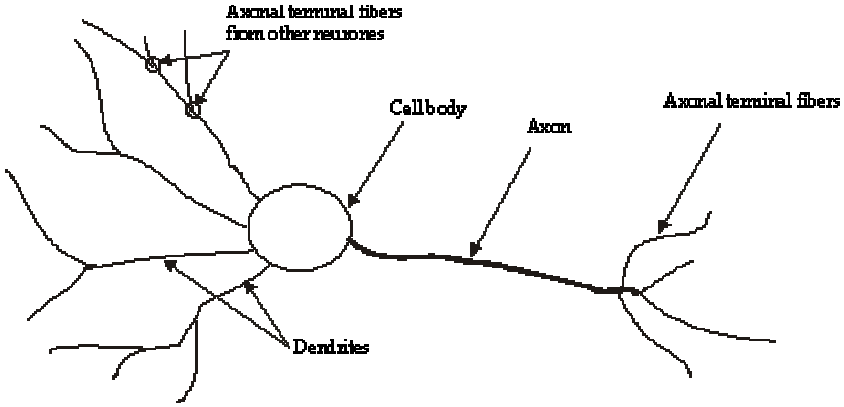
Another way to look at it, is considering the difference between inventing something, and learning how to use the invention. Anyone and his grandmother are using computers today, but no single person have enough knowledge to be able to build one from scratch, let alone invent one. It is the labors of many brains over long periods of time that has resulted in a product like a computer.

This is an important issue, so let's make yet another analogy. The Internet is also a complex adaptive system, analogous to *le langue*. We use it, among other things, to send and receive e-mail – a form of communication. Let's imagine, for the sake of the argument, that the Russians during the Cold War era developed their own style of computer – let's call it an RC – with an architecture different from the PC or Mac processors that are 'the brains' in Western computers. Let's also suppose that these RC's were not linked to the Internet. If someone claimed that the RC could not be used to send e-mail, they would be absolutely correct. But there would be two separate, potential reasons for why they could not be used for sending e-mails. The first would simply be that no one had ever bothered to make the interface necessary to connect them to the Internet. An Internet service provider (ISP) is an absolute necessity if you want to send e-mails. This would, in the case of the RC, be an obvious thing to try. The other possibility would be, that the architecture of the RC would be so different that it wouldn't work anyway. It was somehow designed so differently, that it would be necessary to restructure the whole thing in order to make it work, and it wouldn't be worth the effort. If a group of researchers were told to look into the matter, they would no doubt see the duality of the problem, and consider the processor architecture as well as the ISP issue.

After having separated the carrier of knowledge (a brain) and the source of the knowledge (a collective system), the natural question becomes: "What prevents dolphins from taking part in this collective cognitive system?" In the next chapter, we will investigate how brains work and look for similarities in processor architecture. The conclusion will be that they are similar enough to try, and to look at the ISP end of the problem. The Human-Dolphin Dialog Protocol is an ISP solution to the problem.

– 2 –
Brains and Neural Networks: Pattern-matching Devices

A brain consists of a large number of brain cells called neurons, connected to each other in a network. Here's a sketch showing the basics.



The cell body is a simple processing unit that is hooked up to thousands of other neurons with 'electric wires' called dendrites. Through electric impulses, they deliver input information from other neurons to the cell body . If the stimulation exceeds a certain threshold, the cell body will fire an electric impulse through the axon to other neurons. If the incoming stimulation is too low, the impulse will not be transported any further. This simple construction is able of storing and learning an amazing number of things. Of course, this is a gross simplification of what's really going on in a brain. The individual neurons are very complicated with a myriad of parts, sub-systems and control mechanisms. There are many different kinds of neurons, and the information being passed between them is being conveyed via a host of electrochemical pathways. But this will only cloud the basic mechanics involved. The key issue here is that many 'dumb' units connected together in a network somehow manages to store and process information, and to learn. The name of the theoretical study of this is *connectionism*.

2.1 The Functionality of a Brain

The traditional way of viewing the functionality of a brain, is that brains have evolved layer by layer, adding more and more neurons together, and local patches of neurons have specialized in processing certain kinds of information. Some parts of a brain are used for motor control of body parts, and other parts are used for processing sensory input. The nerve endings from the eyes for example, will be directed to a specific cluster of neurons that will process the data, take it apart and pass information about colors, contours and movements on to other further specialized regions of cortex, which in turn will send out a result of some kind, to yet another cluster. Exactly how the information flows in the brain is still uncharted territory, but a lot of work is being done. PET scans, MRI and other modern brain imaging techniques have shown this activity in impressive photos. Certain cognitive experiences will activate certain areas of the brain, and this will show up as colored patches on a screen. Better equipment has provided a higher magnification of the brain, as subtler and subtler cognitive aspects are being sought to map.

By performing brain surgery under local anesthesia researchers have, with the help of a probe and a weak electrical current, been able to ask patients to name objects presented on slides in front of them. When specific parts of the brain get probed, the patients are no longer able to name them. Similarly, scientists have been able to activate body parts or facial muscles and thereby in detail map which parts of the brain will trigger certain anatomical parts of the body. They have found great uniformity between human brains on a certain level of cognition. The visual cortex is found in the back, Broca's area (associated with language expression) just above the left ear, and so on. Vision, hearing, reading, understanding of spoken language, motor control and so on, are all mapped and named areas of a human brain.

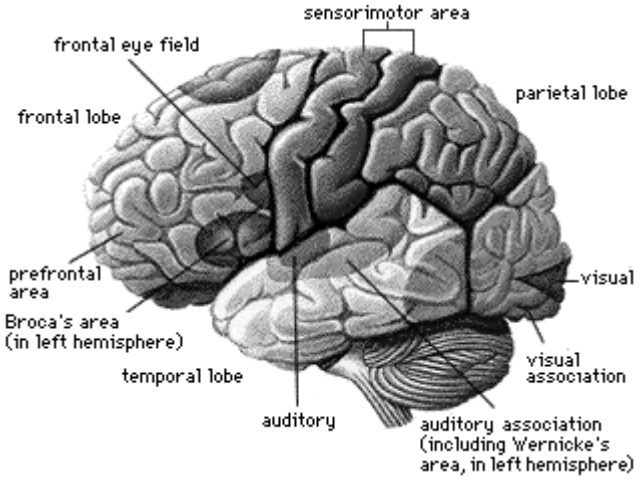


Figure 5. Traditional view of the brain with its specialized areas.

2.2 Plasticity of the Brain

But this uniformity only holds true to a certain extent. Although it has been shown that there are specific language areas in human adult brains, they do not exist from birth. They are structured along the way [9]. Young brains are astonishingly plastic. Hemispherectomy of the left brain (surgical removal of the entire hemisphere of the brain, sometimes used as a cure for uncontrollable epilepsy), has resulted in severe damage in both generation and comprehension of language in adults. This is because the language areas of the brain are removed. However, when this is done on young children, they have been able to learn language using the remaining, right hemisphere. The earlier in life the surgery is done, the better the recovery of the language ability. This is an indication that perhaps the fixed positions of specialized organs are an illusion. A brain's plasticity goes even further, as shown by several (in my opinion horrible), experiments on animals: [10]

- Reduction of the size of thalamic input to a region of cortex early in life, determines the subsequent size of that region (Dehay et al., 1989;O'Leary, 1993; Rakic, 1988).

- When thalamic inputs are 'rewired' such that they project to a different region of cortex from normal, the new recipient region develops some of the properties of the normal target tissue - e.g., auditory cortex takes on visual representations (Sur, Garraghty, & Roe, 1988; Sur et al., 1990).

- When a piece of cortex is transplanted to a new location, it develops projections characteristics of its new location, rather than its developmental origin - e.g., transplanted visual cortex takes on the representations that are appropriate for somatosensory input (O'Leary & Stanfield, 1989).

- When the usual sites for some higher cortical functions are bilaterally removed in infancy, (i.e., the temporal region with primary responsibility for visual object recognition in monkeys), regions at a considerable distance from the original site can take over the displaced function - e.g., the parietal regions that are usually responsible for detection of motion and orientation in space (Webster, Bachevalier, & Unger-leider, in press).

Windows of Opportunity

It turns out that most sensory systems do not develop properly unless there is appropriate exposure during a particular phase of development, what is called a 'critical period'. Critical periods can be seen as windows in the development, in which specific learning *has* to take place, if at all.

For a couple of hours every day, newborn kittens (!) were confined to a room consisting of only vertical black and white lines. The rest of their day was spent in darkness. After a certain period, the neurons in the visual centers were probed, and it turned out that the great majority responded only to vertical lines. Subsequent trials showed that the kittens were unable to navigate in rooms with only horizontal objects. The kittens clearly did not see them. By stimulating only certain types of neurons

during the critical period, It was demonstrated an expansion of these at the expense of neurons responding to horizontal lines. [11]

The well-known saying "use it or loose it" is quite telling in these cases. There seems to be a kind of internal competition, deciding how various cortical areas will be utilized. Some functions seem to - crystallize after an initial tune-up period of the brain, and will loose their flexibility once this window of opportunity closes. Other localized functions will uphold their flexibility throughout a person's life. Exercising only the middle finger for example, will enlarge its territory on the cortical map, and also rearrange all other fingers as well as the thumb-face and hand-wrist borders. Mapping the sensory strip in blind people who have learned to read Braille, shows that their finger areas are larger than average. If an arm is lost, its space in the sensory strip seems to be taken over entirely by the lower part of the face and so on. [12]

Genie

So is there a critical time for language acquisition? We have moral restrictions that prevent us from making certain experiments with human children. We can not lock them up in rooms with only vertical lines, or rewire their brains to see what happens, but there are some cases where mentally ill parents have locked their children up, and where social workers have found them years later. The most famous example is a girl called Genie, who had been kept in isolation from human company, harnessed to an infant's potty seat by day, put to bed in a straitjacket-like sleeping bag at night, and beaten if she made any sound. She was found at the age of 13 and efforts were made to teach her to understand language. Although the literature on the topic varies, it seems Genie did learn some non-trivial aspects of language, even if it was far from perfect. The emotional scars of such an appalling childhood might, however, interfere with learning to such a degree that its effects on her limited language acquisition must be considered. [13] We will however, see in section 4.2 that language acquisition in pygmy chimps was greatly improved when they were exposed to language at an early age.

Phonetic Maps

Although there are no distinct borders between the different sounds that make up a language, sounds will be classified by the speakers into clear categories. Studies of language acquisition in children of different languages have shown that at some point in early childhood, the child creates a map of the available phonemes in their parent's language scheme. Every language has its own set of phonemes. In Japanese, there is a phoneme somewhere between the English /r/ and /l/. A Japanese person who has not been exposed to other languages while growing up, will as a result have difficulties distinguishing between /l/ and /r/, simply because they don't exist as different entities in the Japanese language. The same is true about English speakers having difficulties with certain Hindi or Portuguese phonemes.

Experience Shapes the Brain

It seems as though brains to a large extent are made up of 'general purpose' neurons and connections that can take on any cognitive task, participate in any specialized region and adjust its behavior accordingly. New findings such as the ferret experiments showing that the auditory cortex can reorganize to handle visual input, implies a whole lot more plasticity in early development than traditionally expected, and contradict popular theories on how brains consist of specialized regions for seeing, hearing, sensing touch and, in humans, generating and understanding language. Dr. Kaas, professor of psychology at Vanderbilt University in Nashville puts it this way:

> "The cortex can develop in all sorts of directions. It's just waiting for signals from the environment and will wire itself according to the input it gets. Genes create a basic scaffold, but not much structure. Thus, in a normal human brain, the optic nerve is an inborn scaffold connected to the primary visual area. But it is only after images pour into this area from the outside world that it becomes the seeing part of the brain. As the inputs arrive, the cells organize themselves into circuits and functional regions. As these circuits grow larger and more complex, they become less malleable and, probably with the help of changes in neurochemistry, become stabilized. This is why a mature brain is less able to recover from injury than a very young brain."

To illustrate this: people who suffer from a stroke in a special area, will loose the ability to read. The words are not blurred; they can name the letters, but not the words [12]. Writing was only invented some 5,000 years ago, and very few people have been literate until the last few centuries. Millions of people as still illiterate today. So how did we get reading areas so quickly? Hereditary tendencies do not develop in such a short time. Genetic alterations by natural selection take time, a long time. People have reading areas in their brains because they are being exposed to books. We can suspect that people have language areas in their brains because they are being exposed to *le langue.*

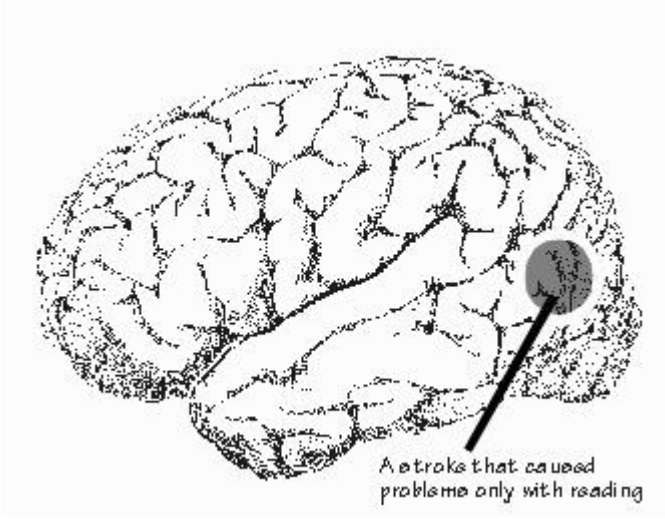


Figure 6. A stroke in one particular area of the brain caused problems only with reading.

The uniformity of the architecture of human brains must be seen in the light of the uniformity in the environment that these brains are exposed to. This does not hold only for human brains, but as we have seen, for brains in general.

This ability of brains to adapt to their environment does not come as a great surprise to people who have worked with artificial neural networks.

2.3 Artificial Neural Networks

The design of an artificial neural network (ANN) is based on what's been found in neuroscience about the brain, and the idea that a system only made up of connections with variable and changing strengths could be able to store information, and to learn. A computer program can not send an electric impulse between its various program lines, so this, among other things, is simulated. We let an electric impulse be represented by a number between –1 and 1. Then we let the computer make some abstract units that can hold a variable, connect these units together with weights that can vary in strength, pass some numbers between them and have the computer calculate the result. The resemblance with brains is thus not physical, only simulated.

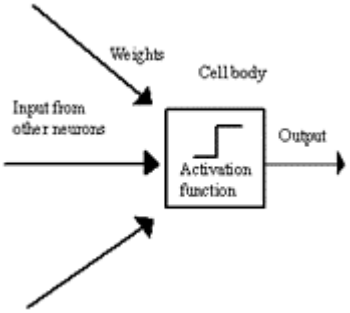


Figure 7. An artificial neuron is conceptually similar to a biological neuron.

There are many kinds of ANN's, but the idea is basically the same in all of them. Neural networks offer a different way to analyze data, and to recognize patterns within that data, than traditional computing methods do. It should be pointed out that ANN's are not the answer to all computing problems. Traditional, rule-based computing methods work well for problems that can be formalized, like balancing checkbooks, searching through text, or performing complex math. Basically, neural network applications are used with problems that are hard to formalize, for example the ability to process real world inputs. In principle, sensory data are all distinct and never repeated. Problems such as classification, data association, data conceptualization, data filtering and prediction are the domain of ANN's. The processing of sensory data can be seen as a process of abstraction. This is what brains do, and this is also a domain for ANN's.

With the same general ANN architecture, researchers have made artificial neural networks perform a myriad of different tasks. They have taught ANN's how to recognize faces from pictures, recognize speech, recognize hand-written characters, compose music, determine illnesses from analyzing patient's breath, dock a space shuttle to a space station, or steer a car, to mention but a few. They have also taught ANN's many language-specific tasks. They are used for recognizing patterns and generalize from these patterns.

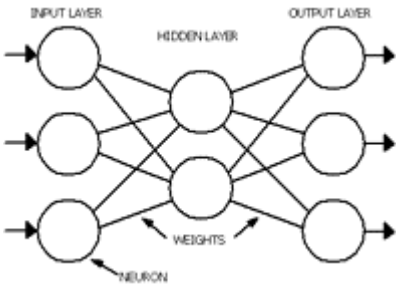


Figure 8. A small three-layer, feed forward network.

ANN's are not given any rules or other forms of explicit information on the topic that they are supposed to learn. They just receive training data and adjust their weights according to the same algorithm, whether the patterns they are about to learn to recognize are pictures, text, speech, music or anything else.

Let's take a simple network like the one in Figure 8 as an example. A feed-forward, back-propagation network consists only of nodes and weights. The nodes in the input layer get its values from the data that should be classified, for example an array of waveform features. Each of the nodes in the hidden layer are connected by weights to the nodes in the input layer, and will get its values by adding up the value in each input node multiplied with its corresponding weight. The nodes in the output layer will get its values by doing the same operation with the values now stored in the hidden nodes, and the weights between the hidden and the output nodes. So we send in a bunch of numbers, and we get a bunch of new numbers out. What's the deal? Assign a category to each output node, and ideally, when the input array of a given type enters the network, the output node with that category will get the highest value, and hence classify the input token as belonging to that class. Initially the weights are given random numbers and will of course not classify correctly right away. In order for this to work, training data is needed where we know what output category the input data belongs to, and the weights are adjusted so that the calculations described above result in the desired output. This is called to train the network, and the back-propagation algorithm is used for doing that. The weights are not adjusted dramatically. The output is compared with the desired output, some weights get punished (their value reduced), and some weights get rewarded. A new array is sent in and the process is repeated. Eventually, the weights in the network get their desired values and unknown data can be fed to the network and be classified.

ANN's will either learn in a supervised or in an unsupervised fashion. Supervised learning means you know in advance what kind of categories you want, and the training data is labeled, as in the example above. Unsupervised learning, as is the case for Kohonen networks, are used in cases where you want the network to find patterns or categories in unstructured data.

No neural network has learnt language in human-like manner, but there are many examples of ANN's having captured aspects of language. Below follows a few examples where ANN's have successfully managed to capture the underlying patterns of some language-specific tasks.

Pronunciation (Net-Talk)

Sejnowski and Rosenburg taught an ANN how to pronounce English correctly from written text. It used 1,000 sample words with the instructions of how they were supposed to be pronounced, and after training the network one could give it words that were not among the samples, which would be pronounced correctly. The network learned the difference between soft and hard c's (city/cat), the different a's (save/ball/have) and so on. When mistakes were made, they would resemble mistakes made by children, or non-native speakers, and mistakes made from not taking context into account. (E.g., 'lead' was pronounced differently depending on whether it's a noun or a verb). [14]

Syntax/Semantics (Websom)

Kohonen's Self-Organizing Map (SOM) is one of the most popular *unsupervised* artificial neural network algorithms. It has, among other things, been used to create 'word category maps'. In a study by Honkela, Pulkki, and Kohonen, the input for such a map was the English translation of fairy tales by the Grimm brothers. The purpose was to see what happens if you classify words based on contextual information only. The network was given a triplet of words as input, where the middle word was to be classified. The word in front and after served as context. No syntactic or semantic categorization was given. Yet the overall organization of the map reflected both the syntactic and semantic categorization of the words. The verbs formed an area of their own in the top of the map, whereas the nouns could be found in the opposite corner. Modal verbs formed a collection of their own among the verbs. Connected to the area of the nouns were the pronouns. A distinct group of possessives could be found, and so on. Inside the large syntactically based groups of the map, fine structures of semantic relationships could also be discerned. Inside the group of nouns for example, animate and inanimate nouns formed separate groups. Sets of closely related pairs could also be found. Father/mother shared the same node, so did day/night, child/son, woman/man, head/eyes, forest/tree and so on. This phenomenon can be explained statistically. The context of a word is, quite naturally, dependent on the syntactical restrictions that govern the positions of the words in the text. The presence of various categories in the context is only statistical.

> "The experimentally found linguistic categories, being determined only implicitly by the self-organizing map, seem to correlate with the categories of concepts occurring in actual cognitive processes. It may also be argued realistically that, in human language learning, naming of explicit syntactic relations is not needed. Expressions heard in the proper context may be sufficient for creating a working model of language." [15]

Past Tense Acquisition in Norwegian

English verbs are divided into two groups: regular (or *weak*) and irregular (or *strong*). In English, past tense of weak verbs is formed by adding the suffix –ed to the stem, as in jump – jumped. Strong verbs are all the verbs that do not follow this rule. The Norwegian past tense system is similar to the English one in that the majority of Norwegian verbs is regular and form the past tense by suffixation. Norwegian regular verbs are however, divided into two subgroups, often referred to as the *larger weak* (WL), and the *smaller weak* (WS).

Although the majority of verbs in both languages are regular, the most *frequent* verbs are irregular (see section 1.3 for a discussion on this). Past tense acquisition in children goes through several stages. The first verbs the child learns (the most common ones) are irregular, and the child learns to correctly conjugate these verbs. At a later stage, when the child has learned more verbs including regular ones, a phenomenon known as *over-generalization* occurs. What is seen both in English and Norwegian is that when children make over-generalization errors, i.e., when they apply the wrong pattern to an existing verb, they almost always over-generalize the regular pattern (e.g. run-runned). In Norwegian however, over-generalization becomes more complex because of the two different groups of regular (weak) verbs. Studies of children have shown that at one stage of development (4-6 years old), over-generalization happens mostly from the WL class. At a later stage (6-10 years old), patterns of the WS class are predominantly over-generalized.

A back-propagation neural network with 100 input nodes, 60 hidden nodes, and 101 output nodes was created by Anders Nøklestad, to see if it could acquire the Norwegian past tense system and to investigate the phenomenon of over-generalization. The model was trained on phonological representations of 1,709 Norwegian verbs taken from a frequency list. The present tense form of the verb was given as input and the network was trained to produce the correct past tense. When tested, the ANN answered correctly 94% of the time (equivalent to adults under time pressure).

Trying to simulate the vocabulary growth of a child, the model was first trained to a hundred percent correct performance on the 20 most frequent verbs, and then the size of the vocabulary was gradually increased. The phenomena of over-generalization occurred in the network as well. Both types of over-generalization were observed, at different stages of the training. Graphs of children's over-generalizations

superimposed on graphs of the network's performance at different stages, reveal an impressive correlation as to which kinds of errors were done. [16]

Similarities with brains

Besides the obvious structural similarities – both brains and ANN's are distributed networks – there are some interesting functional similarities between ANN's and brains, which we don't find in regular rule-based systems.

- ANN's are able to generalize, and thereby capable of dealing with new and unfamiliar input (within their field of expertise).

- ANN's are insensitive to localized damage.
  Up to 10% of an ANN can sometimes be removed without considerable reduction in performance, regardless of which part is removed. This is because *the information is being distributed.*

- ANN's are fault-tolerant. That is, they can produce a reasonably correct output from noisy and incomplete data.

In addition, the autodidactic (self-learning) and pattern-matching capabilities of neural networks seem to be crude versions of similar, but infinitely more advanced processes in the human brain.

Impressive as they might seem, the neural networks we are able to build in a computer today, are pathetically small and simple compared to any mammalian brain. A large ANN will consist of maybe 2,000 neurons. The latest estimates on a human brain are approximately 200 billion neurons all in all, and about 30 billion in the cerebral cortex alone.

Again it must be stressed that the comparison between neural networks and brains is only valid as an abstraction, and that there's no resemblance in a real sense. Making a strict numerical comparison between the two in therefore not appropriate. It is only made as an indication of the complexities of real brains. Is such a numerical comparison between brains relevant?

2.4 Does Size Make a Difference?

There seems to be an intuitive correlation between brain size and intelligence. More neurons, bigger brains – more processing power, more intelligence. But this is only half the story, and a half-truth, befuddled with paradoxes and criticism. There is no correlation found between brain size and intelligence in humans. Some very smart people had relatively small brains, and vice versa. Elephants and some cetaceans

(whales and dolphins) have bigger brains than humans do. Seeing that these animals are very big, it has been argued that it seems reasonable that the larger the body, the larger the brain needed to operate it. To get a correct measurement of a species' intelligence, some say we must look at the brain/body ratio. This places humans well above both whales and elephants, but with a closer look, if we are to believe that the brain/body ratio will give us a correct estimate, a small African rodent will outsmart us all. Further adjustments are needed in order to get humans on top of the list. The math that does the trick is called the Encephalization quotient (EQ). It is a measure of brain size normalized to body size in such a way, that an EQ of 1 implies a brain size consistent with the average of relevant comparison animals of the same size. Humans have an EQ of 7. EQ is the standard measurement used today, but is befuddled with criticism for a number of reasons. Since body weight is part of the equation, one supposedly gets more stupid if one gains a few kilos. Cows, being grass eaters, for example, have disproportionally large digestion systems, and get a lower EQ score for that reason. Body weight follows great seasonal variation in many animals, and varies a lot between males and females. There seems to be some common sense argumentation that puts the brain/body ratio calculations in a questionable light.

Sperm whales are basically big blobs of fat (blubber). Why would it require a big part of the brain for housekeeping this? Fat pretty much takes care of itself, as far as we know. They do not need complex motor control to carry their bodies around, as a monkey swinging from tree to tree needs. Still the sperm whale has the largest brain known by all animals. It can weigh up to 9,000 g compared to a human brain, averaging 1,300-1,400 g. If you must have a big brain in order to household a big body, we could not have had big dinosaurs with brains the size of walnuts. What the whale is doing with its big brain we don't know, but it does not seem reasonable that it is for housekeeping purposes only.

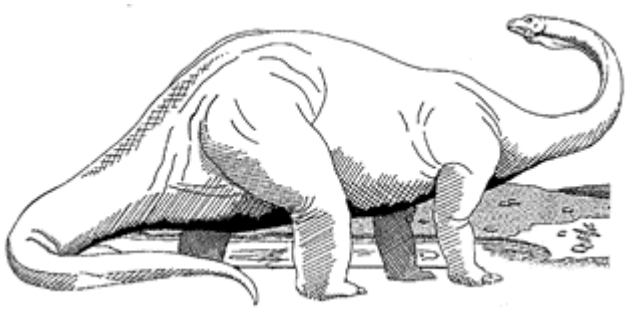


Figure 9. Apatosaurus: A small brain in a big body, contradicting the notion that one needs a big brain to manage a big body.

Another indication of intelligence has been said to be how convoluted (physically folded) the brain is. Brains do most of their information processing on the surface. So species that, in the course of evolution, have come to need more brainpower, have developed brains with creased, convoluted cortexes that pack more surface area into the same volume of skull. With this point of view, surface area is more important than weight. The question of whether an absolute measurement or a measurement relative

to body size or weight should be applied, is relevant here as well, since the reason for involving body weight is that a bigger body is said to need more processing power. There is however, no consensus reached on how to include surface area in the calculation.

The Dolphin Brain

Bottlenosed dolphins have big brains. Slightly bigger than the size of humans, they average 1,600 g. Their EQ is 4,5 (the highest of any non-human animal, twice that of higher primates). The bottlenosed dolphin's brain is also more convoluted (physically more folded) than that of any other mammal, including humans. So if any known measure of brain size is a relevant factor, dolphins are well equipped.

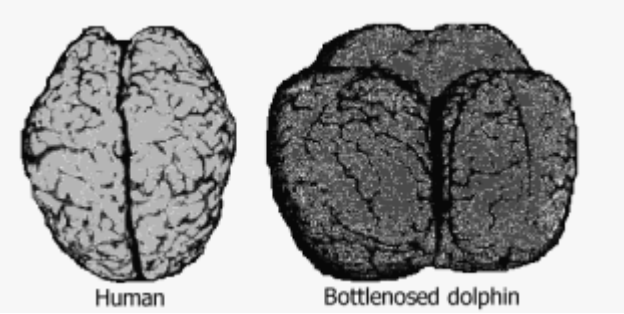


Figure 10. If brain size alone would decide, dolphins would be better equipped than humans.

As the previous passage has indicated, we should be careful about ascribing too many qualities from brain size alone, absolute or relative. The correlation of neural density, the number of connections between neurons, surface area of the cortex, and the internal structure of the neural network, is all very complex, and not well enough understood to make bombastic conclusions in any direction. With an analogy to computers: not only are computers more powerful today, with a fraction of the size of computers 20 years ago, but improved software is also pointing out that performance is dependent on how the information is structured. We can only conclude that cognitive abilities cannot be described by physical measurements alone. Are there other ways to look at intelligence?

– 3 –
Creative Analogies

The cognitive scientist Douglas Hofstadter has written extensively on the processes that characterize human cognition, and how AI could be used in an attempt to model these particular processes. Such mimicry of brain processes is Hofstadter's recommended strategy towards successful AI endeavors, based on the premise that any intelligence, human or non-human, will necessarily display some of the same essential characteristics, such as pattern-matching. More importantly, he has led numerous pattern-matching AI projects in order to better understand cognition and intelligence in general.

According to Hofstadter, the very core of intelligence is pattern-matching, a process which in turn is based on the ability to make analogies: "…analogy-making lies at the heart of pattern perception and extrapolation […] put together with my earlier claim that pattern-finding is the core of intelligence, the implication is clear: analogy-making is at the heart of intelligence." [17]

Further pursuing the rationale that artificial intelligence would have to possess some analogy-making ability in order to qualify as intelligent, we could infer that the same would likely have to be true for animal intelligences as well. In the words of Hofstadter:

> "I believe that intelligent creatures, wherever they are, when faced with complex situations, will see similar things in them – if they're really intelligent. Life is such that we have a certain way of filtering situations and finding what's at their core. Good intelligences, the kind that really survive well, will probably all be very similar to one another in their ability to do this." [18].

Tabletop

Hofstader's 'FARG' group has developed the computer program Tabletop, an effort towards AI analogy-making within a very restricted domain: two persons, Henry and Eliza, at each side of a coffee-shop table, with a similar (but not identical!) lay-out of coffee cups, saucers, utensils etc. in front of each. Henry, the challenger (representing the problem to be solved), starts the game, pointing at one object: "Can you do this?" Eliza (representing the analogy-making artificial intelligence) points to the

corresponding object on her side of the table, and has solved the challenge successfully. True analogy-making starts when Henry points at an object in front of him, which Eliza does not have in her array, or is in a different relative position. "Can you do this?" She will then have to choose a different object than Henry, which can be thought of as being somehow analogous through a process of abstraction: for example by also being to the left of the fork, or by having a similar position in the total pattern of objects. The more different their respective arrays of objects are, the higher-level abstraction is needed to make a Tabletop analogy. [17]

I have experienced a series of interactions with a dolphin, which at the time seemed to be a simple game similar to Tabletop, with analogous – but not identical – solutions to the challenge: "Can you do this?"

3.1 Maui's Game

In the summer of 2001, I spent five weeks in Hawaii, working with Dr. Ken Marten at Earthtrust, a non-profit organization devoted to cetacean research and wildlife protection. Earthtrust has a dolphin observation laboratory located in Sea Life Park Hawaii. It has sound recording, video and technical equipment with large underwater observation windows, used in behavioral psychology observation studies, and recently also language studies of the dolphins. I assisted Dr. Marten with analyzing a large amount of sound/video recordings in order to look for 'signature whistles' of the dolphins currently staying in the pool. The term 'signature whistles' derives from a theory postulating that dolphins have unique individual whistles used as an identification mark, i.e. a name of sorts. For me it was a good opportunity to do some fieldwork, collecting data for this thesis and at the same time assist Earthtrust in their efforts to learn more about dolphins. The pool with the observation windows is large enough so that the dolphins are not forced to face the windows. Most of the time they swim around, interact with each other or play with toys in the pool. Occasionally however, they will swim over to a window to 'check out the humans' and see if there is anything fun going on over there. As with children, it is quite difficult to hold their attention for a long time. They will quickly get bored. A recurring topic for me whenever a dolphin would visit the window, would be to try to entertain them in ways that would make them stay longer.

The story I'm going to tell is not intended to be used as evidence in any way, but as an example of how interaction with dolphins often show indications of playfulness, humor, curiosity and analogy-making of the sort Hofstadter describes as the hallmark of generic intelligence. It is also an example of how important language is for assessing the intentions of another being, since it can be argued that I have no way of knowing how well my version of the story correlates with reality.

On one occasion, one of the residents of the pool, a 10-years old male, captivity-born dolphin called Maui came over to the window, bending his head down and gently bumping the top of his head on the window as a greeting to call for my attention. As I had done many times before, I walked up and brushed the window, caressing him through the glass. The dolphins' hearing is much more sensitive than ours, and the movement of the hand against the glass makes a sound that apparently is mildly interesting to them. When moving the hand over the glass, they often follow the movement of the hand with their beak. This time however, I stopped my hand abruptly at a certain point and looked at him motionless. Maui also floated still for a moment, and when he opened his mouth, I responded to his movement with a quick move up and down with my hand. He noticed the action, and as if to test if there was a correlation, opened his mouth again. I responded in the same way and we did this a few times. I stood motionless with my hand on the glass holding eye contact, waiting for a cue. Maui opened and closed his mouth and I tried to respond as quickly as possible with an analogy-making movement of my hand. After some rounds, Maui grew bored and left the window. The essence of the interaction seemed similar to Hofstadter's Tabletop analogy-making challenges: "Can you do this?"

Maui returned to the window only few minutes later and placed himself in the same position, as to invite for another game. I complied of course, and we went through another set of rounds. This time didn't last as long as the previous. It seemed to be just a confirmation of the unspoken rules we had established.

That first day Maui came back a third time and this is when the first twist occurred. Same procedure as last time for two rounds: Motionless waiting, he opened and closed his mouth; I moved my hand up and down. But then, the third time he tricked me! He moved his head a tiny bit and made me believe he was going to 'strike'. I quickly moved my hand and discovered he had not opened his mouth. I had been deceived. He bent his head, touched the window with the top of his head and pushed back, drifting slowly backwards a few meters and then stopped and looked at me for a moment, before he swam away, as if saying "Gotcha"! He had introduced a new aspect of deception to the game. I was impressed.

The next day Maui came up to the window almost immediately after I had entered the lab. He greeted me with the same bumping of the window with the top of his head and placed himself in position for another game. The first two or three rounds went just like the previous day, but then something unexpected happened again. After having stood motionless looking at each other for a moment, Maui suddenly opened and closed his mouth very quickly, a series of times. His moves were so quick I had no way of following up with a corresponding number of moves with my hand. Maui repeated the 'gotcha' maneuver and swam away. Set, game and match Maui again! The following days we continued the game but Maui always took the 'lots of opening the mouth in a quick succession' approach which I could not match. I felt almost as if he

was cheating. He had introduced an aspect to the game that made it impossible for me to follow. Since the game didn't evolve further, we eventually stopped playing.

3.2 The Need for Language

Of course, I can't really know the validity of my interpretation of what went on in Maui's mind. But if I'm right, Maui displayed both a *creative analogy* (beak movement is analogous to hand movement) as well as a *fluid concept* (sometimes real beak movements, sometimes deceptive trick ones). Hofstadter's book on analogy as the core of intelligence is, incidentally, called 'Fluid Concepts and Creative Analogies'. If this were indeed what Maui displayed, it would indicate the presence of intelligence in Hofstadter's terms.

Another way of looking at it is with the Fregian gaps and layers in mind (section 1.2). Are we dealing with cognitive creatures? Analogy making is in the cognitive realm, and acting with the intent to deceive is perhaps in the praxis realm. Whether a case like Maui's game was an example of analogy making or simply a series of unrelated movements, is a natural topic for discussions. The epistemic gap (classification and ordering of sensory data) is bridged by all kinds of animals, but some would claim that very few creatures (only humans?) are able to bridge the cognitive gap and the praxis gap. This cannot be solved by philosophy alone, but empirical experiments are needed in order to test this.

The way to investigate this is to make a leap of expectation, and assume that they are cognitive creatures and act accordingly. The way to investigate whether someone can speak Chinese or not, is to expect that they can speak the language, address them in Chinese, and look for their response. If the investigator knows the language, there is very little room for misinterpretations. It is hard to fake a language ability if you are talking to someone who knows the language. By just asking someone in English "Can you speak Chinese?" the person could simply lie (either way) and the investigator would draw the wrong conclusions. The measuring of brain size or other organs will not be enough to conclude anything either. Whether its language, the knowledge of algebra, or any other cognitive trait, it is best investigated by assuming its presence and acting accordingly. By inviting for a game of analogy making, I expected Maui to be able to respond to it – or the event would never have occurred. What prevents me from assessing whether my version of this interaction is similar to Maui's, is the lack of accessible representations of Maui's intentions. This illustrates well the function of language: to enable one to understand another's intentions by representing these.

Where possible misinterpretations of the data is great, there will naturally be more objections to the way the data was interpreted. The goal of any such experiments would be to minimize the possibility for misinterpretations. The Human-dolphin

dialog protocol (Hddp) and GAPR, the software introduced in chapter 7, is a way to minimize the possibility for such misinterpretations.

Before we get to that there are other requirements for language acquisition that are still not discussed. That is the topic for the next chapter.

– 4 –
Further Requirements for Language Acquisition.

We have so far identified two important requirements for language acquisition:

- Access to a linguistic system (i.e. *le langue*)
- A way to store and process at least a subset of this system (i.e. a brain)

4.1 Interface: Input/Output Devices

As mentioned previously, one of the core elements in language is the concept of symbolic representation. It is a way to externalize our inner thoughts by finding and agreeing on an external representation for an internal concept/idea. We have also discussed the arbitrariness in the link between the two. It must be stressed again that any agreed upon representation will do. Human languages have in their original form used sound, and later adaptations use sign language, Braille or letters. So one very important requirement would be the ability to produce and interpret the kind of symbols that the language is built up around. Human language has evolved in a direction favoring sounds that are easy for humans to make. This will certainly be a disadvantage for other species with a different physiological structure than ours. For a human language, physiological requirements are a vocal tract and ears to start with. The human vocal tract is different from our nearest cousins the chimpanzees in ways that makes it impossible for the chimp to produce the many sound distinctions that we do. Not only is the physiological existence of a vocal tract (or what ever performs the symbolic manifestation) needed, but also voluntary motor control of that apparatus, i.e., a link between the apparatus, the part of the brain that controls its movements, and the part of the brain that is associated with volition of some form.

Voluntary Motor Control of Speech Organs

Although our hearts move (or we would not be alive), it is an organ that most people have no voluntary control over. A Morse code language based on heartbeat, would

therefore be impossible. It would admittedly be a very impractical language, but the point here is not concerned with the impracticalities of such a language. The point is that several movements that our bodies are doing, and that theoretically could be used as an interface between our internal thoughts and external communication partners, is outside the realm of our volition. The organ that we use as an interface for communication – the larynx – appears to be outside the realm of volition for many other animals. It might seem like a minor requirement compared to the cognitive aspects of language, but voluntary motor control is a prerequisite for language much in the same way that hands are a prerequisite for making gadgets. The human ability to create gadgets that help us manipulate our environment is dependent on intelligence and hands. Gadget-making is, apart from language, one of the main things that has put the human race in such an outstanding position on the planet today. Without our hands there would be no space ships, French fries, books, or computers. We usually put more emphasis on the intelligence part, but without our hands (or perhaps our opposable thumbs), we could be as intelligent as we'd like, and nothing would come out of it.

Dolphins have no hands, and therefore are lousy gadget makers, but they have evolved several features that make them very well-adapted for linguistic communication. Deacon mentions in 'The Symbolic Species' [2]:

> "In addition to our own species, there is one group of mammals that exhibits some significant degree of vocal flexibility and learning ability: the cetaceans (whales and dolphins). In many ways, they are the exception that demonstrates the rule of visceral versus skeletal motor control. Though much of cetacean sound production remains poorly understood, it is generally believed that the many sounds they are capable of emitting probably are not produced by the larynx, instead, dolphins and whales appear to generate squeaks, clicks, and whistles within an elaborate system of sinuses that are located in the front of the skull and that feed into the blowholes on the top of the head. This is probably controlled by passing air through constrictions between the sinuses tightened by the contraction of underlying muscles. These blowhole muscles correspond to face muscles in other mammals, and are almost certainly controlled by the skeletal motor nuclei of the brain stem." [p 241]

Incompatibility

Although dolphins have voluntary motor control of their sound production, the sounds they make are incompatible with human linguistic sounds. It is not possible to use the human organs to make dolphin sounds, and it is not possible to use the dolphin organs to make human sounds. This is an interface problem that becomes a

key issue in the design of the communications bridge. The General Audio Pattern Recognizer (GAPR) - the software that bridges this gap - as well as more details on dolphin sounds, will be introduced towards the end of this thesis.

Input Device

Ears - or something equivalent that can be used to interpret the signals on the receiving part of the bridge, is of course also necessary. Dolphins are highly acoustic animals - it is their primary sensory modality – with hearing far superior to humans, so it will not be the weakest link in the bridge. The hearing range of the bottlenosed dolphin is 20 – 200,000 Hz – ten times that of humans.

4.2 Communication vs. Language

With three requirements in place, it is worth noticing the inter-dependency between them. Lose one, and the whole thing collapses. We must also note that they are not all-or-nothing criteria. It is possible to have graded versions of each one of them. For example, a huge system, a huge brain and a lousy interface. Will any combination of these three requirements, constitute language? A rudimentary grunting by a small-brained animal understood by a receiver as "GO AWAY", is that language?

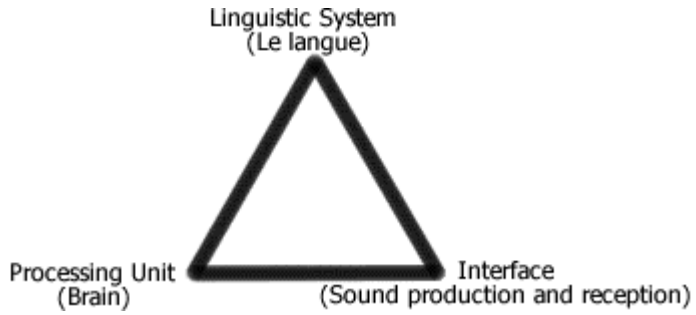


Figure 11. Three equally important requirements for language acquisition. Access to a linguistic system; a way to store and process a subset of this system (e.g., a brain); and input/output devices (e.g., ears and voluntary motor control of a sound-producing organ).

Mind-Reading

Most animals give off signals that correlate with actions about to take place, like a posture indicating attack, or hesitation indicating fear. The ability to use those cues is called mind-reading. Many animals use this so that they can predict what other individuals are going to do. When it is advantageous to be mind-read, e.g., if you want another animal to back off before starting a fight, we get ritualization (the exaggeration of the clues that can be read). This ritualization leads to the production of signals, which are produced in order to be read.

Call Systems

Many animals have developed a repertoire of specific sounds as a response to a specific situation. Vervet monkeys for example, have been found to have vocal symbolic communication about predators. Researchers have identified calls for

leopard, eagle, and snake, respectively, which make the monkeys climb a tree, look up and run to bushes, or look down. Playback experiments with a tape recorder have been shown to be sufficient to induce the same behavioral responses. Chimps in the wild have been found to use about three dozen vocalizations. Each call is meaningful in itself, but combinations have not been found to be used for special purposes. If this is true, then the combinatorics utilized in human languages which enables us to create an infinite set of sentences from a finite set of sounds, is not present.

Bee Dance

First observed by Karl Von Frisch in the 1940's, researchers have uncovered an elaborate symbolic communication scheme in honeybees. When the honeybee returns to the hive, it informs the rest of the bees about the location of a food source, by performing a figure-8 waggle dance. The angle of the figure-8 axis points toward the food and the duration of the dance is proportional to the distance from the hive. The dance is not only utilized to locate food, but also to tell the location of a water source. Another version of the waggle dance tells the bees to commence a once-in-lifetime swarm.

Alex the Parrot

Alex is an African Grey parrot who is the subject of research conducted by Dr. Irene Pepperberg at the University of Arizona. Alex has learned a 70-word vocabulary that includes thirty object names, seven colors, five shape adjectives, numbers up to 6, and a variety of other words including the abstract concepts of same and different.

Transcript of dialogue [19]:

Irene: Okay, Alex, here's your tray. Will you tell me how many blue block?

Alex: Block.

Irene: That's right, block - how many blue block?

Alex: Four.

Irene: That's right. Do you want the block?

Alex: Wanna nut.

Irene: Okay, here's a nut. (Waits while Alex eats the nut.) Now, can you tell me how many green wool?

Alex: Sisss...

Irene: Good boy!

Kanzi the Bonobo

Bonobos, sometimes also called pygmy chimps, have been shown to be more adept at learning communicative skills than normal chimps. The interesting thing with Kanzi, is the way he was taught. Sue Savage-Rumbaugh and her team devised a keyboard with arbitrary symbols that chimps, lacking the vocal apparatus necessary to talk English,

can press, and in that way use as a substitute for words. They had been working on Matata, Kanzi's foster mother, for quite some time without much success, while Kanzi as a baby had been with her, hanging out on her back. To the researchers' big surprise, they one day found out that Kanzi had acquired the symbolic communication scheme all by himself, and was quickly surpassing his mother's abilities both in comprehension and production. They decided to abandon the structured training approach, and let Kanzi, not an experimenter, decide which words were acquired. While still using the keyboard as a communicational device, the trainers would also speak in English the whole time. To their amazement it started to look as if Kanzi also was acquiring an understanding of spoken English. Subsequent tests confirmed their suspicions. In Sue Savage-Rumbaugh's opinion, Kanzi's linguistic abilities (comprehension) are on level with a two-and-a-half year old child. He can interpret sentences he has never heard before, like "Kanzi, go to the office and bring back the red ball". [20] The way Kanzi has acquired his linguistic skills also gives pointers about greater plasticity in younger brains, and critical time discussed in section 2.2.

Koko the Gorilla

The gorilla Koko was brought up and trained in American Sign Language by Francine Patterson. After four-and-a-half years of instruction, Koko had learned 132 words, and leveled out at approximately 500 words. Koko could be productive in her sign language, making new words to describe new objects, by combining known ones. 'Eye-hat' for mask, 'white-tiger' for zebra, and 'finger-bracelet' for ring, are among the words Koko is reported to have created. Patterson also reports that Koko uses her signs for such purposes as to swear, rhyme, joke and lie. Sometimes she signs to herself when she is alone. [21]

Phoenix and Akeakamai the Dolphins

A group of researchers at Kewalo Basin Marine Mammal Laboratory led by Louis Herman, conducted experiments using two different types of artificial languages to test the bottlenosed dolphins abilities to understand syntax. One of the languages was based on sounds and was taught to a dolphin called Phoenix. The other was based on gestural signs and was taught to another dolphin called Akeakamai. The two languages used different word order to investigate how this would affect the dolphins' ability to acquire the languages. One of the languages used a left-to-right word order (Phoenix), and the other the inverse. Both languages included words representing agents, objects, object modifiers, actions, and conjugations that were recombinable, using sentences from two to five words in length. The word order was shown to be understood by testing the dolphins with semantically reversible sentences. These are sentences for which the subjects and objects cannot be interpreted by meaning alone, but only by the use of syntactic knowledge. In a sentence such as: "The cat chased the mouse", world knowledge is helping us to infer who is likely to chase who and would not qualify as a semantically reversible sentence. The dolphins were presented and acted correctly on sentences such as "Pipe hoop fetch" (take the hoop to the pipe) as well as

"Hoop pipe fetch". The dolphins also responded correctly the first time they were exposed to new sentences such as: "Person left Frisbee fetch", on the basis of previous understanding of the words, and their relationships in a command structure. The dolphin that was taught a sign language did also respond correctly to commands given to her on a TV monitor viewed through an underwater window already first time with no previous experience of TV. [22]

4.3 Defining Language

Many readers will feel that, "Well this is kind of like language", as they traverse the above list of communicative abilities in animals. The definition of language in Webster's Collegiate Dictionary is:

> "A systematic means of communicating ideas or feelings by use of conventionalized signs, sounds, gestures, or marks having understood meanings."

This definition would encompass several of the examples above. But for some, the definition would be much stricter. It isn't really language unless it has *syntax,* is not an unusual claim for many linguists. It is not without reason they have put their emphasis on syntax, because it adds a tremendously powerful aspect to communication (I will get further into the power of syntax in chapter 5). But Alex, Kanzi, Phoenix and Akeakamai are all displaying the ability to use syntax to a limited degree. There are ongoing debates in scientific journals, on whether these animals should be considered able to use language or not [23]. But there are other things than the animals language ability at stake here. The two dolphin languages were using an interface style that did not allow for two way communication. Furthermore, the linguistic system vas very small. A cognitive structure consisting of only Frisbee, ball and hoop, can only discuss Frisbee, ball and hoop topics. The examples above can of course not be compared to the works of Shakespeare, or even the mindless jabbering of some foul-speaking teenagers. I am in no way trying to claim that the examples above are coming anywhere close to the complexity of how people talk, but trying to show that the difference seems to be one of degree, not of natural kind. The difference lies in the *expressive power* of the different communicative schemes. Street signs can not express feelings, body language can not express past events and so on. Different styles of communicative bridges are able to transfer different things.

There is today no consensus on how to define intelligence, and no consensus on how to define language. No wonder people disagree. The parrot Alex's talents in a pea-sized brain raise some very interesting questions. Perhaps the cognitive processes required for symbolic communication are less complex than we tend to assume? Or perhaps, more important than brain size, is the mode in which that particular brain structure processes its input data. It is conceivable that other factors such as cultural

aspects, and the need for social interaction, favors higher plasticity, and that small brains with a high level of plasticity, or with plasticity during a longer time, might be better at detecting language patterns, than bigger ones but with a more rigid, dedicated mode of operation.

4.4 Hardwired vs. Softwired

Chomsky's Perspective

Many linguists today are comfortable with the idea that language is somehow genetically hardwired in humans. Noam Chomsky, perhaps the best-known and most influential linguist of the second half of the twentieth century, is largely the one responsible for this view. He suggests that what is required for language acquisition is a 'Language acquisition device' (LAD). When born, humans have a set of rules about language already built into their brains. These rules give human beings the ability to learn language. It's some kind of predisposition to discover grammars in one's surroundings. Chomsky calls this the Universal Grammar (UG). The Universal Grammar is the basis upon which all human languages are built. Although the grammar varies in different languages, this is only on the surface according to Chomsky, and by setting a few parameters the child can transform the UG into its mother tongue. Acquisition of language through exposure alone would simply not allow for the sort of communication we see between humans, Chomsky claims. If humans were born with a clean slate, there would be no sense of consistency in human thought patterns, and communication would be far too abstract to even occur. He claims that it would be little short of a miracle if children learnt their language in the same way that they learn mathematics or how to ride a bicycle. Chomsky's premises (according to Geoffrey Sampson [13]) for arriving at this view are:

1. Speed of acquisition
Children learn their first language remarkably fast. Language acquisition contrasts in this respect with the acquisition of other bodies of language, for example, knowledge of physics.

2. Age dependence
Language acquisition in childhood works differently than in later life. There seems to be a critical period for learning a language, as is true quite generally for the development of the human body.

3. Poverty of data
Children are usually given little or no explicit instruction about the structure of their first language." It is clear that the language each person acquires is a rich and complex construction hopelessly underdetermined by the fragmentary evidence available".

4. Convergence among grammars
Children in a language community all acquire essentially the same language as one another, despite differences in intelligence and conditions under which language was acquired.

5. Language universals
All languages resemble one another with respect to a number of structural features that are by no means necessary properties to any conceivable language – a system would not have to have features these in order to be called a language, but in practice all human languages have them.

6. Species specificity
Members of species other than man do not master human-like languages even when given access to experience comparable to that available to human children. (referring to chimps being raised as family members)

Two Opposing Schools

There are then two opposing schools dominating the study of language acquisition. One claims that language is a social phenomenon, like culture, and the other holds that the grammatical rules on which we build a language are something we are born with. The debate has been going on for more than forty years since Chomsky published the book *Syntactic Structures* in 1957. Newer books, like Steven Pinker's *The language instinct* [24], and Derek Bickerton's *Language and Species* [25], are following in the nativist footsteps of Chomsky. The debate is in a way part of a much bigger debate about whether cognitive structures in general are innate and hardwired; or cultural, softwired creations.

As if to echo Chomsky's 'poverty of data' problem, it seems that the problem of how we acquire language - both from the nativist and empiricist point of view - must be solved by the fragmentary evidence available. It doesn't seem like there will be a consensus on this question in the near future. My personal opinion on the matter is that we are equipped with a PAD (pattern-acquisition device), not a LAD.

Hockett's Definition: What is "Language"?

Charles Hockett was another linguist preoccupied with finding out what separates human language from other modes of communication. Hockett, like Chomsky, envisioned an all-or-nothing barrier to language. In 1960 he made a list of 16 design features that he felt characterized language. Hockett's claim is that in order for something to be considered a proper language, it has to fulfill all 16 items of the list. The list is often quoted in arguments for the claim that only humans have language.

1. **Vocal-auditory channel** (communication occurs by the producer speaking and the receiver hearing)
2. **Broadcast transmission and directional reception** (a signal travels out in all directions from the speaker but can be localized in space by the hearer)

3. **Rapid fading** (once spoken, the signal rapidly disappears and is no longer available for inspection)
4. **Interchangeability** (adults can be both receivers and transmitters)
5. **Complete feedback** (speakers can access everything about their productions)
6. **Specialization** (the amount of energy in the signal is unimportant; a word means the same whether it is whispered or shouted)
7. **Semanticity** (signals mean something: they relate to the features of the world)
8. **Arbitrariness** (these symbols are abstract; except with a few onomatopoeic exceptions, they do not resemble what they stand for)
9. **Discreteness** (the vocabulary is made out of discrete units)
10. **Displacement** (the communication system can be used to refer to things remote in time and space)
11. **Openness** (the ability to invent new messages)
12. **Tradition** (the language can be taught and learned)
13. **Duality of patterning** (only combinations of otherwise meaningless units are meaningful-this can be seen as applying both at the level of sounds and words, and words and sentences)
14. **Prevarication** (language provides us with the ability to lie and deceive)
15. **Reflectiveness** (we can communicate about the communication system itself)
16. **Learnability** (the speaker of one language can learn another)

As far as lists go, Hockett's list is reasonable. The list does indeed capture some essential aspects of spoken human language. Written language does not qualify because of entries 1,2,3, and 6, and would require another list with perhaps some additional items. Using this new list, spoken language would not qualify. Sign language would require its own list. Bee-dance communication could also be the topic for a list including entries 2, 3, 4, 6, 7, 8, 9 and 10 from Hockett's list and an additional number of items, such as a modification of 1: "Communication occurs by the producer performing a waggle-dance", excluding both vervet monkey calls and human communication.

These kinds of lists are the result of the pattern-matching nature of our brains. To classify and make groups that include some things and exclude other things, is in the nature of human beings. The world is full of classifications such as: "There are seven kinds of races among humans" or "There are seven colors in the rainbow". There is nothing objectionable about Hockett's list as such. That is: Hockett's list as an attempt to identify the smallest common denominator of spoken human language. It is however, a categorical mistake to confuse Hockett's list with a set of criteria that any hypothetical type of communication would have to meet in order to qualify as a language. A list of characteristics of apples cannot be used to identify fruits. What is objectionable, is to attribute absoluteness to an arbitrary list. There is a whole spectrum of various kinds of communication. It is of course possible to draw a line somewhere and claim that anything above this line is called language and everything below is not, but what is gained from that? Where to draw the line seems arbitrary. If

we define language as something only done by humans, it will come as no surprise that only humans have language.

Characteristics of Non-Human Language

What would be the characteristics of a non-human language that would allow us to instantly recognize it as a language like form of communication, even if it were quite alien with respect to all human languages? This is a question Terrance Deacon asks himself in 'The Symbolic Species' [2]:

> "A language-like signal would exhibit a combinatorial form in which distinguishable elements are able to recur in different combinations. It would exhibit a creative productivity of diverse outputs and a rather limited amount of large-scale redundancy. And although there would be a high degree of variety in the possible combinations of elements, the majority of combinatorial possibilities would be systematically excluded [...]
>
> If a radio-telescope observer identified a signal emanating from distant space with these characteristics, it would make world headlines, despite the fact that the meaning of the signal would remain completely undecodable. With far more to go on than this, in even superficially studied animal communications, we can be reasonably sure that for the vast majority of likely candidate species such a signal has not yet been observed. Instead, though highly complex, and sophisticated, the communicative behavior in other species tend to occur as isolated signals, in fixed sequences, or in relatively unorganized combinations better described by summation than by formal rules." pp. 32-33.

Dolphin Language?

Although I think language could take on many other forms than what Deacon describes, the findings of Markov and Ostrovskaya from the USSR Academy of Sciences [26], are matching the requirements of Deacon very well, but haven't made the world headlines that Deacon anticipated.

> "It is rare that bottlenosed dolphins produce single signals. As a rule, this is typical of very young or isolated adult animals. In normal communication, the intensity of signalization is very high, reaching sometimes 50 signals per minute.
>
> In free dialogue (for instance, during communication of isolated animals through electroacoustic communication link), signals with different structures are combined into groups, the way human words are combined to construct phrases. Grouping is well pronounced in normal conditions of communication between calm animals but it drastically changes or disappears in stressed situations, when the frequency range of communication is severely restricted or communication between individuals is broken.

> In a number of situations, when dolphins mostly are using tonal signals, one can identify sets of tonal signals with a common structure component. The analyses of variability of signals from the set have shown that their middle sections are most stable, while edge sections (especially those at the end) are extremely variable.
>
> Variability behaves differently in groups composed of different signals. This allows one to assume that the order in which signals follow each other in groups, is meaningful for the animals and that the described variability depends on the interaction of signals and, consequently, on the existence of organization in a sequence of signals. This assumption is supported indirectly by dolphins producing groups with identical composition, sometimes consisting of signals with very complicated structure.
>
> This assumption is rather non-trivial and actually recognizes the ability of bottlenose dolphins to generate organized messages (text)."

4.5 Barriers of Disbelief?

We often come across statements like: "What separates humans from the other animals is....." – Or "only humans have..." We seem to have a strong need to draw a line between us and the other animals. "Only humans make tools", was a claim that held on for a long time. But after it became clear that this was not the case, the claim was rephrased as "Only humans make tools to make tools". That one disappeared as well. Another popular one has been "only humans are self-aware" but both the great apes and dolphins have been proven to be self-aware [27]. It seems like the last stronghold to uphold our position as different from the rest is" Only humans have language". A human-centric definition of language is logically hard to falsify, but doesn't give any substantial comparative insights. To define language as that what is spoken by a particular group of language users yields a circular definition. How are we then to define the language users? As the ones that use language?

I was once told this story: When the European settlers first came to the 'New World' of Australia, they were met by the local aborigines, who of course at that time did not speak English. The ethnocentricity of the settlers however, left the aborigines ostracized from the English language for more than 50 Years. Simply because no one even considered the possibility that these "half-apes" would have enough brains to be able to learn English, no one even bothered trying teaching them. I don't know if the story is actually true, the point is, -it could have been. Similar beliefs of racism has in our past stopped 'inferior races' from getting an education, because it was assumed they were too dumb to learn anything anyway. I'm not trying to make a foolish comparison between aborigines and dolphins, but pointing out that perhaps our assumptions of limited abilities are holding us back in other domains as well. As

mentioned in section 3.2, we must 'make a leap of expectation' in order to investigate the unexpected.

At one time it was clear that the Earth was the center of the universe, and humans (white males) were made in the image of God. This anthropocentric notion has slowly faded away, as we have discovered that indeed the Earth is only a small planet in the outskirts of a medium-sized galaxy, and that the human race evolved together with the rest of the animals. It might sound like I'm predicting the same thing will happen again, and that inter-species communication will change the way we look at our selves. Although I see the possibility, I wouldn't go as far as to make any predictions. What I want to draw attention to is the possibility that dogmas - i.e. opinions or principles based not proof, but something culturally postulated and treated as a given – can act as barriers and prevent bridge building.

Having language, and having the ability to acquire language should be kept apart as two separate things, much in the same sense that being able to invent something and learn how to use that invention, are two separate things. An argument against other animals being able to acquire language is that - if they had evolved the capacity to learn language, why has it not emerged? If indeed there was a physical organ such as a LAD, then the argument would be valid. A race would not evolve the physical equipment for something, and then not use it. But seeing language as a system of information, language acquisition is done by tapping into this system. Cultural evolution needs brains as carriers, but evolution happens on another level, and live its own life, so to speak. We have evolved the organs necessary to invent computers, and have walked around with them for a hundred thousand years without inventing it. As seen in section 2.2, a human being such as Genie, the girl who grew up in isolation, will not have language no matter how able, unless exposed to a linguistic system (le langue). Likewise, a stone-age girl imported to NYC (section 1.4) would learn to use a system far more complicated than she would ever be able to dream up.

We have looked at the requirements of language acquisition and seen that the bottlenosed dolphin meets the requirements pretty well. Apart from vocal motor-control, big brains, and a cognitive understanding of symbolic reference and syntax, bottlenosed dolphins have excellent memory, live in complex social groups, vocalize a lot and are good at mimicking.

Markov and Ostrovskaya claim that their analysis of dolphin vocalizations indicates that dolphins do have language (in a Deaconian way). If that is true, it would of course be possible to make an attempt at decoding their language. The breaking of codes is however a very different approach, and in my opinion much more difficult than the approach suggested here.

Instead of trying to learn their language, I suggest that if we try to teach them ours, perhaps they would be willing and able to learn (on the condition that we have some

technical apparatus available that could help us overcome the interface barrier, and produce and analyze the sounds they make, into sounds that we can make). I think it could be fruitful to give dolphins the benefit of the doubt, and make a leap of expectancy. That is why I propose the language scheme in this thesis. By introducing an artificial, but full-fledged language to their environment, we might be surprised at what we find.

> "Eventually it may be possible for humans to speak with another species. I have come to that conclusion after careful consideration of evidence gained through my research experiments with dolphins. If new scientific developments are to be made in this direction, however, certain changes in our basic orientation and philosophy will be necessary. We must strip ourselves, as far as possible, of our preconceptions about the relative place of Homo Sapiens in the scheme of nature." John C. Lilly, Man and dolphin [28].

Before we turn to the next chapter a quick recapture seems to be in place:

We have seen that we need a linguistic system, a brain and an interface.

Anything else needed? Perhaps a Language acquisition device (LAD)? – We don't know enough about what a LAD should be or where we should look for it, to be able to investigate it. Perhaps it is just a dogma? There are many scientists who don't see the need for a LAD, claiming that a 'pattern acquisition device' (PAD) i.e. a brain will do. We'll put the LAD aside for now, and work on the hypothesis that a dolphin brain is sufficient, LAD or no LAD.

Deciding on the interface.

The interface is for the symbolic reference part. Dolphins don't have hands, so a sign language doesn't strike me as a very good symbolic system to use. They make sounds, and have voluntary motor control over their sound productive organ. A natural suggestion is to use some of the sounds they make as symbolic message units (SMU's). We need some equipment to record and recognize the sounds they can make, and we need to be able to classify them somehow as SMU's. We also need to be able to generate such signals. This is the topic of Chapter six and seven. If we assume that we have this in place, we have the sounds that make up part of the bridge, but we need to decide what they should represent.

Deciding on the system

The system is an agreement between individuals about what these symbolic representations should stand for. What more do we know about this system that manages to transfer ideas and intentions across individuals? Is there any structure that we can build on? Any rules? The next chapter will look at this system from a linguistic point of view, to see if it is possible to use the same rules in a human dolphin language as what is used in human communication.

– 5 –
Linguistic Aspects of the Bridge

Language study is, in linguistics, commonly divided into sub-fields. We are able to conceptually separate languages into layers, and in that way study particular phenomena belonging to one layer without the interference of other aspects of language. Exactly how many layers, and where the borders are, is not always clear.

Many Levels of Arbitrary Components

The great innovations of symbolic reference and syntax are at the core of the structure that enables the passing of intentions between people. To continue our bridge metaphor: phonetics, morphology and syntax provide the structure of the bridge, and semantics and pragmatics are the packages that traverse the bridge. The right focus for discussing language in this context, is to pay attention to the arbitrariness of the building blocks of language. I.e., it doesn't matter which ones you choose, as long as there is an agreement. Actually, even the division into layers is arbitrary and criticized by some linguists. [29]

Phonetics/Phonology

Phonemes are the basic building blocks of every human spoken language. These are sounds that do not have meaning in themselves, but make up words when combined with other phonemes. Most languages have between thirty and fifty building blocks, that is to say, sounds that the speakers of that language makes a (phonological) distinction between. This is an important point. There are no clear borders between different sounds, but the speakers will classify them into clear categories. Let's take an example. Although the English language use both aspirated k as in 'cat' (with a puff of breath after) and unaspirated k as in 'back', these different sounds never distinguish one word from another, and are by speakers of English considered to be the same thing. Speakers of Hindi, however, consider the two to be different sounds and have words where the only difference is aspirated k as in 'khal' (skin) and unaspirated as in 'kal' (time). [30]

Norwegian speakers distinguish between nine different vowels, whereas the English language only has five, and so on. The reason we have around fifty phonemes is probably because of the anatomy of our throats and ears, and their ability to produce

and discern roughly that amount of sounds, rather than being a particularly well-suited number of building blocks for languages.

One could imagine a language built on much fewer phonemes. Two, for example. Morse, the old telegraph language, is such a language. Morse is only used as a step between, in order to translate to and from alphabetic letters (which in this respect is similar to phonemes). So it's not really a true two-phoneme scheme, since it is using its two signals on a lower level of interpretation than the phoneme level in question. But we can see how it's possible to make an infinite number of words with the number of combinations you can make with only two different sounds. The basic building blocks in a computer is another example of how any text can be represented by only two distinct types. We could also imagine having many more than our fifty phonemes. Say we used a much more refined way of dividing our noises, so that the vowel distinctions we make today of only 5-10 in the confined room of possible vowels, could be extended into 60 different distinctions, or 600?

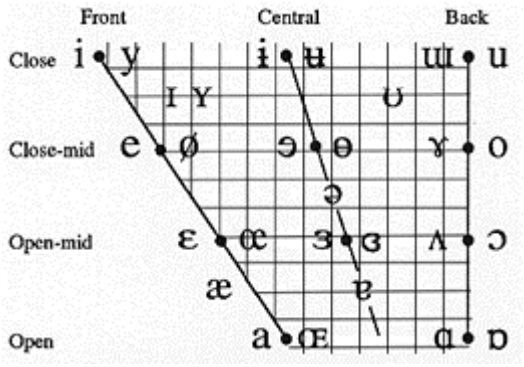


Figure 12. A linguistic drawing of the mouth. The cavity allows for making many different vowels by moving the tongue (front, central, back) and opening and closing the mouth. What if the vowel distinctions we make today could be extended into 60 different distinctions, or 600?

We would have a very efficient language were we could say a lot in a short time, but we would not be able to say more. The number of phonemes in a language is arbitrary, and has nothing to do with the expressive power of the language.

Morphology

Phonemes combine to make words. This is done in all human languages, and is the first level where infinite use of finite media comes into play. This can also be described as the power of combinatorics. Thirty phonemes freely combined in strings varying in length from 1 to 10, makes more than 30^{10} = 590.490.000.000.000 possible combinations. Far from all combinations are possible though, and in fact different languages will have different restrictions on what combinations are possible to make. In Mandarin Chinese, words never end with b, p, m, f, d, t, l, g, k, or h sounds, for example. In English, you will never find more than four consonants in a row. Still, there is no risk of running out of possible combinations in any language. One reason of course, is that there's no upper limit to the length of words.

What all languages have in common, is that the relation between the sound of a word, and the meaning of the word, is arbitrary, i.e., there is no natural link between the two. Except for a few words known as 'onomatopoeia', which sound like the concepts they express (e.g., boom), convention alone decides what the label of a concept should be. What combinations of phonemes are used to represent a concept doesn't matter, as long as the convention is established.

Phonemes can also be used as morpho-syntactic markers – i.e., prefixes and suffixes, which do not carry meaning in themselves, but add functional information to a word. In English, -s indicates plural, un- makes a negation, and -ed indicates past tense, for example. This is of course also an arbitrary choice, and how it's done varies from language to language. In fact some languages, like Mandarin, hardly use it at all.

Syntax

The other great innovation, on which all human languages rest, is syntax. Words combine to make phrases, and phrases combine to make sentences. The power of combinatorics comes into play again. Language, to use the words of Pinker [24], is a "discrete combinatorial system". When words combine, their meanings do not blend into one another, as colors do when they are mixed. Red paint and white paint combined makes an intermediate pink paint, but the combination 'a black cat"' does not identify a concept intermediate between black and cat. This combinatorial structure allows the expression of more complicated ideas, and allows for an unlimited number of new sentences to be created. As mentioned earlier, chimpanzees in the wild have been reported to use around three dozen different vocalizations, but combinations of these have not been observed. If that is true, and the words mean the same independent of context, then three dozen calls will only communicate three dozen concepts/events/things.

To be able to use a language properly, you have to know how many participants are involved in for example, the event that is labeled by a verb.

'Cry' only takes one argument: the one doing the crying (actor).

'Hit' takes two arguments: the hitter (actor) and the person or thing being hit (undergoer).

'Eat' can take one or two arguments: 'the wolf ate' or 'the wolf ate cranberries'.

'Give' takes three arguments: a giver, a recipient, and a thing given.

Argument structure specifies not only that all arguments associated with a verb be mentioned in the sentence, but also that those arguments be linked to particular syntactic positions. The linking of an argument to a syntactic position depends on its thematic role (the role it plays in the event).

'The wolf ate grandma' and 'grandma ate the wolf' are clearly different things. A convention must be established to decide who is the actor, and who is the one acted upon (the undergoer). In many languages, word order decides this. The word order is arbitrary and varies from language to language. Declarative sentences in English use SVO (subject-verb-object), as in the examples above. Hixkaryana spoken in the Amazon basin in Brazil, uses the opposite (OVS). [30] There are six possible permutations of the units used for word order: SVO, VOS, OVS, VSO, OSV, and SOV. Each permutation is found in some human language.

Not all languages use word order as a distinguisher. Case marking, i.e., affixes to the words themselves, is another way to remove ambiguity. It is a common way to resolve 'who did what to whom with which means' in many languages. In such languages, the ambiguity issue is resolved on a morphological level instead of a syntactic level, and the word order is used for other things. English once had lots of case markings, and allowed all possible word order permutations. Today there are still traces of it, as in "he" if the person is subject, and "him" if he is object.

Another way to divide lexical items is in the distinction between open and closed-class items, i.e., items that primarily carry meaning and items that primarily provide structure. The first type is called open-class because it readily admits new members and includes nouns, verbs, and adjectives. The second type includes prepositions (e.g., in, under), determiners (a and the), quantifiers (e.g., some, many), and morphological markers (e.g. -s for plural –ed for past tense). It is called closed-class because it strongly resists the introduction of new members. Together they cooperate in syntax to communicate meaning in a structured fashion. Closed-class items as a group have highly abstract meanings, and function as scaffolding for open-class items.

Semantics

Semantics is the field of study that deals with meaning. This is the reference side of the concept of symbolic reference.

Cognition is autonomous from the code of grammar described above. We can make sentences that are grammatically correct, but nonsensical like Chomsky's classic example "colorless green ideas sleep furiously", and we can make sentences that are ungrammatical, but still makes sense, as in: "Welcome to Chinese Resturant. Please try your nice Chinese Food with chopsticks: the traditional and typical of Chinese glorious history and cultural". [24] So the question of how important the syntax really is, arise. Children, immigrants, and others with a faulty grammar are still able to use the communications bridge in order to express their intentions. Naturally, if we started mixing the English SOV word order with the Hixkaryana word order of VOS, things could get really confusing. But are some aspects of grammar redundant? It seems some aspects of syntax is allowing for a more streamlined transfer, and perhaps increasing the bandwidth, but are not essential for transfer to occur.

So what about the arbitrariness of semantics then? In some sense, it seems more real and absolute than the other layers. We are referring to real things. A wolf is a wolf in all languages. Although it will have different names, the sense and reference is still the same. The idea of a wolf is not arbitrary. Arbitrariness comes in because when we say 'cry wolf' we are not necessarily talking about wolves.

Pragmatics

The same sentence means different things in different contexts, and means different things in different cultures. The sentence "Norwegians eat more fish than Americans" is ambiguous, but would only be misunderstood in a culture where cannibalism was common. "Have you eaten yet?" is a polite form of saying "How are you?" in Chinese. The direct semantic meaning of a sentence is often not the intended message the speaker wants to put forth. Bridging the 'praxis gap' mentioned in section 1.2, is the mental exercise of translating a sentence like "It's cold here..." into "Ah, she wants me to close the window" in one instance, and "Ah, she doesn't think the atmosphere is romantic enough" in another. There is a gradual shift, without any clear border between pragmatics and semantics. This is partly because if you say what you mean, no lying, no irony, no head games, and no sophisticated hints about what you want, the two levels collapse into one. "I want you to close the window!" can have the same semantic and pragmatic reference.

5.1 Design Considerations in a New Language

We have seen that phonetics, morphology, and syntax constitute the structure of a communications bridge, and that semantics and pragmatics constitute the packages that get transferred. The process of designing an artificial language for inter-species communication must look at all these aspects of human language, and make some choices concerning design and structure of this new bridge. The software described in chapter 7 is only one aspect of the bridge, and further work is required in choosing what kind of grammar to use, which words should be incorporated, and so on. It would, of course, be nice to make it as easy as possible to acquire this language, so it would be advantageous to look at some existing artificial languages, as well as advantages and disadvantages with certain implementations in natural languages when deciding the structure of the language scheme. Is it possible to cut some corners somewhere? A full implementation is outside the scope of this thesis, but I will go through the levels of language one more time, this time with some design considerations in mind.

Although English is today de facto *the* international language, it has not become so because of its merits of being so easy to learn, but rather in spite of being so difficult. Critical mass does not necessarily favor products of excellence, as anyone in the

software business is aware of. Esperanto is the most famous constructed language, but not the only one. IDO, interlingua, Occidental, Lincos, Glosa, Lojban, Novial, Vorlin, Gilo, NGL, Ceqli, Arulo, Cosman, Frater, Folkspraak....there are over a hundred constructed languages to get inspiration from.

Phoneme Considerations

As mentioned earlier, the external representation of a language is arbitrary, and can be represented by sound, as in normal languages, or with hand signals as in the sign language of the deaf, Braille the written language of the blind, colored dots or lines on a paper as in our written languages, or whatever. If we made a language representation of blinking lights of different colors, it would work just as well. It is necessary to find a representation that is possible for the target audience (dolphins in this case) to perceive and produce. We need to be species-specific on this level, and must look at the sound production of dolphins to determine what kinds of sounds they naturally produce. If our job were to design a new language for Japanese people, for example, it would be advantageous to omit the /l/ and /r/ phoneme distinction, since we know that they have difficulties with this. From the set of sounds they naturally produce, we can choose an arbitrary subset, and assign phonetic properties to these. The word 'phoneme' and the word 'syllable' are closely connected with human linguistic sounds, and have been replaced by the more generic term 'symbolic message unit' (SMU). As we saw in the previous section, there is no need to build the language on 30-50 SMU's just because that's the amount we are using in human languages. We can start in the other end, and ask ourselves: "how many words do we need to have?" and then look at the relationship between the length of words, and the number of SMU-combinations necessary to produce that amount of words. Let's say we want to be able to express approximately 2,500 words, and we decide that a word should not be more than four SMU's of length. We can cover this with all possible combinations of seven SMU's ($7^4+7^3+7^2+7= 2{,}800$). With ten SMU's, and words no longer than four SMU's, we would have 11,110 placeholders for concepts.

Solresol

An interesting constructed language to mention in this context is Solresol, developed by Jean François Sudre between 1817 and 1866. To avoid giving any national language an advantage, François Sudre created a language that does not resemble any other, built on seven notes of music: Do, Re, Mi, Fa, Sol, La, and Si. It was the first artificial language to get beyond the project stage and to be taken seriously as an interlanguage. One of the advantages with this language was that it was considered very easy to pronounce, regardless of mother tongue. Another that it could be expressed in several different ways. In writing it was proposed to either use the normal alphabet and simply write for example Misisoldo (meaning 'experience'), or omitting the vowels (except the o of sol to distinguish it from Si) and write mssod. Alternatively one could use the numbers 1-7 and replace Do with 1, Re with 2, and so on. Misisoldo could

then be expressed as 3751. It would also be possible to use the musical scale, and three lines, without having to know music.

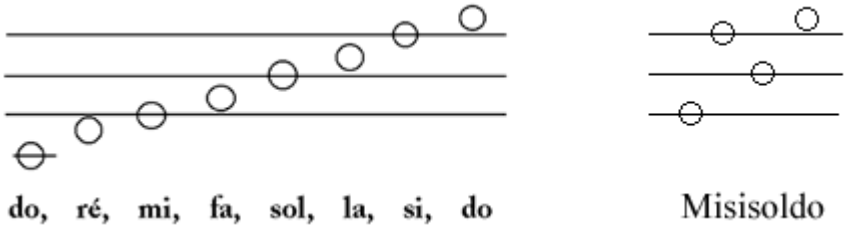


Figure 13. One possible representation of the musical language Solresol.

A kind of stenography alphabet was even created by Vincent Gajewski using seven signs, one for every SMU, that can be connected together to form words.

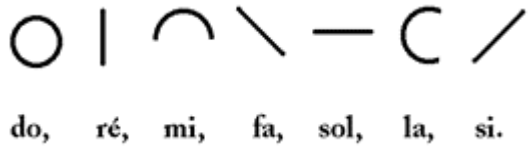


Figure 14. An alphabet for Solresol.

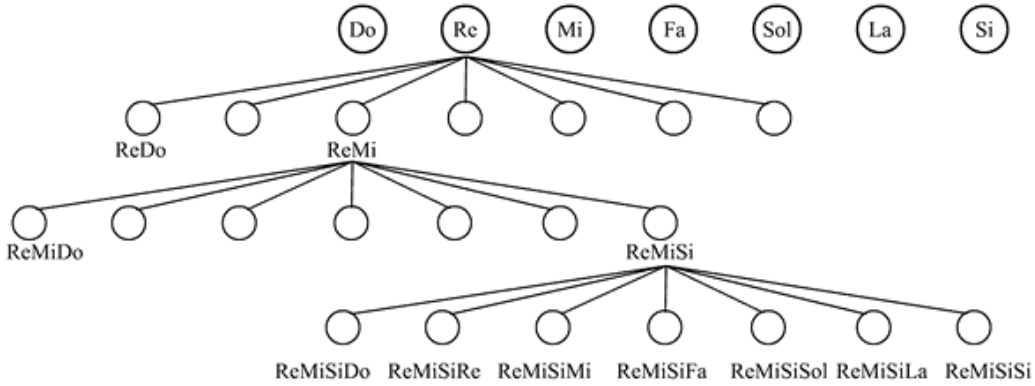
Misisoldo would then look like:

Solresol provided a means for ships to signal, with a color representation for each note. Red for Do, orange for Re, yellow for Mi, and so on. A Universal Mute and Sign Language was developed. The French army considered implementing it as a means of long distance communication, and Sudre worked for the French government for several years, trying it out. Ambitious and promising as it might have seemed at the time, Solresol never became a big hit, and very few people today have even heard about it.

The point in this context is to show that symbolic reference can be manifested in many ways. We can use colors, musical notation, flags, or sign language as SMU's in Solresol. We simply need a number of placeholders for concepts, and a systematic link between them and their physical manifestation. Returning to our constructed language Hddp, seven is of course a completely arbitrarily chosen number.

Morphologic Considerations

Words then, can be made by assigning a concept to a combination of SMU's. Holding on to Solresol for a moment longer, a visualization of the space of possible SMU combinations will look like this:



It will create a tree structure, and as mentioned above, will give room for 2,800 concepts, (if we stick to seven SMU's and restrict the length to four). Which concept we place in which node is of course also arbitrary, but with such a rigid structure in hand, it opens up for the possibility to group similar concepts in nodes close to each other. It could for example be grouped so that all living things started with 'Re' and inanimate objects were in the 'Mi' category and so on. It is not clear however, that this semantic grouping is a desirable feature. Perhaps it will confine more than it will help, by adding an inappropriate impression of linkage between sound and concept, and thereby remove the arbitrariness that is giving languages their openness.

Another aspect of morphology in human languages, is the use of morpho-syntactic markers (-s for plural, -ed for past tense and so on) mentioned in the previous section. Many languages use them extensively, but not all. Chinese for example, has no morpho-syntactic markers. The character system used as a written representation in the Chinese languages creates a natural restriction for these kinds of features, since it is not possible to add prefixes or suffixes to them. In some cases the languages have evolved alternative ways to express functionally similar features, but often they are simply not present. This kind of features is a subject of constant struggle for Chinese adults trying to learn English, complaining about unnecessary rules that doesn't add any functionality, only complicating things. Westerners trying to learn Chinese on the other hand often find the straightforward grammar of Mandarin Chinese a pleasant surprise.

- Mandarin does not have a singular-plural distinction. They say: one chair, two chair, one car, five car.
- The verbs don't have any tenses. They say: today I go, tomorrow I go, and yesterday I go.
- Adjectives are all regular and made by a sort of 'more, most' form.

The abundance of irregular verbs in English, such as 'I am, you are, he is, they were, it has been, and she will be', is another topic of frustration for foreigners. In Mandarin, the verb stays the same, regardless of gender, time, or person. The ambiguity, if any, caused by less complicated grammar is resolved largely by context, or by explicitly stating what is needed to clarify, when context is not enough.

The strong focus on grammar being the essence of language, fails to recognize the fact that there are human languages with much less, and simpler grammar, but with the same expressive power. The simplicity and elegance of Mandarin morphology serves as a reminder that unless we gain something from making things more complicated, we're better off keeping it simple.

Redundancy

We have in all human languages a large amount of redundancy, i.e., elements that are not absolutely necessary in order to interpret the message. It would be possible to randomly remove 50% of the letters in a text, and in most cases the text would still be readable. Or we could for example remove all vowels, and still understand the text. Ths s smpl sntnc whch y cn prbbly ndrstnd wtht th vwls. The vowels make it easier to understand though, especially in a noisy environment. Noise in this context is anything added to the signal that is not intended by the source. According to Claude Shannon and Information Theory, redundancy is crucial for clear communication in a noisy environment. The best way to counter noise is through repetition, or redundancy.

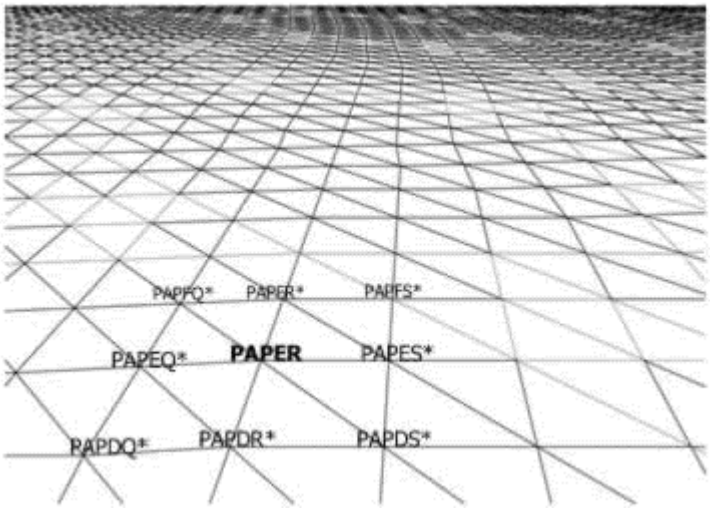


Figure 15. The space of possible letter combinations is sparsely populated by nodes that have concepts attached to them.

If we imagine the vast space of all possible combinations of letters, we see that it is sparsely populated by nodes that have a concept attached to each one. I.e., most phoneme combinations don't mean anything. A message being transmitted must be selected by the receiver from a set of all possible messages. If we have a noisy signal,

the message could easily be misinterpreted if there was another word close by. By increasing the distance to the nearest neighbor, we also reduce the chances of misunderstandings, and the language model becomes more robust. We see also that in cases where words do have neighbors close by, e.g., 'help' and 'kelp', their semantic meaning and contextual occurrence will be distant.

In contrast, the language scheme suggested in this thesis is very dense. What if all combinations of syllables are nodes that have concepts attached to them? This could open up for possible misunderstandings and be a great disadvantage. There is a relationship between the number of SMU's, the length of words, and the density of the map. The advantage of reducing the number of SMU's must be weighed up against the disadvantage of a map that is too dense. One way to, at least initially, work around the problem, is to populate the space gradually, and place the initial words to be taught as separate as possible. Another way could be to increase the number of SMU's or the length of words.

We have redundancy on several levels in our languages. On word level, languages like English have morphological markers that can be considered redundant. In a sentence like "last week, I bought three cars", the verb 'to buy' is in past tense, signaling that it is an event that has already happened, but it doesn't tell us precisely when it happened. We must add the 'last week' part to clarify. Since we have to mention the time specifically anyway, the past tense marker is redundant information. Similarly, the plural -s in 'cars' is only telling us there is more than one car, and we need to add the numerator 'three' to be precise, hence the plural –s is redundant. From an evolutionary perspective, we might suspect that aspects that did not have any advantage for the language users would, as the language evolved, disappear from the language. But if we look again at Mandarin Chinese which does not have these markers, and directly translated would say "last week I buy three car", these aspects of the English language seem truly redundant. Also on a phonetic level, Mandarin Chinese holds the redundancy down to a remarkably low level compared to other languages. The low levels of redundancy in Chinese seem like a paradox, considering Shannon's ideas about redundancy as a tool to overcome noise. Is this compensated for on some other language level, or do some languages have more redundancy than others overall?

Syntactic Considerations

The experiments being done in the Kewalo Basin Marine Mammal Laboratory, discussed in sect. 4.2, showed the ability of two bottlenosed dolphins to understand imperative sentences expressed in artificial languages. One dolphin was taught a left-to-right language, and the other the inverse. Both languages included words representing agents, objects, object modifiers, actions, and conjugations that were re-combinable, using sentences from two to five words in length. This indicates that arbitrary syntactic rules can be understood, and indicates a freedom, as far as the dolphins are concerned, to choose whichever style we want. If there is nothing to gain

from avoiding the most common SVO style, there are advantages on the human side of things to use the SVO style.

There are other aspects of syntax that differ between languages. Western languages change the word order to turn a declarative sentence into an interrogative sentence, as when "This is a good sentence" becomes "Is this a good sentence?". Chinese, once again, do not need to change the word order, they make interrogative sentences by either adding both the verb and a negation of the verb, saying: This is, is not a good sentence" – and where the answer is either 'is' or 'is not'- or by adding a vocal question mark 'ma', saying something similar to "This is a good sentence, right?" Again: ease of use must be highly prioritized provided it doesn't come at the expense of expressive power.

Semantic Considerations

Although the research mentioned above is an important indicator of the bottlenosed dolphin's linguistic potential, it hasn't taught us very much about what is going on inside the mind of a dolphin. The biggest reason being that languages previously taught to dolphins, haven't allowed them to talk back. Since Hddp enables two-way communication, what we are faced with, is to decide which words to incorporate in the language.

If we are to try to communicate with someone very different from ourselves, no matter how intelligent, can we assume that they share, or are able to acquire the same concepts as we have? Thinking in the terms of Hofstadter (chapter 3), there are reasons to believe we can. At any rate, it is possible to assume that they are able to acquire the same concepts as we do. That is, after all, one of the main things we want to test.

How many, and which concepts, need to be incorporated in the language in order to express anything substantial? What does expressing something substantial mean? Many words can be expressed by a combination of other words. 'Hospital' for example, is something like 'big house where you bring people when they are sick so that they can get well'. Is there a small subset of our vocabulary that is sufficient to explain our full vocabulary? It turns out that there are many people who have worked with these questions, so let's look at some examples.

- **Basic English**, developed by Charles K. Ogden and released in 1930, was the first serious effort to make a working subset of English, and consisted of 850 words. In his own words:

  > "If one were to take the 10,000 word Oxford Pocket English Dictionary and remove the redundancies of our rich language and eliminate the words that can be replaced by combinations of simpler words, we find that 90% of the concepts in that dictionary can be achieved with 850 words."

- **Essential World English** created by Lancelot Hogben is a revision of Ogdens list. In place of Ogden's 850-word list, he offered, a 1,300-item list of what he called Essential Semantic Units.

- **The Longman English Dictionary** uses a small set of words to express all its definitions. It consists of 2,197 words, 10 prefixes, and 39 suffixes.

- **Special English** is a subset of the English language developed by the United States Information Agency for worldwide news broadcasts on Voice of America. This limited vocabulary of approximately 1,400 words is used to talk about current events and everyday activities. Special English is an example of a controlled language, i.e., a carefully engineered version of a natural language.

Another, but related question, is whether there is a core of basic words that can't be explained by the help of other words. Such a set of words is often referred to as semantic primitives. Let's take a look at one such set of suggested semantic primitives.

aUI – The Language of Space

aUI was constructed in the 1930's by linguist and psychologist John W. Weilgart. The entire vocabulary of the language is built on the foundation of 31 basic elements. These basic elements serve as phonemes and morphemes, and all other words are built from these. In addition to this there are the numbers zero to ten

The Basic Sounds and Categories of aUI [*]

a	Space	n	quantity (plural)
A	Time	o	life
b	Together	O	feeling
c	Being	p	before
d	through, by	Q	condition
e	Movement	r	positive, good
E	Matter	s	thing
f	This	t	to(ward)
g	in(side)	u	(hu)man
h	Question	U	mind, spirit
i	Light	v	active (verb)
I	Sound	w	power
j	Equal	x	relation (relative pronoun)
k	Above	y	negative, un
l	Round	z	part
m	quality (adjective)		

[*] short vowels are written in lower-case, long vowels in upper-case, and nasalized vowels used for the numbers 0 – 10 (long and short with *).

Examples of Vocabulary

aUI	English	Literal translation
iA	Day	light-time
yiA	Night	unlight-time
Ib	And	sound-together
ha	Where?	which place, what space
hA	when?	what time
yg	Outside	opposite of inside
yga	Surface	outside space
jE	Liquid	equal (even) matter, since liquid at rest is level (of equal height)
jEn	Water	liquid (in) plenty, the plentiful liquid
ua	Dwelling	human-space
waubos	Dog	power-space-human-together-life-thing
bo*zvos	Cat	together-five-part-active-life-thing
ukbos	Horse	human-above-together-life-thing
ko*zvos	Lion	above-five-part-active-life-thing
kEos	Bird	above-matter-life-thing
jEos	Fish	horizontal-matter-life-thing, even-matter-animal

More complex words become more elaborate compounds, and I must admit quite a few are not easy to figure out without the translation. "power-space-human-together-life-thing" sounds far fetched as a self-explanatory word for dog, but if it is only seen as the etymological history of the word, then 'waubos' is as good a word as 'dog'.

There has apparently not been introduced any – or hardly any - new characters to the Chinese language in the last 1,000 years. New 'words' are introduced all the time, but they are created from combinations of the already existing building blocks. Computer is called 'electric brain', telephone is 'electric talk', and so on. But this is not how people think of the word, any more than English people think about the word 'telephone' having its roots in Greek and Latin terms for distance and sound.

The semantic primitives in aUI is an interesting list of words, and it's worth noticing how abstract all the words are. It is however also saying something about why we have more than 31 words in our dictionaries. On the one hand: why have a special word 'hospital' when we can just say 'big house where you bring people when they are sick so that they can get well'? On the other hand – and this is precisely the point of language – if there is an established concept that is frequently used, we can just make a single word for the concept, and save some breath. A complex compound of concepts - in this sense meaning an aggregate of concepts (such as 'hospital') - does not need to have a complicated label. The openness of our languages is a very important point. Once a concept is established, we can give it a label. Insisting on a semantic relevance between the label and the concept is missing the point. 'Hospital could just as well be called 'pish-posh' or something else. It is just a convention.

During the second world war (WW2) the Allied forces had a hand symbol of two fingers shaped as a v for victory. It meant something like "we are together against a common enemy, and we're gonna get him!". After WW2 the use of the symbol died out for natural reasons, but in the 60's it was taken over by the flower power generation, this time meaning "peace and love". Today we see it used by kids in photographs, but the meaning has changed again. Once a complex concept is established, it can be substituted by a single sign. What sign we choose is arbitrary. A dog could be trained to understand the 'v' sign as "time for a walk in the park". That the dog understands it as a symbolic reference, in the same way as humans, is a controversial claim, but that the dog owner will have this understanding is not.

The discussion about semantic considerations in a new language is both important and interesting. Finding the optimal set of words to include is not however an issue that can be resolved without a much bigger investigation. We can round off the linguistic aspects of the bridge instead with the insight that it doesn't really matter so much what we do, as long as we leave the system unlocked. Robustness can come from openness of the system instead of strict rules. In the beginning of this thesis we recognized language as a 'complex adaptive system' (section 1.3). A key point would be to allow it to be adaptive. Starting out with for example, Special English as a base, new words can enter the system when appropriate and words that are not used will die out.

With a group of dolphin sounds in our hand – if we followed and agreed upon a structure like what human languages are built on – any combination of these sounds could be a label for any concept, and new more complex compounds of concepts could get new combinations of sound units as the concepts arrive.

– 6 –
Computational Aspects of the Bridge

As mentioned in the introduction, the tremendous advances in processing power the last few years, as well as new computational algorithms, have made it possible today to bridge gaps that were previously unbridgeable. The key issue of the Human-dolphin dialog protocol and of this thesis is the creation of GAPR, a software that enables otherwise incompatible language interfaces – such as the human and the dolphin sound organs – to meet, and engage in exchange of information. GAPR, introduced in chapter seven, is built on an understanding of digital signal processing (DSP) and classification, so an introduction to the relevant aspects of these issues will be presented first.

6.1 Digitizing Audio

The external representation of sound is a complex series of changes in air pressure. The eardrums in animals are sensitive to this change and will pass its registration/perception to the cochlea, a part of the inner ear, which will split the signal into a set of frequencies. When humans process speech, this information is in turn passed on to a phone-mapping classifier, as mentioned in sect. 2.2. To mimic this process in a computer, here is briefly what happens:

A microphone is sensitive to changes in air pressure, and will translate this into changes in voltage. The voltage, which corresponds to the signal's amplitude, is measured at regular intervals. Each measurement is called a sample, and the number of such samples taken per second is called the sampling rate, and is measured in Hertz (Hz).

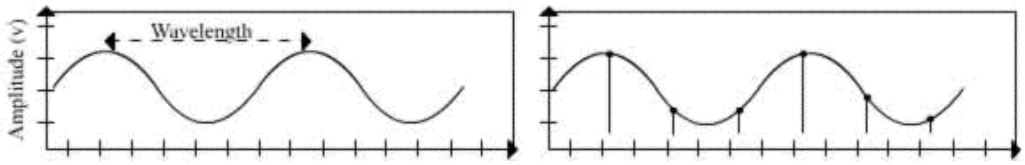


Figure 16. A sinusoidal wave and some sample measurements of its amplitude.

We see that if we take too few samples, we will loose some information. (see also fig. 19) How frequently do we have to sample in order to get enough information to reconstruct the original signal? Two characteristics of sound are of interest here: amplitude and frequency. The amplitude corresponds to our perception of loudness (the strength of the signal) and is measured in volts. The frequency of the signal corresponds to our perception of pitch, and is how many times per second the wave repeats itself. Like the sampling rate, frequency is measured in Hertz.

Natural sound, like speech, will have many different frequencies at the same time, and will change over time.

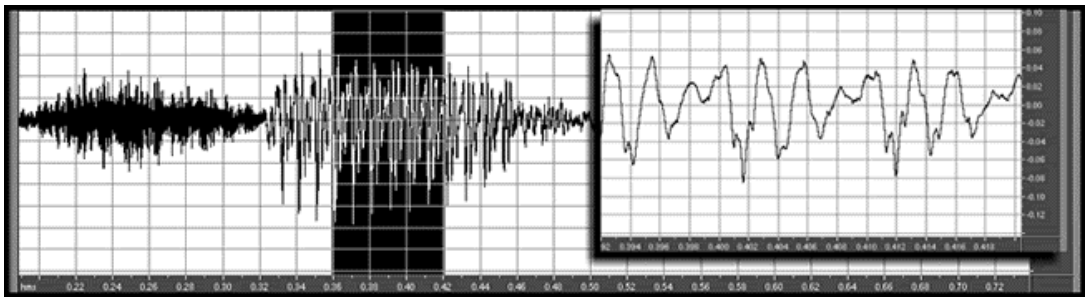


Figure 17. A short piece of human speech

Shannon's sampling theorem states that, in order to correctly measure the frequency, at least two measurements are needed within a cycle.[31] This maximum frequency for a given sampling rate is called the Nyqvist rate. This means that by taking 1,000 samples per second, the highest frequency we are able to measure is 500 Hz. The human ear is able to detect sound of frequencies in the range between 20 and 20,000 Hz. To be able to correctly measure 20,000 Hz, we need a sampling rate of at least 40,000 Hz. The sampling rate on a CD is 44,100 Hz. Most information in human speech however, is in frequencies below 4,000 Hz. This means that a sampling rate of 8,000 Hz is sufficient to get understandable human speech. Telephones are utilizing this knowledge and filter everything above 4,000 Hz to save bandwidth.

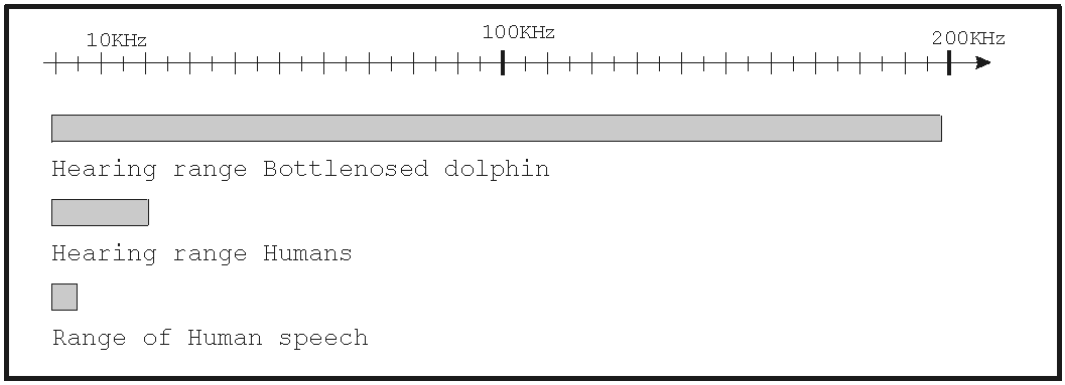


Figure 18. Domain of human speech compared to the hearing range of humans and bottlenosed dolphins.

The hearing range of bottlenosed dolphins on the other hand, is 20- 200,000 Hz. Ten times that of humans, and in order to capture the entire range we would need to make 400,000 samples per second, or 400 kHz. Since we don't know what frequency bands will be interesting to focus on, as we do with human speech, it would be desirable to use the full spectrum. This is unfortunately far too much to ask for a real-time classifier with today's technology. By recording and analyzing dolphin vocalizations at 44.1 kHz, we will see that there is still enough information in this frequency range to get a working model.

Another limitation that we need to consider is the accuracy that the samples should be stored wiyh. Even with a low sampling rate of 8,000 Hz, we have 8,000 discreet numbers that needs to be stored in an efficient way every second. They are usually stored as integers, and the granularity is decided by how many bits you use to store a number. This is referred to as the bit-rate or quantization, and is usually 8, 16 or 32. By only using 1 bit per number, the amplitude can be either one or zero. Clearly too coarse, but a reminder that whatever bit rate we choose will be an approximation of the real value. If we use 8 bits per number, we have 256 values that the measurement can take, with 32 bits we have 4,294,967,296 values.

Finally, the dynamics of the signal, i.e., how sudden changes in amplitude are preserved, will suffer by a reduction in sampling rate.

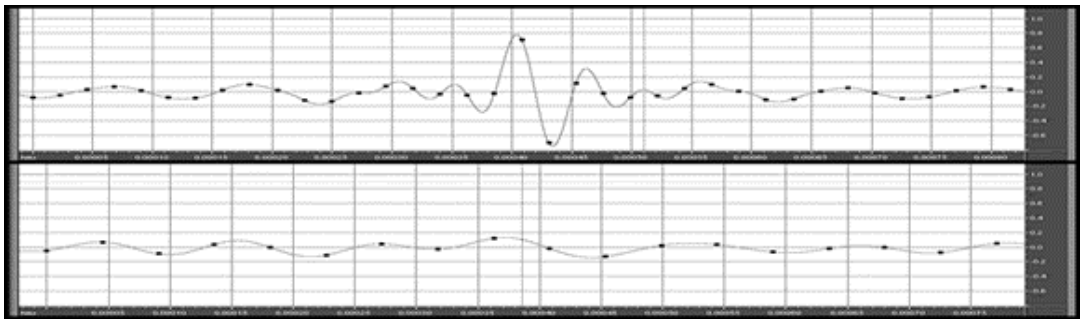


Figure 19. Same signal with different sampling rate. 44.100 Hz (above) and 22.050 Hz (below)

Other compromises that will affect the outcome of our bridge, will be mentioned as we move along.

6.2 Feature Extraction

Although we could have attempted to classify the audio signal directly from the amplitude measurements, both animals and speech recognition applications will first transform the waveform into its spectral features. Since frequency is how many times a cycle repeats itself over time, doing a frequency analysis involves calculations over a certain time frame. In order to compute the spectral features, the continuous flow of amplitude measurements will be treated as a series of chunks of limited and equal length. A time window of 10 ms is common, and the equations performed on each chunk are called a Fourier transform (FFT). The output is called a spectrum and shows the frequency components of the wave at that time.

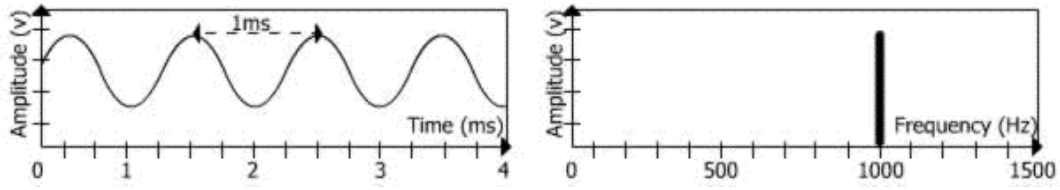


Figure 20. An amplitude/time graph of a periodic sinusoid, and the corresponding spectrum.

Here the x-axis shows the frequency range and the y-axis shows the amplitude for each frequency component. Again, resolution becomes an issue. The calculations involve deciding which frequencies with which to compute an average amplitude. The granularity becomes how many units we decide to divide the frequency range into. We can for example, divide the frequency range into 128 units and calculate the average amplitude on each of these. We can store this information in an array, let the size of the array be 128, and let the value in each index represent the average amplitude for that frequency.

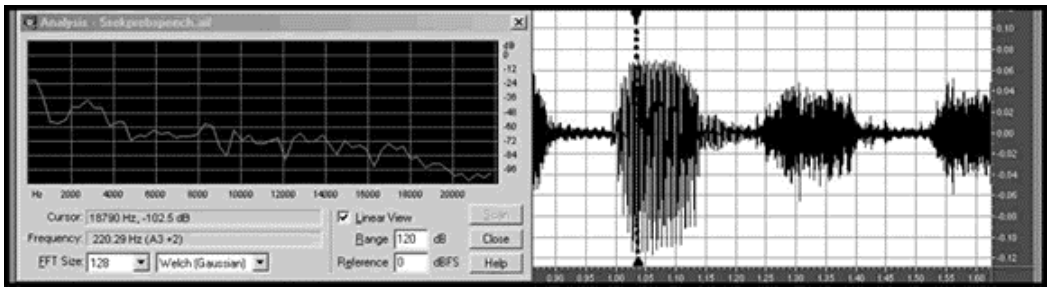


Figure 21. A spectrum of a selection of speech with FFT size 128.

By repeating this process with slight displacements in time, a two dimensional array called a spectrogram is created.

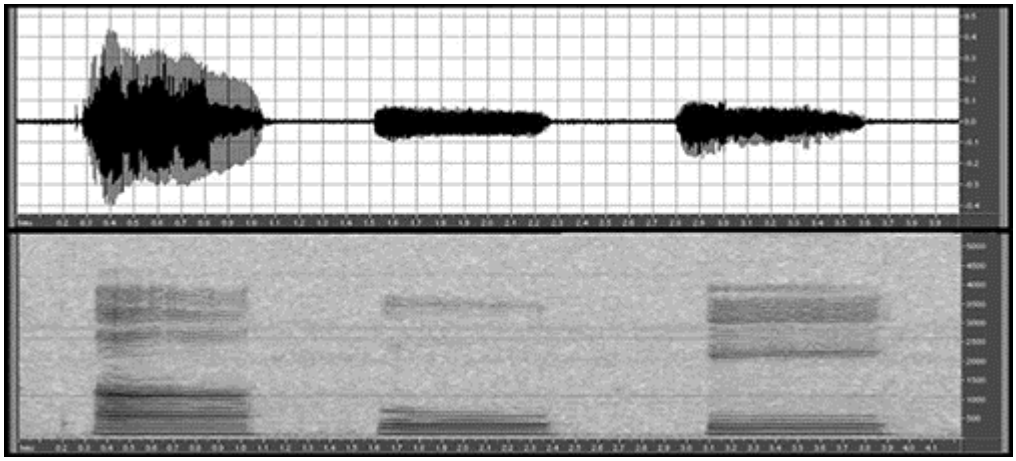


Figure 22. A wave and spectrogram of the vowels A O and I.

A spectrogram shows how the frequencies of the waveform change over time. Here the x-axis is time, the y-axis is frequency, and the z-axis, represented by darker or lighter areas, is the amplitude. Darker areas represent higher amplitude. We are now faced with resolution issues in three dimensions. The x-axis resolution is the amount of displacement in the sliding window, the y-axis is the frequency resolution, and the z-axis is the amplitude bit-rate. Although calculating the spectral features involves a large number of mathematical operations, the graphical representation is quite intuitive to us. Our ears are doing the same transformation before our brains are being presented with the sound perception, so we are quite used to thinking of sound as pitch-change over time. Sheet music and MIDI notation are familiar time-frequency diagrams similar to a spectrogram

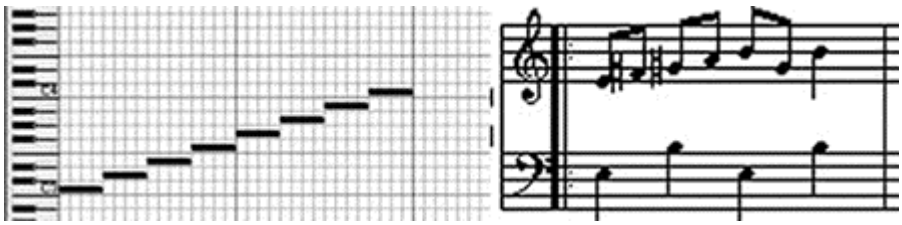


Figure 23. Sheet music and MIDI notation are time-frequency diagrams.

A spectrogram can often be more informative than a time/amplitude graph. The dark horizontal bars on a spectrogram, representing spectral peaks, are in speech called formants. Different vowels have their formants at characteristic places, and a trained person can read a spectrogram of speech, and tell what is being said. This is not possible from an amplitude/time graph. It is also helpful to study spectrograms when analyzing dolphin vocalizations.

Dolphin vocalizations can be grouped into three kinds: clicks, pulses and whistles.

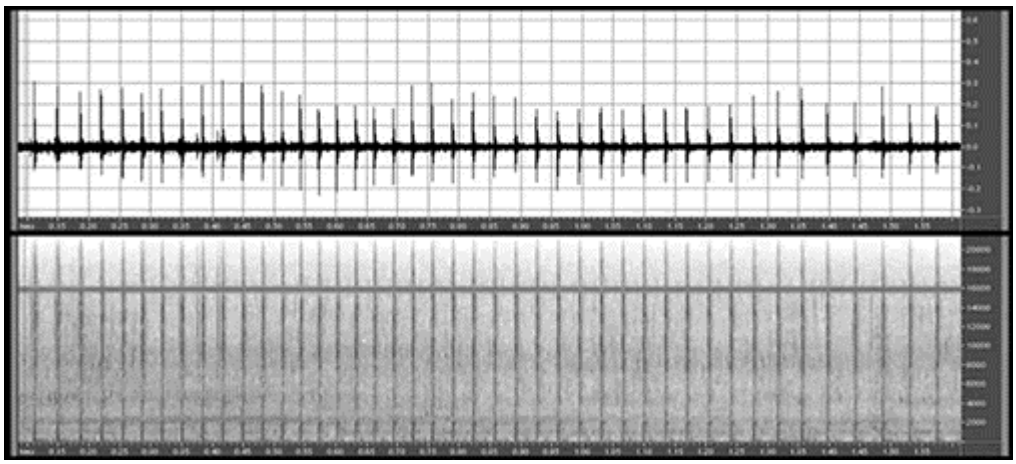


Figure 24. Clicks are used for echolocation, i.e. as sonar, a tool for locating objects, and perhaps also for communication.

Clicks are used for echolocation. These are short, broadband pulses ranging from .2 to 150 kHz. The dolphins' sonar surpasses overwhelmingly all existing technical devices in detecting and recognizing small-size underwater targets [32]. Dolphins are able to click and whistle at the same time.

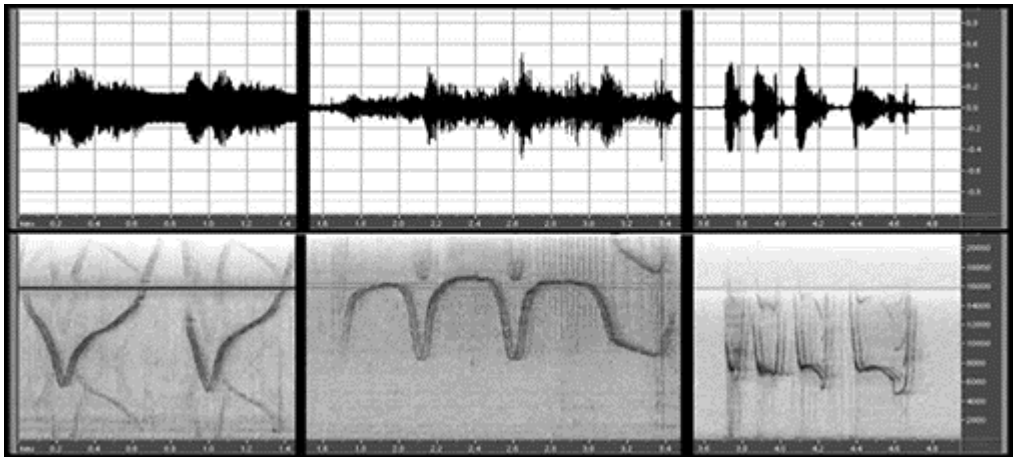


Figure 25. Characteristic whistles known as "signature whistles" of Puna, Kawai and Maka: three dolphins at Sea Life Park, Hawaii. Notice that the frequency is up at 16,000 Hz.

Whistles are used in communication. They are continuous, narrow-band, frequency-modulated pure tones, limited in frequency to the mid- and upper-range of the human sonic spectrum, generally from 4 to 24 kHz, and of 0.5 s in duration. Most of their energy is below 20 kHz.

When analyzing whistles, the spectrogram is a very useful tool and as one can see, much more informative than the amplitude/time graph above. Because they are so easy to visualize, researchers have been able to spot regularities in individual usage and presented the theory that every dolphin has its own signature whistle (see section 3.1).

The signature whistle hypothesis is still controversial and several papers have been written for [33] and against [34]. I have myself studied many hours of video recordings, analyzing dolphin vocalizations, and although a wealth of different whistles were observed, both individual regularities in whistle usage were found, as well as across individuals. Whistles are used in social context, and could be a good symbolic message unit candidate .

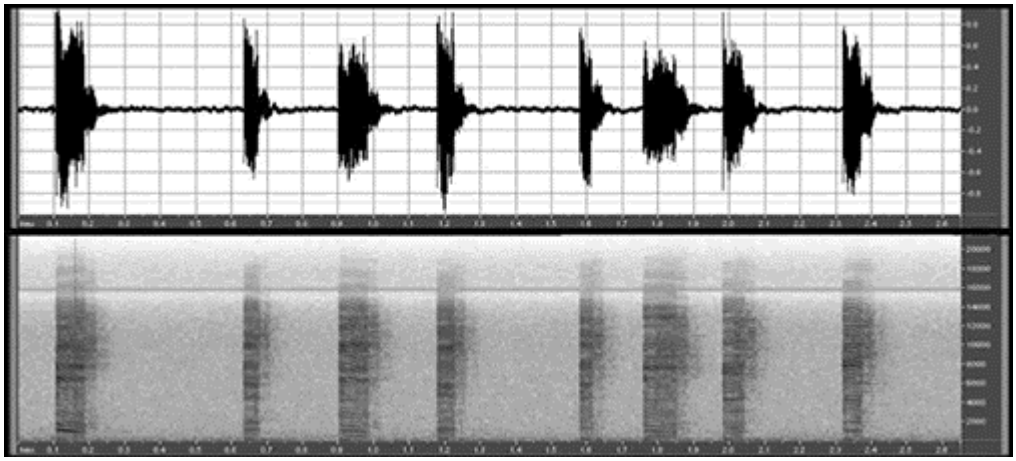


Figure 26. Dolphin pulses are used for communication, and will be used in the artificial language.

Pulse sounds occur commonly in social and emotional context, and are thought to be used for communication purposes. These are trains of clicks with repetition rates of up to 5,000 clicks per second. The high repetition rate gives a tonal quality to the sounds. Pulses are less informative in spectrograms than the whistles, and not so well documented. The spectrogram only goes up to 20,000 Hz and the signal continue well out of range of the picture. There are indications from recordings and personal experience that pulses are good SMU candidates. They are used more frequently during social activities than whistles, and have some advantageous classification properties. Although I will discuss using whistles, and how it would affect the bridge, pulses are the ones being tested in this thesis.

Dolphin vocalizations are crudely classified and still not very well understood. A fourth category is sometime used. This is mainly a bag for all the other kinds of sounds these animals are doing such as: squawks, squeals, snaps, cracks, bleats, barks, moans, and chirps. 'Chirps' are pulsed, frequency-modulated, broadband sounds. Frequency modulation occurs in different frequency bands. Chirps are also thought to be used in communication.

Further filtering

What we have so far, is an array of numbers that works as a numerical representation of the features of the waveform at a given time. In the example above, a straight Fourier transformation (FFT) would make an array with 128 values, each value being

the average amplitude at a given frequency. Digital signal processing involves a number of other algorithms to further process the signal, and it is not uncommon to let the signal go through several other filters before it is being presented to the classifier. This is often done in order to reduce the dimensionality of the data, i.e., if there is data in the resulting array that is not necessary to interpret the signal, it can be removed. This has been found by presenting speech-recordings to test subjects both with and without the filter. If the difference does not reduce the test subjects' ability to understand the meaning of the speech-recording, the filter is considered to be removing only unimportant or redundant data and hence effective. Perceptual linear prediction (PLP) and Mel scale cepstral analysis (MEL) are examples of this kind of psychophysically based transformations.

Another powerful feature extraction technique commonly used today is Linear predictive coding (LPC). It provides good quality at a low bit rate, and is relatively efficient for computation. In the examples in the next chapter, an LPC feature with 13 values is used instead of the FFT array with 128 values. Since one array is computed and fed to the classifier every 10 ms, the reduction in dimensionality will greatly improve the performance of the classifier.

6.3 Classification

The general idea is to take, let's say, seven of the pulses you saw in Figure 26 and assign them the roles of symbolic message units (SMU's). What we want the classifier to do then, is to recognize nine different classes; the seven different pulses as well as silence and garbage, i.e., other sounds that are neither any of the seven pulses nor silence. We want to make a model that is able to generalize; i.e., a slightly different signal should pass as being the same. If the model is too rigid, nothing except the original sample will be accepted. On the other hand, if the model is too loose, it will accept signals that shouldn't be accepted. To be able to generalize, it is necessary to know what it means for two signals to be similar.

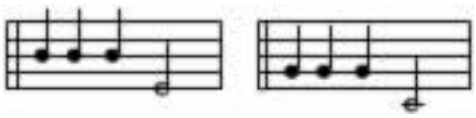
So how will the SMU's we have chosen be perceived? If we make a comparison with human perception for a moment, we can see that at least five alternative interpretations are possible:

1. Information lies in the resonance room, like our phonemes. Each phoneme has its own distinct characteristics. Pitch and length are not relevant features. A child or an adult, speaking fast or slow, are all interpreted as the same type.

2. Information lies in the pitch, like notes in our 12-tone scale. Either:

2.1. Absolute. The notes C and G are distinct even if they are alone. A person with 'absolute pitch' can hear a single tone and tell which tone it is.

2.2. Relative. Most people cannot tell if a tone is a C or a G, but they are still able to tell one song from another. The information is stored in the relationship between the notes.



The song is the same whether played as B-B-B-E or as G-G-G-C.

3. Information lies in the length of the tones/vowels, like Morse code. It doesn't matter whether the signal is in C or in G, or has the spectral characteristics of an O or an I. Either:

3.1. Absolute. The signal holds its information regardless of context.

3.2. Relative. The relationship between the signals, i.e., context decides. E.g., the first can be any length, as long as the other is twice as long.

We have to make a classification hypothesis as to what is important and then test it. One way would be a copycat experiment. If dolphins are able to copy a signal provided by us in such a way that the classifier would recognize what we anticipate, we have a working model. We should remember however, that copying a signal immediately after it's been presented does not necessarily mean the same as being able to recall the same signal from memory. If a human (without absolute pitch) were to do a copycat experiment, and the classification hypothesis was based on absolute pitch, he might be able to correctly sing in the right pitch immediately after, but if asked to repeat the sequence from memory fifteen minutes later, he might recall it in another pitch. Furthermore, if the sequence was presented in a Do-Re-Mi fashion, as "Mi-Mi-Mi-Do", the test subject could also just have remembered the words and not the tonality of the signal (as category 1 above). Our classification would fail even if the test subject did recall the sequence in such a fashion that enough information could have been extracted, to correctly classify it, had we used another classification hypothesis.

General Problems with Classifying

Classifying data involves a decision on what to classify by. For example, let's look at how things are organized in a supermarket. They could have used color as a classifier, putting all red things in isle 4 and blue things in isle 5. Or put everything made of metal in one place and things made of plastic in another. They could classify things alphabetically, after price, size, country of origin and so on, but they don't. Supermarket managers put great effort in grouping things in such a way that customers will find what they're looking for (and buy as much as possible). Partly they

classify by geographic location in the customer's home. Things you use in the bathroom will be grouped together, but things are also classified by functionality: sporting goods together, or by user: children's things together, etc. Certain items can be troublesome because they have conflicting properties. Where do they place a jar of olives? By the fruits? Together with the pizza condiments? In the Gin & Tonic department? In the snacks department?

It is sometimes difficult not to get conflicting data when you make a model. It's like a bad IQ test, where you are asked to classify something, and there are several good alternatives.

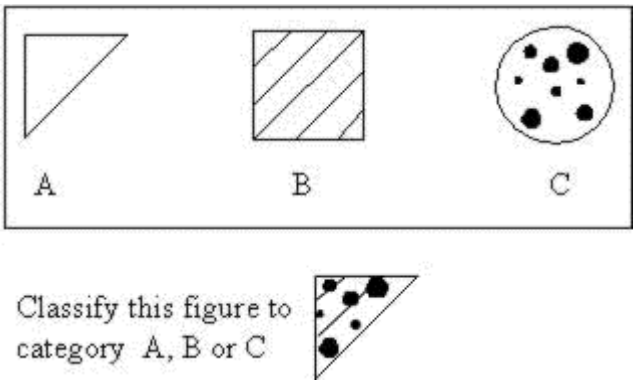


Figure 27. There is often more than one good alternative to a classification task.

Say for example, we record seven different people singing one tone each of "Do Re Mi Fa Sol La Si" in their right tone, respectively. Let person number one say "Do" in C, and call it category 1. Let number two say "Re" in D, and so on. This represents the categories 1-7. Then we let person number one say "Fa", but in D, and ask "Which of the seven categories does this one belong to?" One might argue that it belongs to category 1, because it was spoken by person number one. One might argue that it belongs to category 2, because the tone was a D. And one might argue that it belongs to category 4, because the utterance was "Fa".

All answers are equally right in their own way, and so the interrogator will need to be more specific in order to elude the ambiguity in the question. We encounter similar problems of ambiguity when we try to design an audio classifier. What are we trying to classify in speech recognition?

In standard speech recognition, pitch is not an important feature, since different speakers, a child, woman, or man can say something that should be classified as the same thing. Speech is also (partly) independent of the length of the constituents. 'Speech' and 'Speeeeech' should be classified as the same thing. In other instances, these features might be something to pay attention to. This is something we have to decide when you make our model. The way a domain is represented has a huge bearing on the way it is 'understood'. We have to choose the salient features that

should represent the domain and sometimes, as in our example, we don't really know what the best way to represent the domain is.

6.4 Outline of Standard Speech Recognition

There are a number of different approaches to speech recognition. I will give an outline of one, called frame-based recognition, which is quite common. It consists of a back-propagation neural network (BPN) as a phonetic probability estimator, and a Viterbi search as a word classifier.

The steps involved in making a numerical representation of the speech signal, are identical to what is outlined in section 6.1 and 6.2. We digitize the signal, divide the waveform into frames, and compute features for each frame. We then feed this frame, and a small number of surrounding frames as context, to a BPN. Please refer to sect 2.3 for a general outline of neural networks.

The steps involved in training a BPN to recognize phonemes are as follows:

- Specify the phonetic categories that the network should recognize.

A general-purpose phoneme estimator network must have one output node for each phoneme in the target language (in English 40), and two additional nodes for silence and garbage. (Garbage being any sound that is not speech.) If we however, only want to spot the words 'Yes' and 'No', we only need seven output nodes.

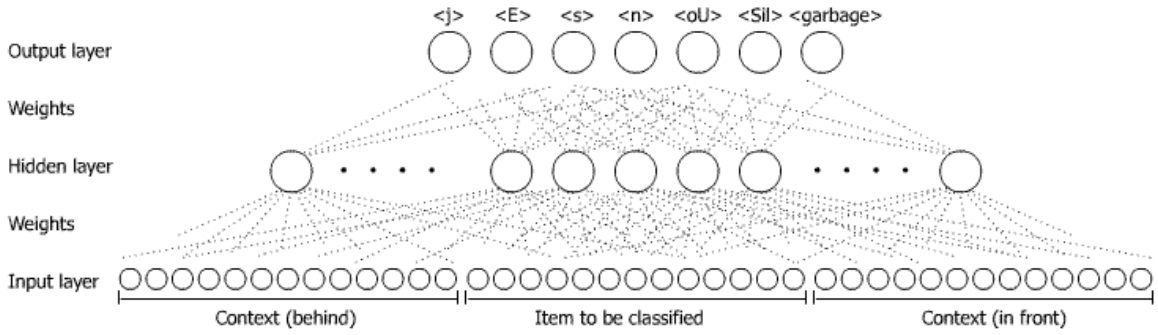


Figure 28. A BPN for classifying 'Yes' and 'No'.

As input, an LPC feature extraction as described in section 6.2, gives us an array of 13 features for a 10ms chunk of audio. It has been shown that the performance of the classifier is greatly improved by adding some context together with the input data. Our network will then have 13 X 3 = 39 input nodes.

- Find many samples of each of these categories in the speech data.

We must manually transcribe all this speech data in order to train the network. If we don't have enough variation in the training data, the neural net will be too rigid, and unable to correctly classify other speech than what it was trained on.

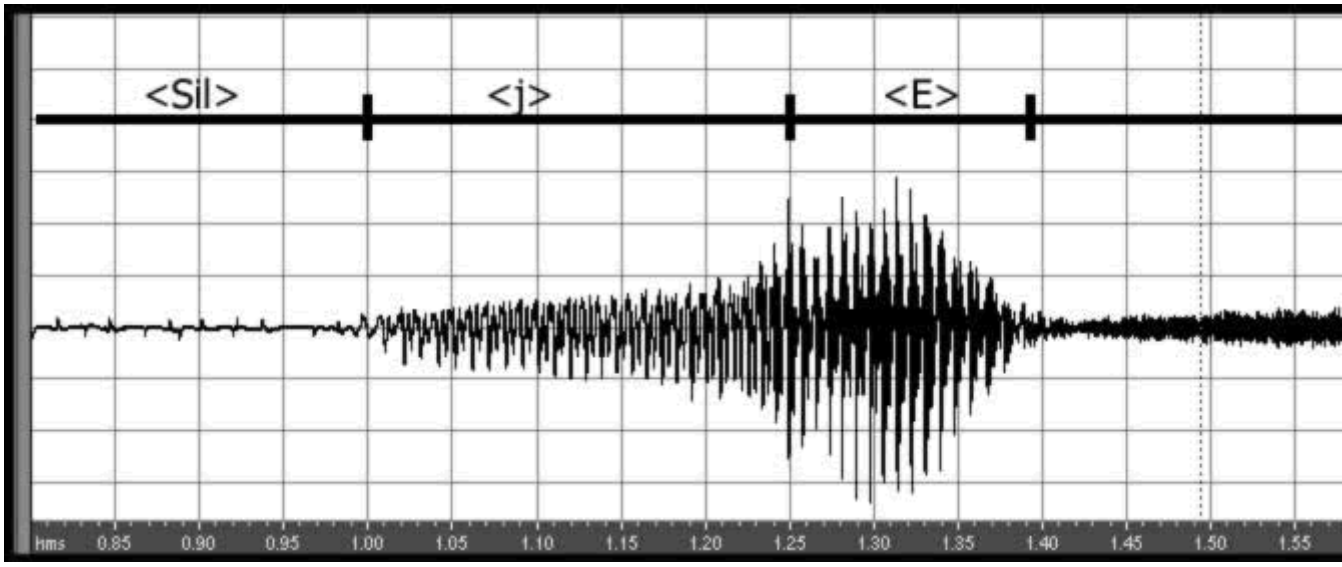


Figure 29. Manually transcribed speech sample of 'Yes' to train the network on.

- Train the network to recognize these categories with the transcribed samples.
- Evaluate the performance of the network by using a test set of transcribed speech that the network was not trained on.

The outputs of the neural network are not used as answers, but as estimates of the probability that the current frame contains that category. Every output node will give a rating of the incoming data, and if the network is properly trained, the node corresponding with the actual utterance will give the highest rating. This way a matrix (two-dimensional array) is created that is presented to a Viterbi search that will find the word(s).

Viterbi Search

A Viterbi search uses the matrix of probabilities and a set of pronunciation models to determine the most likely word(s) by finding the highest scoring path.

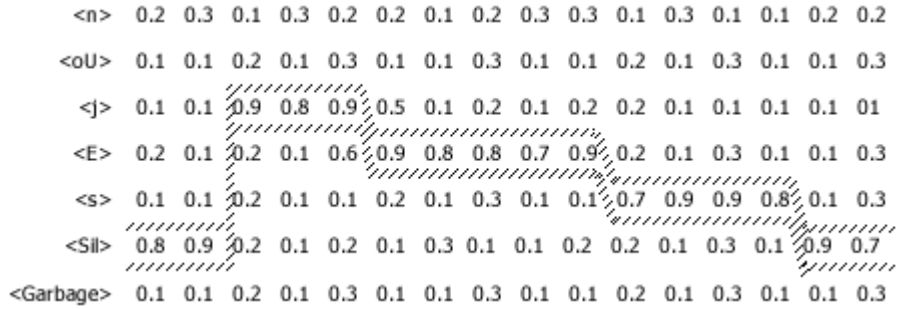


Figure 30. A Viterbi search through a matrix of probabilities.

The set of pronunciation models helps confine the search by not allowing transitions from one phoneme to another which is not found in the dictionary. If, for example, the <s> in the sixth column would get a higher rating than the <E>, the search would not go there, because we have no word transition from <j> to <s> in our model. There are still many possible paths through the matrix, but for any particular phoneme in any given word, there can only be one path to the end that will maximize the score. The Viterbi algorithm takes advantage of this insight to increase the efficiency of the search. If we see two paths come together at the same phoneme, in the same word at the same time, we can discard the path with the lowest score. Since the two paths will behave identically from that point on, we know the lower scoring path will never catch up.

Difficulties Due to Lack of Samples

The task of classifying the chosen dolphin vocalizations, differs from regular speech recognition in several ways. It is easier than regular speech recognition for several reasons. Firstly, because we have chosen to have only a few SMU's. Partly because it would be hard to find 40-50 different ones, and partly to make the classification task easier. (Recall the discussion in section 5.1, about the arbitrariness of the number of symbolic message units needed to code a language.) Secondly, we don't have so many words in the dictionary. Our restricted domain language of >1,500 words is so small that a simple lookup table will do, whereas a normal ASR will have 50-100,000 words to choose from.

On the other hand, it is harder than regular speech recognition, because we can not take for granted that the same psychophysically based transformations that our feature extractions are based on, are valid in the dolphin domain. We must also keep in mind that the hearing range of dolphins is ten times that of humans, and that we are forced not to use a sample rate high enough to cover this.

Most seriously however, is the fact that we don't have a large amount of samples to train the network with. Both back-propagation networks and Hidden Markov Models (another popular technique not discussed here) need many samples of each category in order to work. This makes it impossible to use a standard speech recognition approach, and we have to look at alternative classification techniques.

Dimensionality

We often talk of dimensionality in classification tasks, and people who are not used to it, could get a bit set back when they hear of something having more than three dimensions. If we are to describe the location of an object in space, we can do this by giving its x, y and z coordinates. We can place this information in an array of length 3, and represent it as [13, 28, 16], for example.

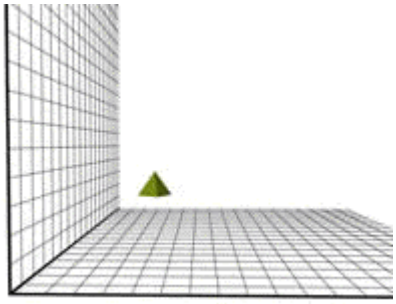


Figure 31. X,y,z coordinates describe the location of an object in space.

This conforms to our normal understanding of the universe having three dimensions. If however, we want to describe an object that moves in this 3-D space, we need to add time as a fourth dimension and for example say [1, 13, 28, 16]; meaning: at time=1 the object was located at [13, 28, 16]. This becomes a little more difficult to visualize, but we can for example make a sequence of 3-D projections and somehow keep the fourth dimension in our mind as the difference between the frames.

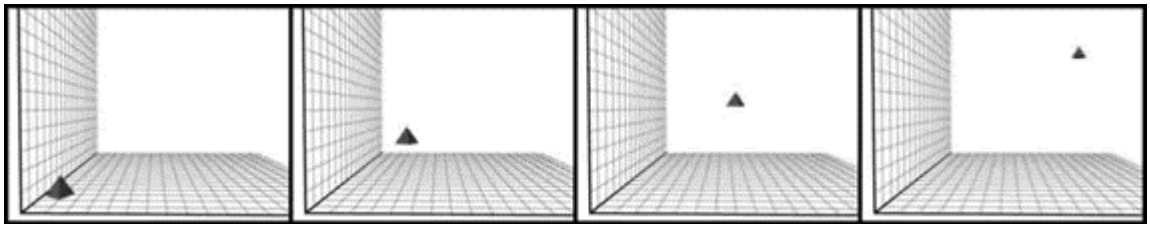


Figure 32. Time is represented as a sequence of three-dimensional images.

Depending on the characteristics of the item we want to describe, we might however have more than four dimensions and the visualization becomes a problem. Taste can for example be represented in five dimensions. We have five basic kinds of taste buds: sweet, sour, salt, bitter, and umami (savoriness). If we, to simplify things, say that each of these taste buds can send out four values (i.e. no salt, a little salty, medium salty and very salty), then every taste sensation can be described as an array of five elements holding a value of 0-4. Pure salt will, for example, be described as [0,0,4,0,0]. We can then talk of cinnamon holding a particular position in this five-dimensional space, but we can not make a visual representation of it. The LPC feature extraction we created from our digitized sound in section 6.2 can be said to be an object with 13 dimensions. Any sound token will have a location in a 13-dimensional space. The classification task then becomes to give labels to areas in this space. This can be done by for example a self-organizing feature map (SOM).

6.5 Self-Organizing Feature Maps

One way to describe the self-organizing feature map (SOM), is to say that it is compressing high-dimensional data into a low-dimensional representation. What this means, is that it finds a relationship among the data in a multi-dimensional space, and can represent this relationship in two dimensions. It is an algorithm both for interpreting and visualizing complex data sets. The SOM has a different architecture than the back-propagation network, and uses a different algorithm for adjusting the weights. It is called an unsupervised neural net, because it corrects the weights of the network without the need of manually classified samples, as is needed with a BPN. There is only one input layer and one output layer. The output layer is usually laid out as a sheet in two dimensions, and often referred to as a Codebook. SOM's have the ability to construct topology-preserving mappings of the kind that are expected to happen also in the mammalian cortex. [35]

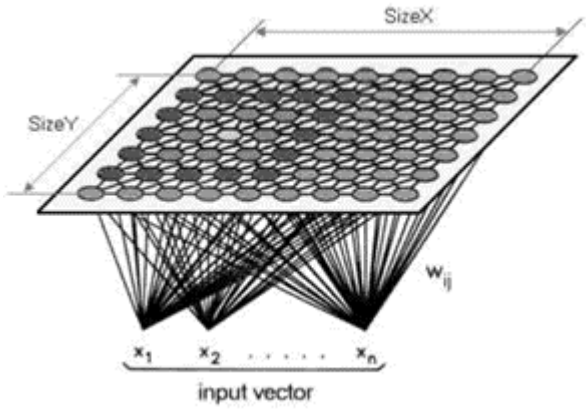


Figure 33. A two-dimensional representation of multi-dimensional data.

All the nodes in the input layer get connected with all the nodes in the output layer, and the weights are given random start values. An array with the same length as the number of input nodes, is fed to the input nodes. After some mathematical operations are performed, where the weights are multiplied with the values in the array, one of the output nodes will get the value closest to the value in the array, and be the winner. It gets reinforced, which means its values will be adjusted slightly towards the input values. Also its neighbors gets reinforced by a Mexican hat function, so that the closest neighbors gets rewarded, while neighbors a little bit further away gets punished.

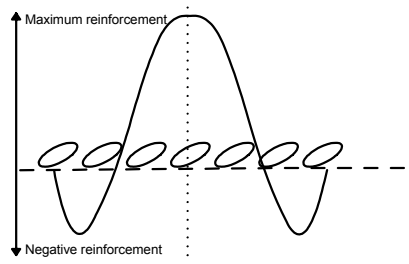


Figure 34. A Mexican hat function distributes the reinforcement.

A new set of input values is fed to the network, and the process is repeated. After letting all the data go through the network many times, a pattern will emerge where similar data will be grouped together in clusters.

This is a very useful tool in cases where we don't know what to make of our data. We may, for example, have lots of data from the control room of a paper-processing plant. Every gauge or instrument will show some kind of measure, but it's hard to get an overall impression of how they all affect the process. We collect this information in the form of an array, and continually feed it to the SOM. If we monitor the paper plant through different phases where some part of the process goes wrong, say the boilers are overheating, or the pulp gets the wrong texture, these different states will give rise to clusters in the SOM which can later be labeled and give us a two-dimensional map of the state of the plant.

This way SOM's can be used as decision-support systems in a wide range of businesses. The financial health of banks have been measured and later predicted in a similar fashion, by analyzing and visualizing sets of statistical indicators in a SOM. The SOM has also been used to classify speech into phonetic groups. Instead of being fed data from a paper plant or a bank, a feature extraction like the one described in section 6.2 is done, and an array of audio features serves as the input.

A SOM could be used in the hunt for samples to use as SMU's in our language scheme. By visual or auditory inspection, we are not able to tell if sample A is of the same type as sample B. Say we recorded a lot of dolphin pulses and fed this to a SOM. Similar input would group itself into clusters, which could then be used to label the data. The process would go in three steps. First we train the SOM with all our available data. Once the SOM is trained we run the data through it again. This time we measure and mark out which nodes were most active, and see if natural candidates will appear in the form of clusters with many hits. We decide which nodes or groups of nodes to use, and label these, for example, 'Do, Re, Mi, Fa, So, La, Ti' (or 0-9). We then let the data through the SOM a third time. This time together with a procedure that picks out the data entries that correspond with a hit on a labeled node, and saves these in a separate file together with its new labels. We will also save the rest of the data together with the label 'garbage' or 'silence' to have as many samples as possible for those categories too. We now have a file with labeled samples that we can use to train a specialized SMU classifier such as LVQ.

6.6 Learning Vector Quantization (LVQ)

LVQ and SOM have, apart from being created by the same person, several things in common. The LVQ architecture is the same as the SOM, having all the input vectors connected with all the cells in the codebook (see Figure 33) Similarly, their learning algorithms perform two important things:

- Clustering of the input data.
- Spatial ordering of the map so that similar input patterns tend to end up close to each other on the map.

The Learning Vector Quantization is however, an algorithm for *supervised* classification. In contrast to the SOM, the LVQ is working on labeled data. If labeled data is available, the LVQ algorithm is preferred because it takes advantage of the extra knowledge that the labels give. During initialization of the codebook, it is possible to survey the available data and assign an equal number of nodes to each class. It is also possible to place special attention to the borders between different classes, and hence arrange the nodes so that the number of misclassifications is reduced.

It is also possible to use the same codebook for both algorithms, and sometimes it is advantageous to first use the SOM algorithm, and later fine-tune the positions of the cells using the LVQ algorithm. [36]

6.7 The Time Domain

The kind of classification tasks discussed so far, has not been concerned with context, i.e., change in time. A different aspect of classification, is considering sequences of patterns. If a classifier is to distinguish between a rising and a falling 'i' for example, it is not enough to go through the steps described so far. The LVQ classifier described above might correctly classify each frame as an 'i', but would not be able to tell anything about the bigger picture, i.e., is this frame part of a riser or a faller?

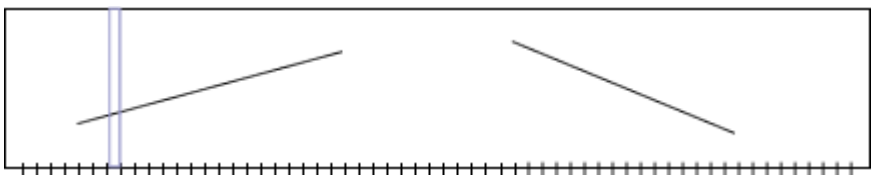


Figure 35. Frame by frame classification can not tell anything about the context that the frame belongs to.

Looking at change over time as something to pay attention to, is to focus on the process instead of on the state of things. The object has to be viewed from the right

distance, for the pattern to be seen. We might call it the classifier's 'attention span'. Time sequences can be anything from a split of a second – as the on-off oscillation in a light bulb, to millions of years – as in the change of galactic constellations. Human perception is suffering from (alternatively, taking advantage of) this. When the change is too fast or too slow, we fail to see it. Attention span is something not well enough understood in the human mind, to correctly model it. We might however create simplified versions of it. This can be done by something called 'template matching', where we create a matrix of sets of features and compare our input data to a set of such matrixes. We need to distinguish between two kinds: fixed length template matching, and varying length template matching. Humans are, within certain limits, quite adept at the latter, which is a harder classification task.

The LVQ classifier described above, could be made to incorporate context. Features from some frames behind, and some frames in front of the frame to be classified, could be incorporated in the input array in a similar fashion as described in section 6.4. The architecture of these ANN's are, however, dependent on having a fixed set of input nodes, and hence view a fixed length of time. The problem is then, that real-life input never comes in fixed lengths.

Attempts have been made to incorporate flexible time input in ANN's, but it is troublesome, and outside the scope of this thesis. I will describe one solution that works for small sets of templates, called 'Dynamic Time Warping'.

6.8 Dynamic Time Warping

Dynamic Time Warping (DTW) is used in template matching to allow for the incoming word of speech and the template to be of different length. The idea is to calculate a numerical difference (Euclidean distance) between two features, and in that way be able to compare them.

Two types of distance is calculated:

1. Local: A computational difference between two features: the incoming feature and a template feature.

2. Global: The overall computational difference between an entire signal and a template (of possibly different length).

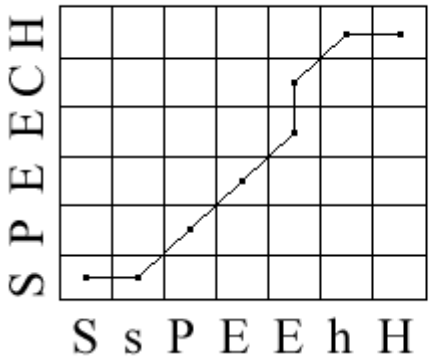


Figure 36. Illustration of a time alignment between a template (SPEECH) and input signal (SsPEEhH).

The task consists of finding the best-matching (i.e., lowest global distance) path between an input signal and a template. Evaluating all possible paths would be possible, but very inefficient as the number of paths grows exponentially with the length of the signal. The DTW algorithm is similar to the Viterbi search described in 6.4, in that it takes advantage of some constraints we can impose on the search which makes it much more efficient:

- Matching paths cannot go backwards in time
- Every frame in the input must be used in a matching path
- Local distance scores are combined by adding to give a global distance

The shortest path is in this way calculated for every template, and the global scores are then compared. The template with the best match is the word estimate. A threshold can be put to reject the signal if the score is too low (i.e., the incoming signal was not one of the words in our dictionary).

This technique works well if the number of templates is small. A 0-9 digit recognizer can for example, be implemented in a DTW. But since it has to go through all the templates for every incoming signal, the algorithm suffers when the number of templates grows. It takes too long to use this technique in a bigger lexicon.

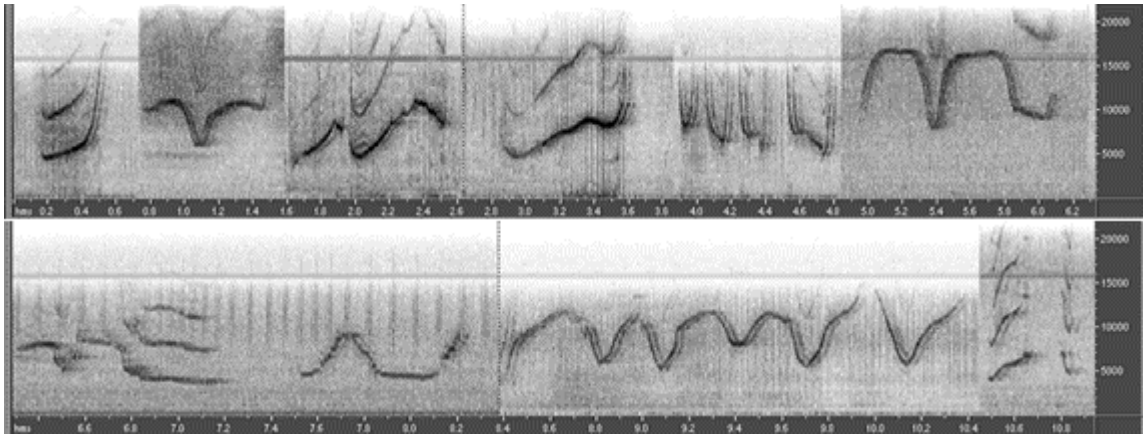


Figure 37. A spectrogram of 10 different dolphin whistles. Dynamic Time Warping could be used to recognize them if whistles were used as symbolic message units in the language.

If the artificial language for this thesis would use dolphin whistles as SMU's instead of pulses, a DTW would be a possibility. It has the advantage of working well without many samples and could, as with the LVQ, make a template from one sample. The DTW in only mentioned here and not implemented. It would be advantageous to have several kinds of classifiers implemented in the software, but that is outside the scope of this thesis.

– 7 –
General Audio Pattern Recognizer (GAPR)

GAPR (pronounced 'gapper' as in someone who crosses a gap) is an attempt to implement the interface aspects of the bridge that I have described in this thesis.

As shown in Figure 11 on page 47 there are in my opinion three equally important requirements for language acquisition. Access to a linguistic system; a way to store and process at least a subset of this system (e.g., a brain); and input/output devices. Dolphin sound production is incompatible with human linguistic sounds. It is not possible to use the dolphin organs to make human sounds, and it is not possible to use the human organs to make dolphin sounds. This interface problem must be addressed and overcome somehow if two-way communication between humans and dolphins is ever to take place. The General Audio Pattern Recognizer (GAPR) – is a software that bridges this gap. It allows humans to make, and recognize dolphin sounds.

In general, GAPR is a tool for classifying and translating audio patterns into symbolic representations, and for linking combinations of these symbols to words. GAPR can be used both for setting up a language structure, and to run an (almost) real-time translation for dialog or game. It should be seen as a prototype with limitations, which will also be described. It is written in Tcl/Tk, a scripting language well suited for prototyping.

Symbolic message units (SMU's) are being used as a kind of 'Interlingua' i.e. a neutral intermediate representation of a word, that can be turned into dolphin friendly, or human friendly sounds and words. Several instances of GAPR can be linked up with each other, in a server–client relationship, using the Tcp/Ip protocol. The different instances are linked up either locally on the same machine, on separate machines in a local area network (LAN), or through the Internet. SMU's are what is being passed over the network, and each instance of GAPR will decide what to do with the SMU's depending on what mode of operation that particular instance is in.

With a return to where the thesis started, an overview of how intentions are being transferred from a sender to a receiver, and GAPR's position in the flow of intentions can be described as below.

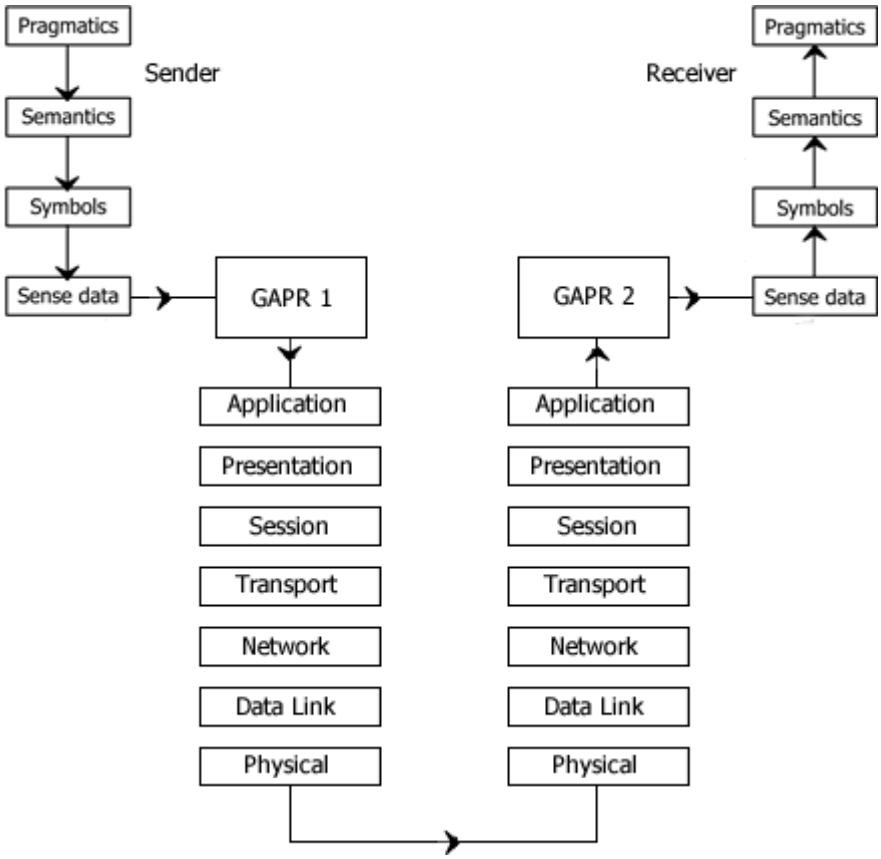


Figure 38. Intentions have a long way to go. Top part is the Fregian internal bridges described in the beginning of the thesis (section 1.2). Middle part is two instances of GAPR. Bottom part is the OSI protocol of data transfer over the Internet.

Starting with the internal gaps of Frege (section 1.2) GAPR-1 is receiving sense data from the sender, transforming it into SMU's, packing it into Tcp/Ip packages, and sending it over a network using the OSI protocol. The process is mirrored on the receiving end where GAPR-2 is receiving Tcp/Ip packages and unpacking them, converting SMU's into sense data suited for the target individual on the receiving end, who in turn hears a set sounds that with internal processing hopefully gets interpreted as the same message as was intended from the sender.

Overview

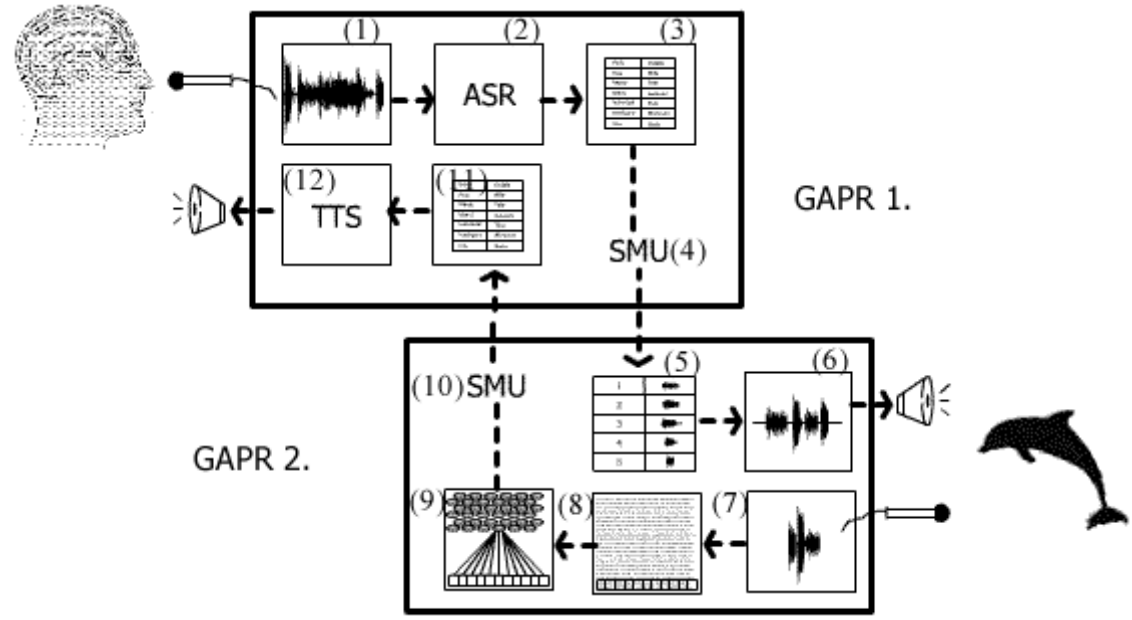


Figure 39. Schematic drawing of how data flows through GAPR in Human-Dolphin dialog mode.

---------- Human sender – Dolphin receiver ----------

1 and 2: Recording and automatic speech recognition (ASR) of human speech. This is not part of the current implementation of GAPR. It is 'off the shelf' software that will help to make GAPR more user-friendly, but is not essential for the functionality. Current version uses keyboard input of text, or selection from the menu of available words from the Dictionary Tab.

3: Every English word in the Dictionary Tab will have a corresponding value in the SMU column. These can be collected into sentences before they are sent away.

4: The Socket Tab will have a record of which port and channel is connected, and will send the SMU's on this channel.

5: A lookup table linking each SMU with a corresponding sound file with a dolphin pulse sound that is loaded into RAM.

6: Concatenating (joining together) waveforms from the sound files into a string of 'words' that is being played to the dolphin.

---------- Dolphin sender – Human receiver ----------

7: Dolphin vocalizations are being recorded.

8: A feature extraction of the recorded waveform is created.

9: A classifier (LVQ) receives features and turns them into SMU's.

10: The Socket manager sends the SMU's over the network to the human side.

11: The Dictionary lookup table converts SMU string to English words.

12: Text to speech (TTS) engine or text on the monitor tells the human what is said.

7.1 Creating a Bridge

The steps involved in setting up a structure in order to use GAPR for dialog will be described in the following section. Briefly the steps are:

- Record the audio-patterns you wish to recognize as the basic symbolic message units (SMU's)
- Assign some sort of label to them
- Decide on which kind of feature extraction to use
- Initialize and train the LVQ classifier on these SMU's
- Add the words you wish to have in the language
- Give each word a unique combination of SMU's as a 'word' in the target language

The Audio Tab

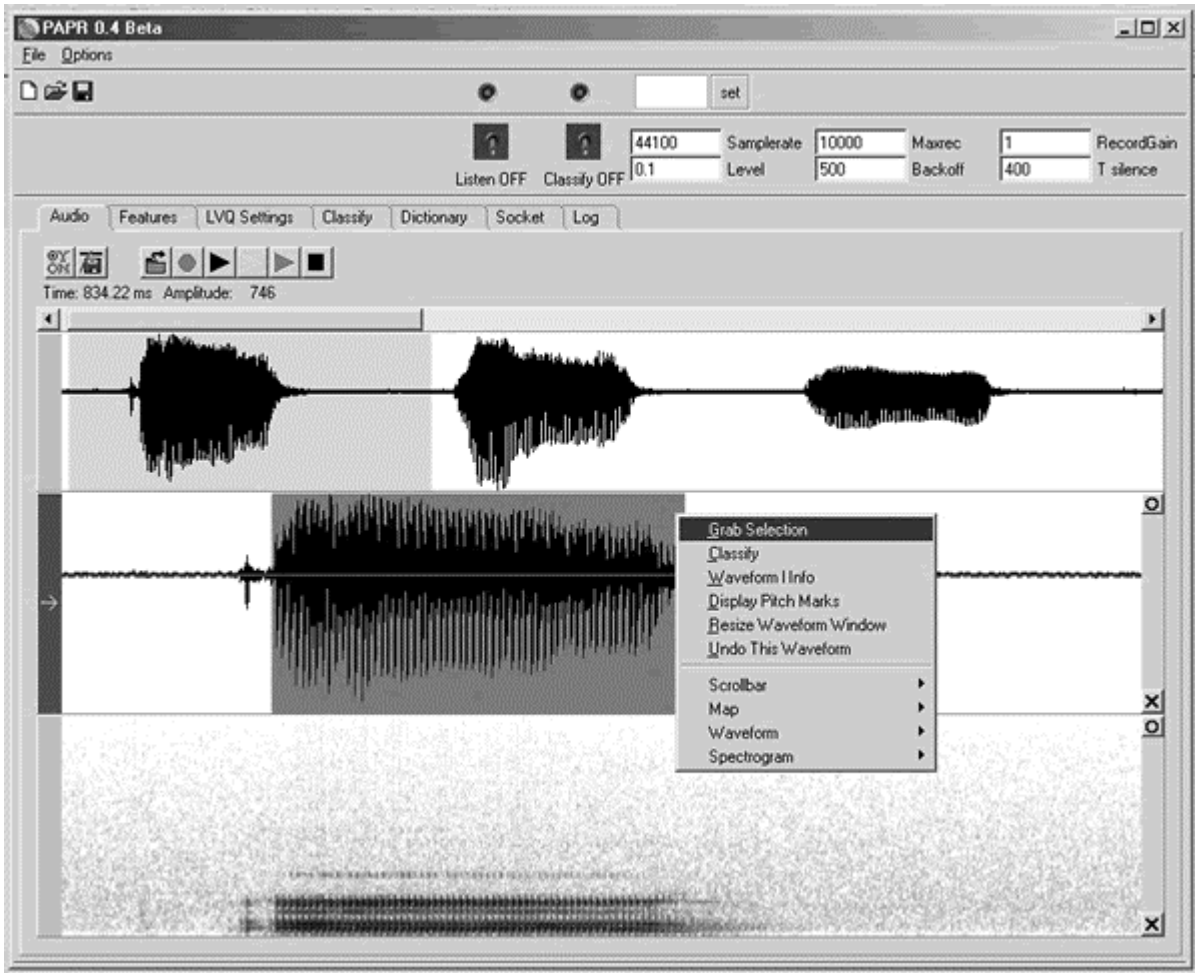


Figure 40. A full screen of GAPR's Audio Tab: Getting ready to grab a selection of a recording of the three vowels a, i, and o.

The Audio Tab is used to collect the audio patterns we want to train our classifier on. The top frame is a map of the entire waveform that is loaded or recorded. The middle frame is a zoomed-in sub-selection of the top frame, and the bottom frame is a spectrogram of the same selection. I will go through the different steps of creating a project, with a simplified set, in order to give an easily digestible introduction to the program. Recall that contrary to regular speech recognition, this work is based on the assumption that we don't have many samples of the same type to train our classifier with. A visual inspection provided by an amplitude/time graph, or a spectrogram, will not give enough information to tell the samples apart. Neither will an acoustic inspection, as it will with human speech. We have a recording of samples that we have decided to use as symbolic message units, one sample of each SMU. In this example it is a recording of three vowels: a, i, and o. After selecting an appropriate part of the sub-selection in the middle frame, right-clicking will open up a menu as shown in Figure 40. Choosing the first item will open up the Grab Selection Dialog, where a label can be given to the selection.

Figure 41. The Grab Selection dialog.

The label serves as a peg for the classifier, and is as arbitrary as the relationship between a sound and its reference (meaning) in a word. I.e., we could have given the audio pattern any label, (for example "SMU1") but for clarity in the example, the vowel 'a' is given the label 'a', and so on. The selection and labeling process is repeated for all SMU's, and in addition a selection of silence (ambient noise) is chosen and labeled 's'. A more robust model would also have other, clearly different, audio patterns grabbed and labeled as garbage, meaning: sound that is neither a chosen SMU, nor silence. It will suffice to mention it here, and leave it out, to keep the model simple. In this example we end up with four files: a.pwf, i.pwf, o.pwf, and s.pwf.

The Features Tab

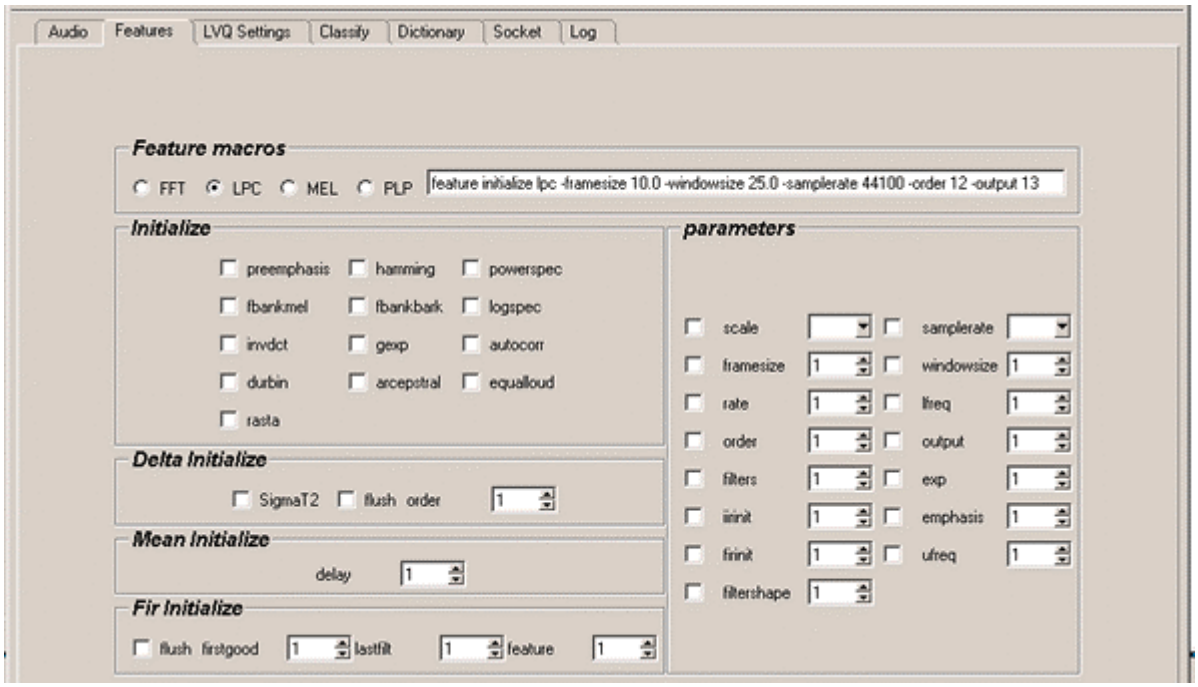


Figure 42. The features Tab offers the possibility to experiment with different settings on the feature extraction.

The Features Tab is a front end to the Features package, provided by the CSLU toolkit. As described in sect 6.2, most pattern recognition tasks are preceded by a pre-processing transformation that extracts invariant features from the raw data. For acoustic signals, this would be some spectral components. Selection of a proper pre-processing transformation requires careful considerations, but no general rule can be given. Since we don't really know what the salient features of our audio patterns are, the Features Tab offers the possibility to experiment with different settings on the feature extraction. Depending on the situation we might be interested in different features. In general, this part can be left untouched, but if the classifier is doing a bad job, there is the possibility that a different kind of feature extraction will give a better result. We will stick with the default settings, and see how the classifier does. The default setting is a Linear Predictive Coding (LPC) creating a 13 dimensional array with a sample rate of 44,100, and a frame size of 10 ms. By pressing OK in the Grab Selection dialog (Figure 41), a feature extraction of the vowel sound 'a' in figure 1 together with its label, is stored in the file a.pwf. The first few lines of the file look like this:

```
13
# feature initialize lpc -framesize 10.0 -windowsize 25.0 -samplerate
44100 -order 12 -output 13
-0.260753 0.298011 -0.145685 -0.051231 0.078632 0.577027 -0.359537 -
0.220445 -0.451845 -0.127602 0.396726 0.719543 0.252523 a
-1.298124 0.710236 0.500421 0.866782 0.710096 0.250810 -0.850528 -0.504889
-0.280059 0.229634 -0.068467 0.186041 0.123383 a
-3.548136 1.373249 0.951680 0.890689 0.467577 -0.127909 -0.521988 -
0.476816 -1.282958 -0.143639 0.244442 0.156188 0.261656 a
```

This file together with similar files for i, o, and silence, is what is being used to train the LVQ classifier.

The LVQ Tab

The LVQ Tab is a front end to the Learning Vector Quantization Program Package (LVQPak). All the .pwf files created from the recording in the Audio Tab are being collected into one file (.dat), which is being used as training data to create a Codebook file (See section 6.6). We decide on the number of codebook vectors (noc) we want, and initialize the codebook.

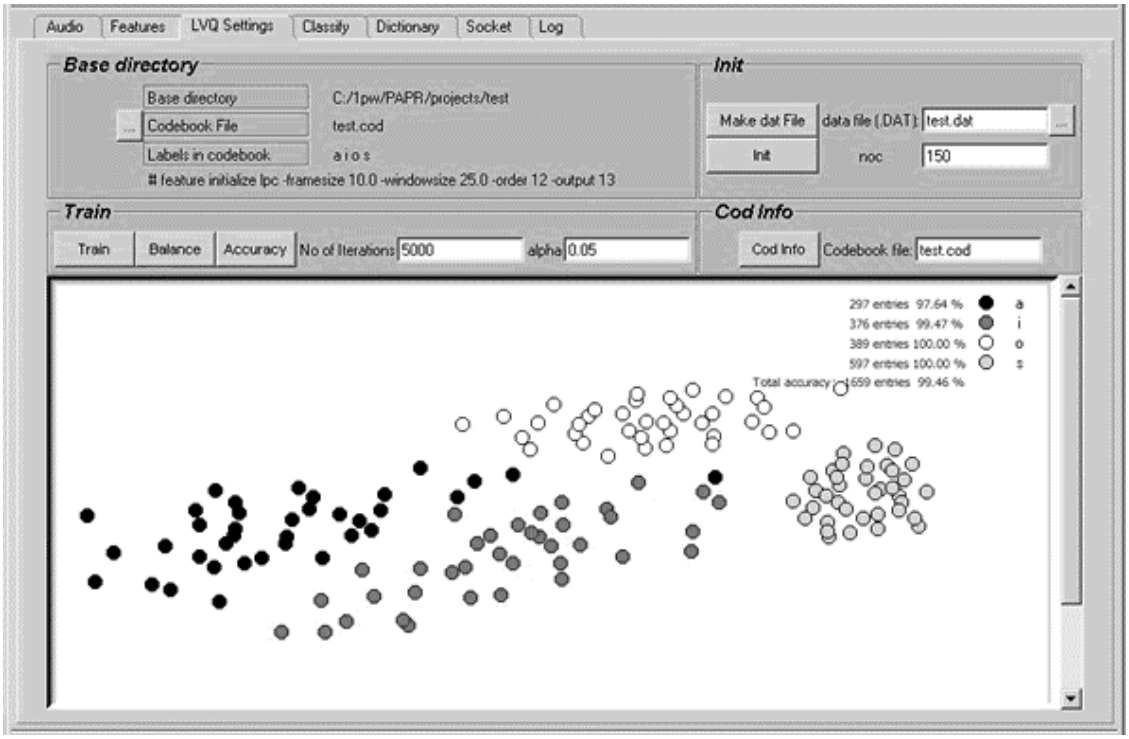


Figure 43. The LVQ Tab showing four distinct clusters in the codebook created from the test data.

The initialization procedure consists of distributing the nodes in the codebook evenly by checking how many classes are in the training data. With for example, 100 nodes and four classes: a, i, o, and s, every class gets 25 nodes. It is also giving the codebook vectors some initial values picked up from the training data, and runs one round of training. The canvas in Figure 43 shows a visualization of how the nodes are distributed after the initialization.

The second stage is to train the codebook, using the LVQ algorithm. The length of the training run (number of iterations) and the learning rate (alpha) has to be decided first. As a general rule, 30 to 50 times the total number of codebook vectors is when the codebook reaches its optimum. As with any neural-network algorithm, the LVQ can over-learn and thereby reduce its accuracy. If the codebook becomes too

specialized on the training data, the algorithm's ability to generalize for new data will suffer from that.

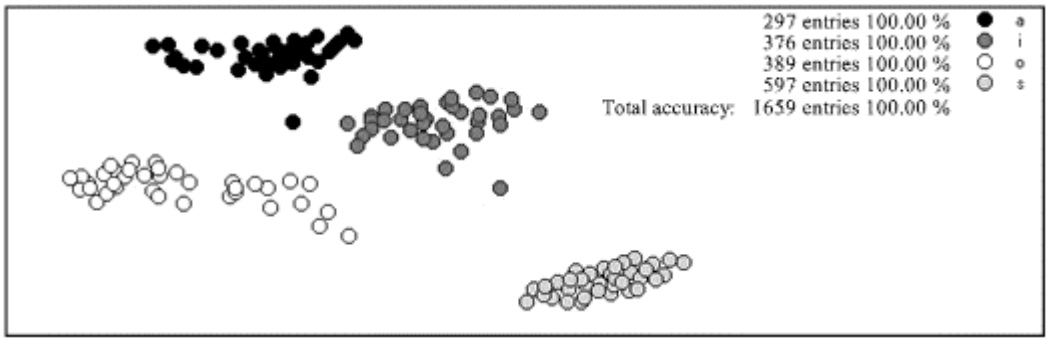


Figure 44. The codebook after training.

The Codebook is now trained and it is time to test it.

Classify Tab

Recording an "a" in the Audio Tab, right-clicking and selecting classify, will send the recording through a feature extraction and onto the classifier.

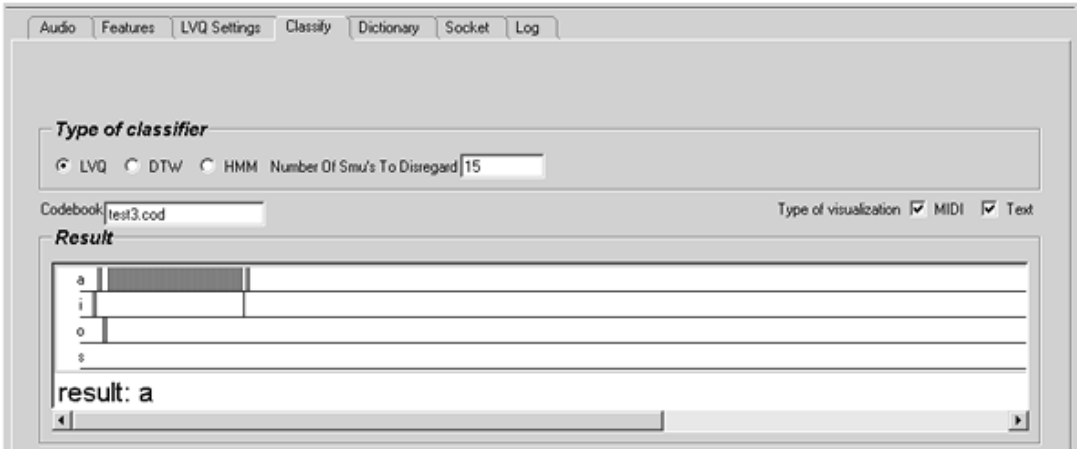


Figure 45. A visual representation of the classification.

The result from the classifier is visualized by a notation similar to a MIDI sequencer. Every class in the codebook gets its own line. Every token of sound (10 ms) being returned by the classifier is represented by a vertical bar on the line it has been classified as belonging to. This means that if every token was correctly classified, an 'a' lasting one second (1000 ms) would be represented as 100 vertical bars on the first line.

Before the result can be passed on to the receiver of the signal, it needs to run through a filter for two reasons. First, in order to reduce the many consecutive symbols to one; and second, in order to remove occasional misclassifications. This is done by letting the stream of symbols run through a pair of regular expressions removing tokens according to a pair of rules saying something like:

1. If there are less than X (in the picture set to 15) consecutive letters of the same kind, we consider them errors and remove them.
2. If the next letter is the same as the last one, it belongs to the same SMU and they can be reduced to one.

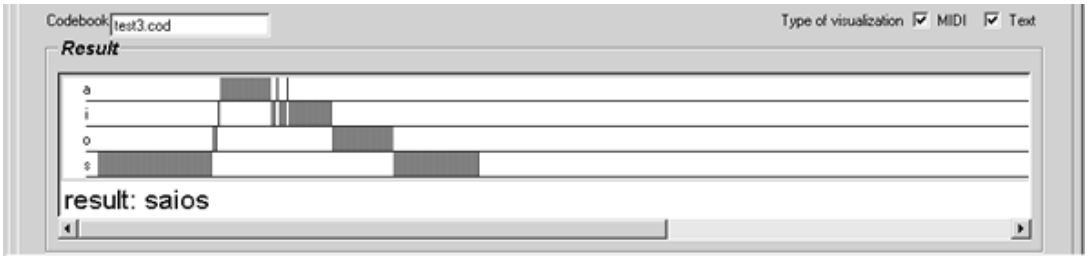


Figure 46. screen shot of GAPR recognizing a combination of vowels forming a possible word.

The resulting string from the classifier in the example above is "silence a i o silence". This combination of sounds could be assigned a meaning in an artificial language. What is needed then, is a dictionary for all the words we want to have in the artificial language.

Dictionary Tab

Depending on its use, a restricted domain language could have anything from two-three words, up to 100,000 words or more.

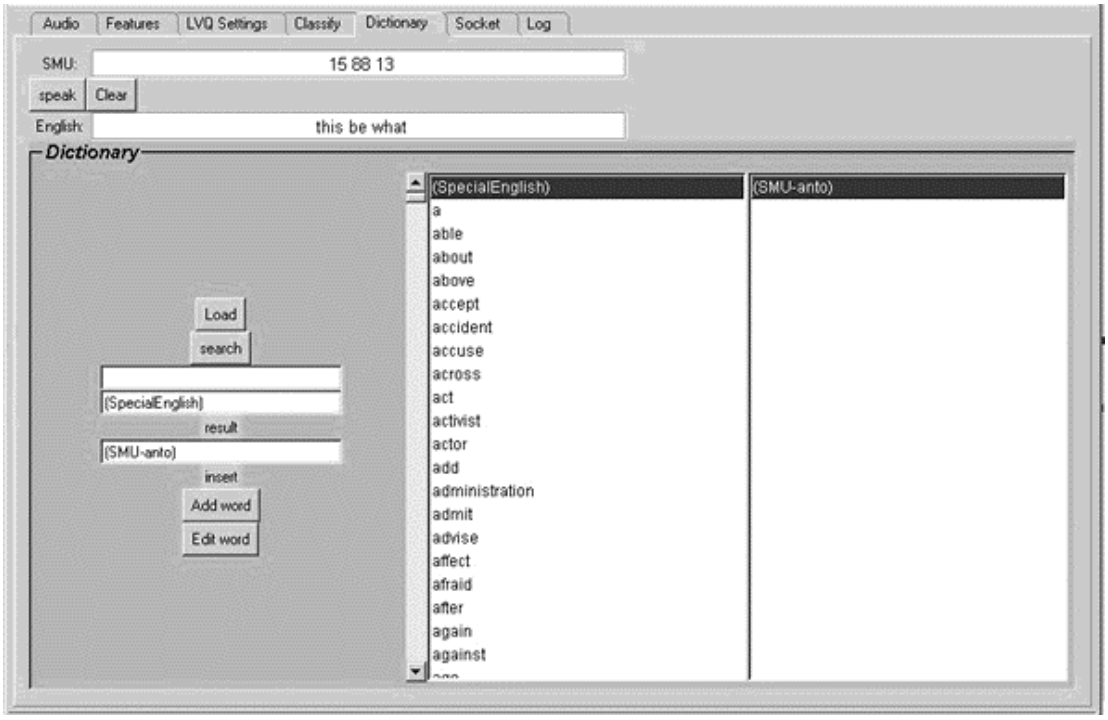


Figure 47. Screen shot of the Dictionary Tab with the Special English wordlist loaded as an example.

Since GAPR is assumed to be used on a relatively small vocabulary (less than 1,500 words), a simple lookup-table with two sides will be quick enough. One side for English (or some other human language), and one for the target artificial language. The Dictionary Tab offers the possibility to load an existing dictionary, to add or to edit words, and to select a series of words into a sentence, and have it played back in the target language. Playback in the artificial language is done by a concatenation of the collected SMU samples. The SMU's are loaded into RAM and joined together in any combination. This technique is used extensively in applications of a restricted domain, like when an automatic telephone operator tells you a phone number. In ext-to-speech applications (TTS), this approach has proven to be too simple. The many phonemes humans make, will sound differently depending on what phoneme precedes, and what phoneme follows. Rising and falling tones and many other factors are involved in order to make a natural-sounding TTS.

Having only three SMU's: (a, i, and o), there are rather limited possibilities of creating many words, unless we allow very long strings, but to follow the example through, a restricted domain language can be created, consisting of the four words up, down, right, and left, used to control movement in a 2D-plane.

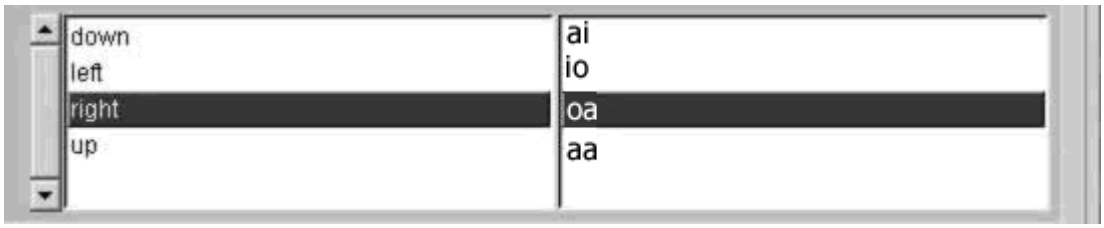


Figure 48. An example of GAPR holding a restricted domain language of only four words.

The building part is now done, and it's time to use the bridge.

7.2 Using GAPR

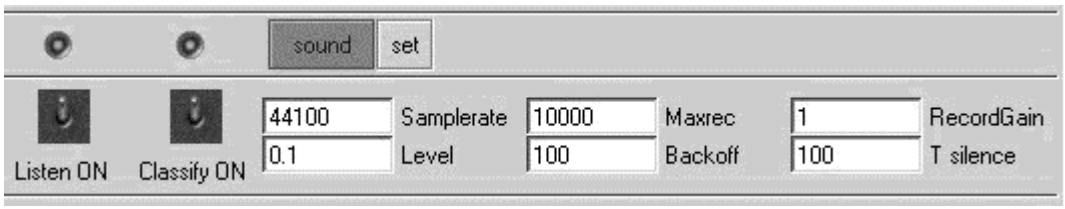


Figure 49. Screen shot of the SDet module.

When GAPR is used in real-time dialog mode, it is possible to use the Signal-detection module (SDet). Recording starts when the audio signal reaches a certain threshold (Level), and continues until there has been silence for a certain time (T-silence), or until a maximum recording time is reached (Maxrec). SDet will hold a buffer of the signal while listening, and insert a small section (Backoff) before the moment the signal reached the threshold, to ensure that the whole spoken utterance is captured. The captured signal is passed in a pipeline to the feature extraction, whose output is sent to the classifier. What happens with the output of the classifier, depends on the settings in the Socket Tab.

Socket Tab

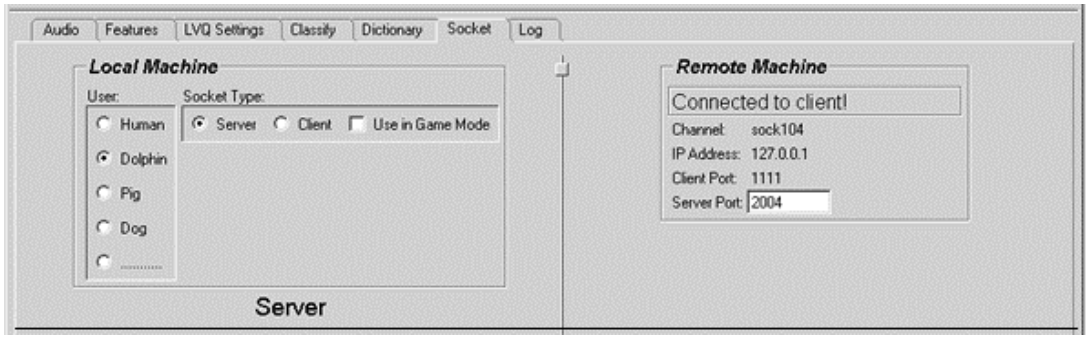


Figure 50. Screen shot of the Socket Tab.

The Socket Tab is the command central, where the flow of the data is decided. Several instances of GAPR can be linked up with each other, in a server–client relationship, using the Tcp/Ip protocol. The different instances are linked up either locally on the same machine, on separate machines in a local area network (LAN), or through the Internet. The socket Tab lets you decide which role each particular instance of GAPR should play.

Open architecture

Each instance of GAPR will only receive and send out SMU's, without concern for who is in the other end. It would for example be possible to have one instance of GAPR loaded with a project suitable for a pig, and another for dolphins. The pig instance would work exactly in the same manner as described above, but would have

recordings of pig sounds, and a codebook trained for those sounds instead. Whether other animals are capable of grasping the concepts necessary for this kind of communication, is not a concern with GAPR. The subject of this thesis is creating the bridge, and thereby enabling such further investigations.

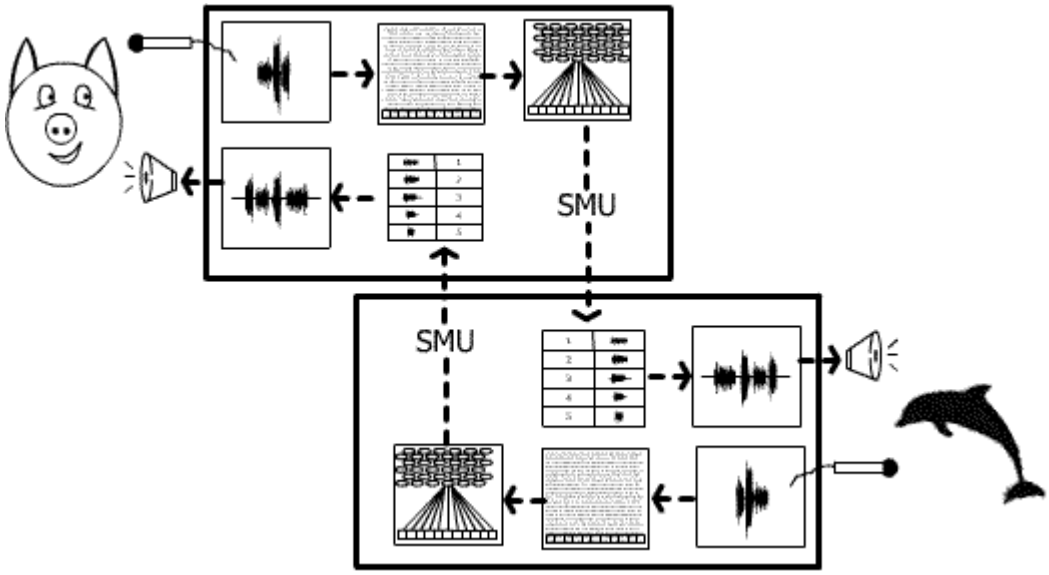


Figure 51. As far as GAPR is concerned, there is nothing stopping a pig-dolphin dialogue.

Dolphin-Computer: Game mode

GAPR can also be used in game mode. Interactive games can be created that will teach different aspects of the language, and perhaps an 'Edutainment jukebox' could let the dolphins discover and select games by their own. Macromedia's Flash and Director are programs that have been used extensively to create interactive games and educational content. With Flash 5 it is possible to send and receive data as XML. This enables GAPR to work in a server-client modus, as a 'plug-in' for Flash, so that events can be speech-controlled by the dolphins.

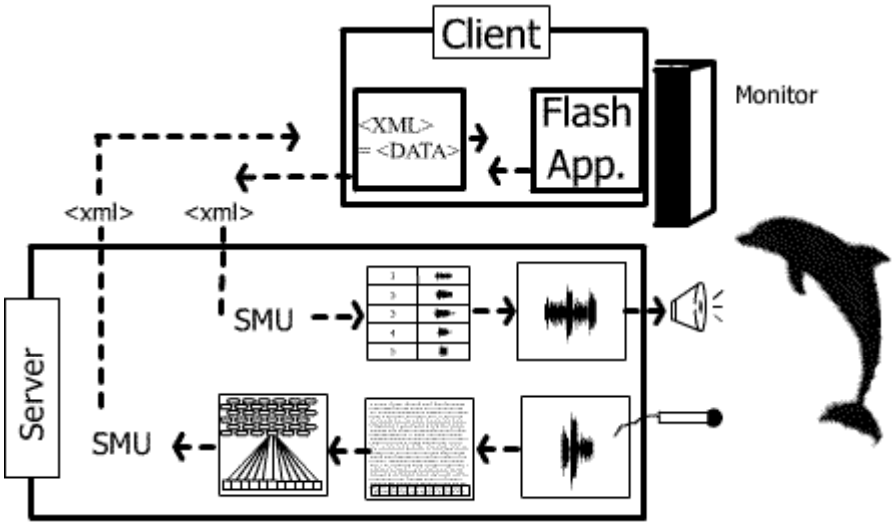


Figure 52. Schematic drawing of how data flows through GAPR in game mode.

If game mode is selected in the Socket Tab, the SMU's that are sent over the network are packed into an XML structure that enables Flash to receive them and parse the message into commands. Flash enables a developer to place symbols – which can be animations, graphics, bitmaps, movies etc – on a time line, and tell Flash to move forward, backward, or jump to a specific frame on that time line. Symbols with their own time lines can be embedded into other time lines enabling complex behaviors to be created. A programming environment called Action scripting is available making more dynamic behavior possible. Interactive movies for entertainment or educational purposes are becoming a big industry on the Internet, but are also very well suited for other aspects.

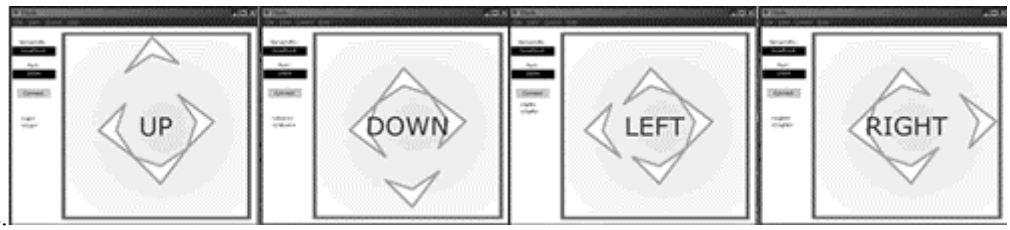


Figure 53. Four screenshots of an implemented example of how GAPR sends, and a Flash application receives XML data, allowing Flash to be used as GUI for educational content.

I have made a demonstration Flash application with four commands, corresponding to movement in four directions, that can be controlled from GAPR.

Internet chat room?

The architecture of GAPR makes it easy to implement other extensions like for example a dolphin Internet chat room. Since the sounds are transformed into symbols locally, only small amounts of data is actually passing through the network as text. If Hddp was taught in various locations, making a chat room would be trivial. As a play

with words, instead of the normal Http:// address, dolphins could get the Hddp:// extension instead.

The use of GAPR in a Classroom scenario

For teaching Hddp to circus dolphins, a good start could be to tap into the concepts they have already acquired from their life in show biz.

Language acquisition on word level can be described as:

1. Understanding of the concept
2. Memorization of the label
3. Linking of label and concept

We can have (1) without (2): knowing something, but not knowing what it's called. For example, being introduced to a novel African instrument, but not to it's name.

We can have (2) without (1): remembering the African word "bawaggawagga" but not knowing what it stands for, or a child saying "quantum physics" without knowing what it is.

Dolphins in show biz have been taught a number of concepts. Jumping through hoops, tossing a ball and so forth. It has been taught to them with hand signals and whistles and what not, so they have (1) but not (2). In order to link label and concept (3), the dolphins could be exposed to Hddp while the trainers did their normal routine and after some time the correspondence would be noticed. Initially there would be both the hand signals and SMU's while hearing and associating the SMU's with old learned tricks and eventually the hand signals could be replaced with SMU's entirely.

Flash games with animations of the same standard tricks could be made and work as confirmations on what we have agreed this combination of SMU's represent. The dolphins should enjoy being able to compose new combinations of sounds and controlling the animated dolphin to perform new acts. As any intelligent being is likely to do: they enjoy controlling their environment. As a child playing with blocks or Lego, an adult playing golf or Chess, captive dolphins play with balls and other toys in their 'spare time' as well. Actually both free and captive dolphins seem more concerned with playing and having fun than us humans. Remember Douglas Adams in The Hitchhiker's Guide to the Galaxy:

> "Man had always assumed that he was more intelligent than dolphins because he had achieved so much... the wheel, New York, wars, and so on, whilst all the dolphins had ever done was muck about in the water having a good time. But conversely the dolphins believed themselves to be more intelligent than man for precisely the same reasons."[37]

A game-boy language teaching toy could be an attractive item. To learn language, there has to be a reason to learn. To talk, to play, to exchange ideas etc. The reason we all learn language so quickly has to do with our desire to communicate, and the joy of communication. This desire is a factor in shaping the structure of the brain – we learn what we want to learn, and deem important – and should be provided for in order to achieve rapid progress.

Weaknesses with GAPR

As mentioned in the beginning of the chapter, the current version of GAPR must be seen as a prototype, with several limitations. I will mention what I see as the most important weaknesses in something of a 'wish list' for a future version.

Classifier for context-dependent data. I decided to focus on dolphin vocalizations in the pulse domain, and found that the LVQ classifier did a good job there. For whistles and other forms of sound patterns that get their properties from change in time, another kind of classifier is needed. A Dynamic Time Warping classifier, as an alternative to the LVQ-classifier for audio patterns with a contextual character, such as dolphin whistles, would have been desirable.

Pipelined recognition. Currently, recording and feature extraction runs in a pipeline, but then the features are written to file and the classifier is reading from file. Furthermore, the classified SMU's are not being sent across the network until the whole chunk is processed. This is creating a delay in the recognition process that is unnecessary, and probably not acceptable in a real-life situation. Since the recognition is frame-based, every frame is treated as a separate unit anyway, and there is no reason to wait until the user has stopped talking to begin recognition.

End-pointing algorithm. End-of-word and end-of sentence are aspects of communication that has not been considered in this thesis. GAPR is currently just waiting until there has been silence for a certain length of time before it sends the SMU's over the network. When building up a vocabulary, there will be a possibility for ambiguity in the interpretation of SMU's. If for example one word is '55' and another word is '5': How will the interpreter know that '55' is not two instances of '5'? A convention like - every word will end with an 'end of word' SMU would solve one aspect of it, but it would be an expensive convention using up one SMU only for an end pointing issue.

ASR not implemented. Automatic speech recognition on the human side would improve user friendliness, and is readily implemented.

Parsing of sentences. Parsing of sentences for translation into other styles of syntax would enable English users to write full sentences in a way they are accustomed to. The current implementation stops at word level, and expects the human users to be familiar with the syntax of the artificial language.

Relevant vocabulary. I have ignored the big job of creating an appropriate vocabulary. Special English, which is used as an example dictionary, contains a vocabulary of 1,400 words that makes it possible to keep a normal conversation. It was however, created by the United States Information Agency and contains words like: gun, tank, missile, air force, and so on. Words for an aquatic environment such as: octopus, whale, shrimp and so on, would perhaps be more appropriate.

Possibility for a bigger dictionary for human users. GAPR is currently also expecting human users to know which words are in the dictionary. It would be nice to create a link to, for example, Longman's semantic dictionary with synonyms, enabling human users to use synonyms that were not in the dictionary, compressing a sentence into its definition words on the dolphin side.

A speaker filter. For making different human speakers appear with a different flavor, it would be nice to have a set of filters that changed the outgoing signal a little bit. That way there would be an individual auditory difference between speakers, enabling the dolphins to hear who is speaking.

SMU recorder. A logging scheme that could play back conversations and view 're-runs' of dolphin game interactions by sending the log and a time code through the game.

Batch processing. To optimize the classifier, a trial and error search for optimal settings is required. For creation and evaluation of many codebooks with different settings, and based on different feature extraction styles, a batch process script would be useful.

Some real games. Edutainment games could be an important aspect of the language acquisition. A lot of work has gone into creating the XML speech interface to Flash, and the simple 'up-down-right-left' application is just showing the possibilities.

Improvement of LVQ classification. I worked out an idea for improvement of the LVQ classification. By collecting the mis-classified tokens from the MIDI slots in the Classify Tab and creating a new .dat file, a fine tuning of the classifier could be done. This would be a useful addition to GAPR.

SOM style data mining. The classification is surprisingly good with only one sample of each SMU, and perhaps the improvements would only have been minor, but it would be interesting to do the SOM style data mining of dolphin recordings described in section 6.5 to see how much it affected the recognition.

Robustness. There is lots of robustness missing. Error handling, internal conflicts in data. E.g. if a non-existent combination of SMU's is uttered, it is not taking any measure. Other similar robustness aspects are not dealt with.

– 8 –
Conclusion

In investigations of animal communication, focus is usually put on the cognitive aspects of the animals in question. Hasty conclusions have sometimes been drawn from failing to recognize the need for two other, equally important aspects of language. Apart from a good brain, a linguistic system and an interface is needed.

A linguistic system is a 'complex adaptive system' not *in* users, but *between* users. The power of such a system has been exemplified (section 1.4) with a stone-age baby moving to NYC. Instead of thinking the stone-age thoughts she was predestined to think, she will be preoccupied with computers, French fries and Hip-hop. Such a system is not something that evolves on a personal, physical level. Although she has made a quantum leap bypassing endless stages of cultural evolution, nothing in her genetic setup has changed.

Brains are carriers of a subset of this system, and the ability of brains to learn how to use the innovations of others comes as a consequence of this. Language is recognized as such an innovation. Brains have evolved the ability to adapt to a changing environment where plasticity is a key ingredient. Brains are seen as 'pattern-acquisition devices' able to lock on to systems of information. Artificial neural networks have been investigated as functionally similar to brains, and their ability as 'general pattern-matching machines' to recognize aspects of language (as well as a myriad of other things) has been documented. A well-developed mammalian brain, such as the bottlenosed dolphins, is hypothetically seen as able to do the same, if exposed to a linguistic system.

The need for an interface to handle the symbolic reference aspect of language is emphasized. The lack of voluntary motor control of speech organs in many species is disqualifying their participation in a language scheme based on these conditions. Bottlenosed dolphins are shown to have such voluntary motor control, but it is not compatible with the sounds used in human linguistic communication. New computing power and algorithms has made it possible to address this issue in a new way, something that could not have been done even few years ago.

These three aspects of language: a system, a brain, and an interface are seen as inter-dependent requirements for language acquisition. We have seen that the bottlenosed dolphin meets the requirements pretty well. Apart from vocal motor-control, big brains, and a cognitive understanding of symbolic reference and syntax, bottlenosed dolphins have excellent memory, live in complex social groups, vocalize a lot and are good at mimicking. This makes them very good candidates for an attempt to investigate inter-species communication of a much higher level than previously attempted.

A way to bridge the interface barrier between dolphin and human vocalization is presented: The General Audio Pattern Recognizer (GAPR). This application, together with general knowledge on how human linguistic systems work, constitute a language scheme - the Human-dolphin dialog protocol.

My hopes are that this thesis will work as a preliminary study, enabling a much larger empirical investigation of the linguistic capabilities of the bottlenose dolphin. I have here provided a tool that makes such investigations possible.

Our history books are full of stories where assumptions of limited abilities have created barriers, where bridges could have been made. With a new bridge at our disposal, investigations into the minds of the dolphin and new knowledge about the minds of men could emerge.

Reference list

[1] Aristoteles **Historia animalium** / ex rec. Immanuelis Bekkeri dokid. 50ka17248

[2] Terrence Deacon: **The symbolic species** W.W. Norton & Company; ISBN: 0-393-03838-6

[3] comparative mammalian brain collections, University of Wisconsin and Michigan State http://brainmuseum.org

[4] Herman Ruge Jervell: **Abstractions and Metaphors on the Internet**, SSGRR-2001, L'aquila, Italy

[5] Kim Sørenssen: **The architecture of Asian Modernity, Memetic Meditations on Modernity and Butterfly Effects**. Cand. Polit. Thesis, Dept. of Social Anthropology, NTNU, 2001.

[6] Christer Johansson: **A view from language**, Lund university press, ISBN 91-7966-421-0

[7] Richard Dawkins **The blind Watchmaker** s.158, Penguin, 2000, ISBN 0-14-029122-9

[8] Daniel C. Dennett: **The role of language in intelligence**, in What is Intelligence? The Darwin College Lectures, ed. Jean Khalfa, Cambridge Univ. Press, 1994

[9] Kolb, Bryan, **Brain plasticity and behavior**, Mahwah, N.J. : Erlbaum, 1995, ISBN 0-8058-1520-1

[10] Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). **Rethinking innateness**: A connectionist perspective on development. Neural Network Modeling and Connectionism. Cambridge, MA: MIT Press.

[11] Blakemore, C.B. & Cooper, G.F. (1970) **Development of the brain depends on the visual environment**. Nature, 228, 477–78.

[12] William H. Calvin: **Conversations with Neil's Brain,** ISBN 1-800-358-4566

[13] Geoffrey Sampson, **Educating Eve – The language instinct debate**, ISBN 0-304-70290-0

[14] Terrence J Sejnowski and Charles R Rosenberg **Parallel networks that learn**

to pronounce English text, Journal of Complex Systems, February 1987

[15] Honkela, Pulkki, and Kohonen, **Contextual Relations of Words in Grimm Tales, Analyzed by Self-Organizing Map**, Proceedings of International Conference on Artificial Neural Networks, ICANN-95, Paris 1995.

[16] Anders Nøklestad, (Hanne Gram Simonsen Rolf Theil Endresen Editors) **A Cognitive Approach to the Verb**, ISBN 3-11-017031-0

[17] Douglas Hofstadter: "**Fluid Concepts and Creative Analogies,** Penguin Books, 1998, ISBN: 0-14-025835-3

[18] Douglas Hofstadter in Wired Magazine, 3/11 –95

[19] scientific American Presents Winter 94
http://www.sciam.com/specialissues/1198intelligence/1198pepperberg.html

[20] Sue Savage-Rumbaugh, Roger Lewin: **Kanzi : The Ape at the Brink of the Human Mind,** John Wiley & Sons; ISBN: 047115959X

[21] Danny D. Steinberg: **An introduction to Psycholinguistics**, ISBN 0-582-05982-8

[22] L.M. Herman et al. **Sentence comprehension by bottlenosed dolphins**, Cognition 16 1984

[23] Edward Kako, Elements of syntax in the systems of three language-trained animals, Animal Learning & Behavior, 1999 27 (1), 1-14, with replies 15-17, 18-23, 24-25

[24] Steven Pinker **The language instinct,** Penguin Books, 1994, ISBN: 0-688-12141-1

[25] Derek Bickerton **Language and Species**

[26] Vladimir I. Markov and Vera M. Ostrovskaya **Organisation Of Communication System In Tursiops Truncatus Montagu**, A. N. Severtsov Institute of Evolutionary Morphology and Ecology of Animals, USSR Academy of Sciences, 33 Leninsky Prospect, Moscow 117071, USSR

[27] Ken Marten and Suchi Psarakos, **Using Self-View Television to Distinguish between Self-Examination and Social Behavior in the Bottlenose Dolphin (Tursiops truncatus)** Consciousness and Cognition (Volume 4, Number 2, June 1995) pp. 205-224, Orlando: Academic Press

[28] John C. Lilly, **Man and dolphin** Garden City, N.Y. : Doubleday, c1961

[29] Ronald W. Langacker: **An introduction to Cognitive Grammar** Cognitive Science 10: 1-40, 1986

[30] Microsoft Encarta

[31] James H. McClellan, Ronald W. Schafer, Mark A Yoder, **DSP First, A Multimedia Approach:** Prentice Hall, ISBN 0-13-243171-8

[32] Tengiz V Zorikov, Nikolai A Dubrovsky, Naira J Beckauri, **Signal processing by the bottlenose dolphin's sonar: Experiments and modelling**. Institute of Cybernetics, Georgian Academy of Sciences, Tbilisi. NN Andreev Acoustics institute, Moscow

[33] Caldwell, M. C., Caldwell, D. K. & Tyack, P. L. 1990**. Review of the signature-whistle hypothesis for the Atlantic bottlenose dolphin**, The Bottlenose Dolphin (Ed. by S. Leatherwood & R. R. Reeves), pp. 199–233. New York: Academic Press.

[34] Brenda McCowan, Diana Reiss **The fallacy of 'Signature whistles' in bottlenose dolphins**. Animal behaviour, 2001, 62, 1151-1162

[35] Tuevo Kohonen, **Self-Organizing maps**, Springer 1995

[36] Jari Kangas, Kari Torkkola, Mikko Kokkonen: Helsinki University of Technology, **Using SOMs as feature extractors for speech recognition**. ICASSP-92

[37] Douglas Adams: **The Hitchhiker's Guide to the Galaxy**, ISBN: 0345391802