Comparing Music and Language: Data structures, Algorithms, and Input-Output Conditions

Roland Hausser
Universität Erlangen-Nürnberg
Abteilung Computerlinguistik (CLUE)
rrh@linguistik.uni-erlangen.de

Abstract

The input-output conditions of music and language resemble each other in that both are time-linear. There are also some crucial differences in their functional structure, however. These are well-suited for bringing out important points regarding their respective cognitive mechanisms.

In this paper, music and language are jointly characterized as being time-linear, with melodies and language signs having only one abstract type and an indefinite number of corresponding tokens. The tokens, in turn, have different realizations in different media, e.g., speech, writing, and signing.

The differences between language and music are remarkable too, however: The distinction between natural languages like English or French is missing in music. Furthermore, the content transmitted by music does not contain any 'facts' from the external world (with the pardonable exception of 'program music'). If language seems to be favored in this comparison, it must be said in all fairness that a few minutes of good music may make up for hours of language in any medium.

1 Introduction

Up to now, scientific analyses of music and language have been mostly limited to structural objects, fixed on paper or magnetic tape. In music, such objects of analysis are exemplified by a single melody or the movement of a sonata, in language by a single sentence or an article in a newspaper. By concentrating on the melody or the sentence, one has attempted to abstract away from the aspect of communication.

The purpose of producing music or language in the first place, however, is interaction between cognitive agents. Furthermore, music and language in their original medium of sound are of a fleeting nature, whereas the agents using them are concrete and comparatively permanent pieces of wetware.

A cognitive analysis of language or music cannot fail to be inadequate,[1] if it does not include the production and interpretation procedures inside the cognitive agents. The question here is not what a piece of music or language *is*, but rather what it *does*, and how it does it by virtue of the way it is.

Our analysis of communication procedures will show that a piece of music or language may exist in many realization-dependent forms (tokens), but has only one realization-independent form (type). Take for example a certain melody. Performed over and over again on a violin, an oboe, a flute, or a piano, the melody is realized in a form which is dependent on the properties of the instrument as well as the player (color of sound, mono- versus polyphony, tempo, phrasing, etc.). The realization-independent form of this melody, in contrast, consists in the structure which all its different realizations have in common. This abstract form may be represented explicitly by using music paper, for example.

Agents must constantly switch between the realization-independent and -dependent forms: During production, a realization-independent structure is realized by the agent in a certain medium as a token of music or language. During interpretation, this realization-dependent token is transformed by the agent for processing – i.e., inferences, storage in memory, etc. – into a realization-independent form.

This paper argues that the production and interpretation circumstances of communication are a topic of considerable importance. Without understanding them thoroughly, one cannot even begin to analyze the structure of a melody or sentence in *functional* terms. The theory resulting from the new approach may and must be verified by constructing computational models (robots), the functioning of which

[1] This use of double negatives is quite mild compared to General Eisenhower's immortal: *I couldn't fail to disagree less.*

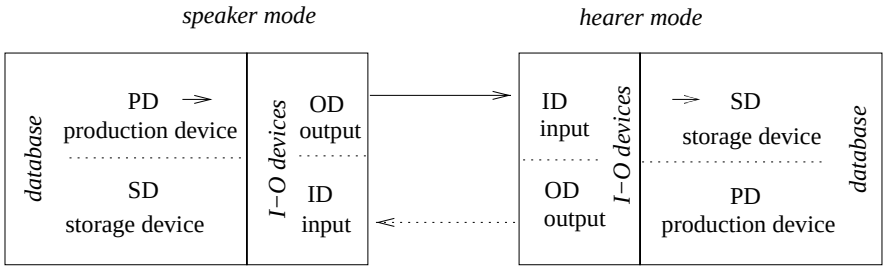
resembles human behavior more and more closely. Thereby, the analysis of the input-output conditions of communication must serve as the theoretical foundation for ensuring a functionally coherent flow of the software procedures and for integrating the recognition and action components of the hardware.

2 Input-output conditions of language and music

Where does a piece of music or language produced by a cognitive agent come from in the first place and how does it get realized (output)? How is it processed by the recipient (input) and where is it stored in memory? In other words, what is the general nature of the relation between production, output, input, and storage of language or music?

At the most general level, these questions may be answered by characterizing the input-output conditions of language and music jointly, as follows:[2]

2.1 BASIC INPUT-OUTPUT CONDITIONS

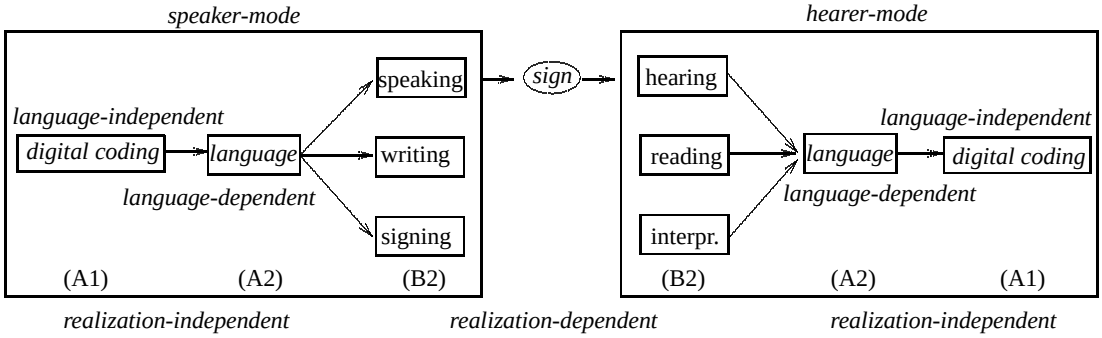


The cognitive agent consists of a database and an input-output (I-O) device. The database is equipped with a production device PD for autonomously generating melodies or sentences from the data structure, and a storage device SD for autonomously storing melodies or sentences in the data structure. Furthermore, the input-output device is divided into an output device OD associated with the production device (PD-OD) and an input device ID associated with the storage device (ID-SD).

3 Formats of natural language production and interpretation

Turning to a more differentiated analysis of the input-output conditions of natural language, we must distinguish between an abstract, language-independent representation of content (A1), a realization-independent representation of content in different languages (A2), and a realization-dependent representation in different media (B2):

3.1 LANGUAGE- AND MEDIUM-DEPENDENCE IN COMMUNICATION



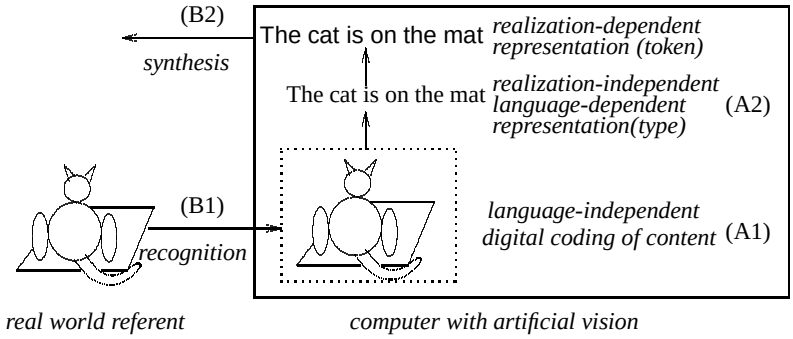
[2]For simplicity, the terms speaker-mode and hearer-mode are used here for the production and interpretation, respectively, of language and music alike, regardless of the medium, i.e., speaking, writing, or signing in the case of language. Later, when we turn to a more detailed analysis of music, the terms music production and music consumption will be used.

Communication requires the speaker to associate an abstract content (A1) with a suitable sign in a particular language (A2), which is realized in a particular medium (B2). The hearer in turn must map the language token realized in a particular medium (B2) into the corresponding realization-independent language type (A2), which is associated with a suitable abstract content (A1).

The content coded by a speaker into language comes from different sources. One is the non-verbal interaction with the external environment, represented in the format (B1) of unanalyzed perception or action. A second source is the retrieval of observations from memory. A third source is provided by actively rearranging structures stored in memory, as in inferencing, planning, or the creation of fiction.

As an example, consider an agent observing a cat on a mat (B1). This observation is represented cognitively as an abstract digital representation (A1) which may be coded neurologically, as in a human, or electronically, as in a robot.

3.2 A CONTENT ORIGINATING FROM OBSERVATION

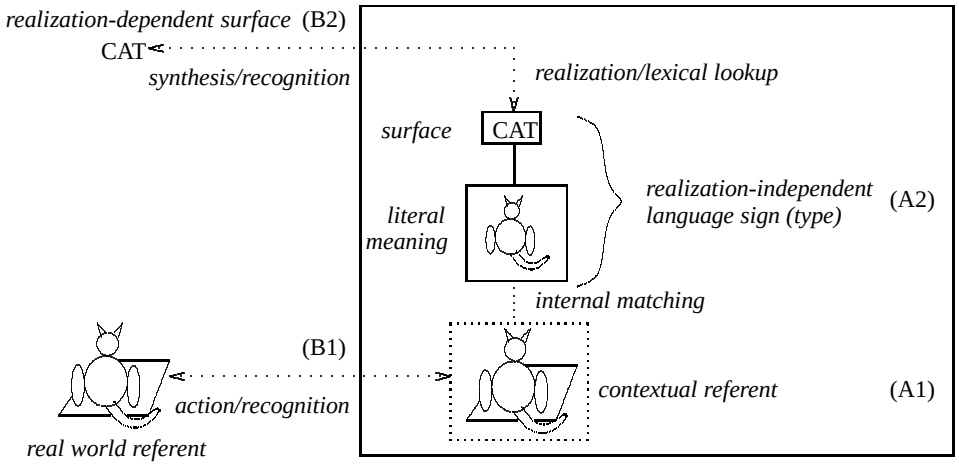


During production, the digital representation of content (A1) is coded into the sign types of a particular language (A2) and realized in the form of surface tokens (B2).

The use of an artificial cognitive agent in this example has the advantage of providing a tangible realization of the abstract, language-independent form of the content. It consists in digitally coded concept tokens which result from analyzing the visual input with corresponding concept types.

The correlation of a language-dependent coding (A2) and a corresponding contextual content (A1) is based on the principle of internal matching. Consider the following example, which is a structural refinement of 3.2 showing the internal matching of the single content word cat:

3.3 THE SIGN STRUCTURE ENABLING INTERNAL MATCHING



In format A1, the context is represented as a structure of concept tokens which represent a cat on the mat. We assume that this subcontext was established via recognition of a real world referent (B1).

The concept types used for recognizing the visual input (B1) and for deriving concept tokens (A1) as contextual referents are *reused* as the literal meanings of natural language (A2). Communication is based on matching the concept types of the language level with concept tokens of the context level.

In order for communication to work it is inevitable that language signs have a two-level structure. One consists of the literal meaning, defined as a concept type. The other consists of the surface type, which depending on the language may be, for example, cat, chat, or Katze. The two levels are lexically attached to each other by means of conventions which every speaker-hearer has to learn.

Different natural languages make it possible to code the same content in different forms. For example, English The cat is on the mat, French Le chat est sur le tapis, and German Die Katze ist auf der Matte are different, language-dependent surface types which code the same abstract content.

4 Algorithms and data structures of language communication

Natural language production and interpretation each constitute a two step process. In the speaker-mode, the two processes consist in (i) the transfer from format A1 (context) into format A2 (language), called production, and (ii) the transfer from format A2 (surface types) into format B2 (surface tokens), called synthesis. In the hearer-mode, they consist in (i) the transfer from format B2 into format A2, called recognition, and (ii) the transfer from format A2 into format A1, called interpretation.

In B2/A2 transfer, each of the language-dependent sign types will be synthesized or recognized as realization-dependent tokens in different media: spoken, written, or signed (B2). Thereby, spoken language is realized acoustically, whereas written and signed language are realized visually; the latter differ in that written signs are static, while signed signs are dynamic in nature.

Furthermore, the realization-independent representation of a language (A2) is usually associated with a particular realization-dependent format (B2): In speech, each language is associated with a language-dependent pronunciation system. In writing, some languages are associated with a particular writing system, e.g., Greek, Hebrew, Arabic, hieroglyphs, pictograms. In signing, some languages are associated with a particular system of sign articulation, e.g., ASL (American Sign Language), LSF (Langue des Signes Française), DGS (Deutsche Gebärdensprache) – requiring the 'interlingua' sign language Gestuno, which is equivalent to 'Esperanto' (for the media of speaking and writing).

From the interpretation structure of a single word (cf. 3.2 and 3.3), let us turn next to the interpretation structure of a *sequence* of words (or rather word forms). Let us also include the useful distinction between the speaker and hearer modes, which has been omitted in 3.3 for the sake of abstraction.

The speaker mode originates at the level of context. As already mentioned, the required content may have several sources: these are observation, memory, and operations on memory. In the SLIM theory of language (Hausser 1999/2001, 2001), they are all uniformly treated as a navigation through a special kind of database called a word bank, which contains concatenated propositions.[3]

[3]The data structure of a word bank uses the alphabetic order of concept *types* (using basic English words like give, love, man, Mary, Peter, walk, woman, etc. as names or 'handles' by which to manipulate (compose) their meanings.) Each type is followed by a sequence of corresponding tokens, arranged in the order of their arrival in the database:

type token token token token ...

Each token is part of a proposition. For example, Peter loves Mary would be represented by the tokens Peter, love, and Mary, each stored in their appropriate *token line*.

The tokens are called proplets. The proplets belonging to – or making up – a certain proposition are held together by having the same proposition number. Furthermore, the proplet Peter is a feature-value structure (frame) which says: *my functor is* love. The proposition number and the functor type name love are sufficient to travel from Peter through the data structure to love.

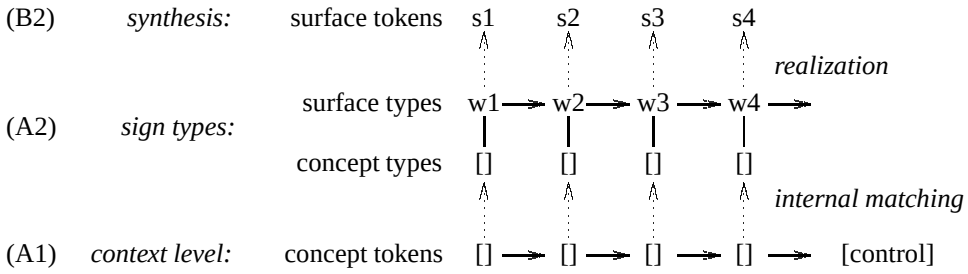
Conceptually speaking, this is accomplished by going from the token (proplet) Peter to the corresponding type at the beginning of the token line, from there alphabetically to the token line of love and from there linearly through the token line to the proplet with the same proposition number (whereby proposition numbers are strictly linearly ordered by time of arrival).

When retrieved, the proplet love specifies its arguments, here Peter and Mary, the first confirming the first valency filler's address of origin and the second providing the address of the second filler. The surface Mary and the proposition number are enough to travel from the frame love to the frame in question, i.e., Mary, thus traversing – and activating – the whole proposition.

In addition to the intrapropositional navigation described above, there is also extrapropositional navigation from one proposition to the next. This is based on two types of bidirectional address-specifications: *identity* between objects (nouns, arguments) and conjunction between *relations* (verbs, functors).

The continuous intra- and extrapropositional navigation of a focus through the database is driven by the time-linear motor algorithm of LA-grammar, whereby its control is located at the level of context (A1).

4.1 SCHEMA OF NATURAL LANGUAGE PRODUCTION (SPEAKER MODE)

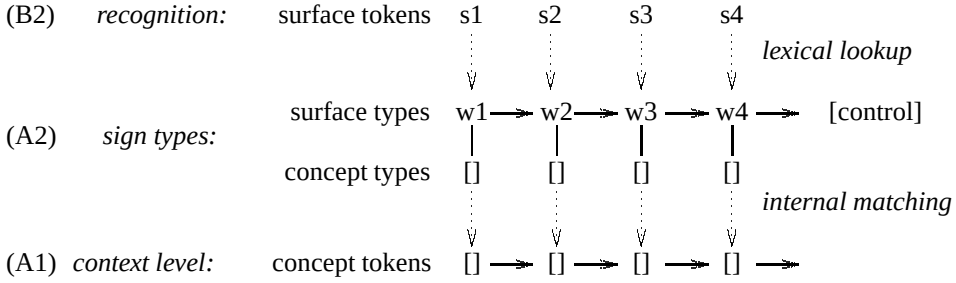


The navigation activates sequences of propositions (thought), drives inferences, answers queries, and – last but not least – provides *conceptualization* for natural language production. Given the two-level structure of natural language sign types, production simply matches[4] the concept tokens traversed at the level of context (A1) with suitable concept types (A2) serving as the literal meanings of signs. This results in a sequence of associated surface types of a particular natural language. The surface types are realized as surface tokens in a particular medium (B2).

The hearer mode constitutes the inverse procedure, starting with the incoming unanalyzed surface tokens (format B2), which are recognized by matching them with the surface of suitable lexical items. This procedure depends on the medium, i.e., speech, handwriting, print, or signing.

The resulting sequence of sign types (A2) controls the associated navigation through the context (A1). This is based again on the internal matching of the signs' concept types with suitable contextual referents, defined as concept tokens.

4.2 SCHEMA OF NATURAL LANGUAGE INTERPRETATION (HEARER MODE)



This overall structure provides the speaker with two opportunities for creativity and the hearer with one. The speaker may be creative in the choice (i) of the navigation (i.e., the thought and thus the conceptualization) and (ii) of how to code the navigation into the natural language at hand (verbalization). The hearer may be creative in how to match the incoming language signs with corresponding context structures (interpretation) – which is the inverse of verbalization.

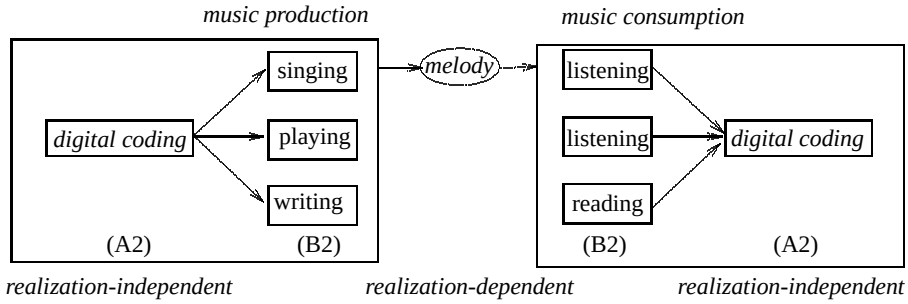
5 Components of music production and consumption

Next let us apply the same analysis to music. Music resembles natural language in that it is time-linear as well. However, while natural language production is based on *choosing* words or word forms with lexically predefined surfaces, combinatorics, and meanings in order to *model* a contextual structure (content), music production is based on *creating* surfaces by defining their pitch and length within a framework of measures and keys.

This is possible because in music there is no distinction between content and its coding in different languages. Instead the abstract type of a particular piece of music *is* the content. The goal of music is to define a sequence of surfaces, called notes, which is pleasing or interesting to the hearer.

[4]The historian philospher Charles Taylor 1975, p.7., calls it "*projection of meanings onto things*."

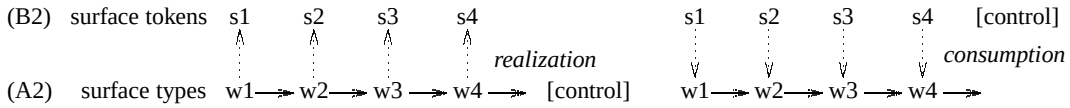
5.1 MEDIUM-DEPENDENCE IN MUSIC



Like a sentence in a particular language, a melody exists in the forms of many realization-dependent tokens (B2) and one realization-independent type (A1). However, while language production and interpretation distinguish four formats, namely B1, A1, A2, B2, music production and consumption distinguish only two, namely A2 and B2. A2 is the abstract, realization-independent coding of, for example, a melody, while B2 is the concrete realization in a particular medium, e.g., singing.

During music production, an abstract type is realized in a particular medium as a token. During music consumption, a realization-dependent token is mapped into an abstract type. As in language, the origin of a melody may have several sources.[5] Consider the following analysis of music production and consumption, by analogy to the analysis of language production and interpretation in 4.1 and 4.2 above:

5.2 SCHEMATA OF MUSIC PRODUCTION AND CONSUMPTION



In music, there is no internal matching between meanings and contextual referents as in natural language. There is no other content than the surface to be transmitted. In this sense, music is not a language, and a fortiori not a universal language. Instead, the surface sequence with its melody and its rhythm affects the cognition of the hearer directly.

Music may seem to differ from language also in that music is realized mainly in the medium of sound, whereas the realization of language is distributed more evenly between the media of sound, writing, and signing. There is, however, the special case where professional musicians write sheet music on paper which is read by other musicians without any realization in sound.

Furthermore, while the realization of an abstract melody requires only one output device, e.g., a bird using its voice, music is frequently realized by combining an output device, for example a human player, with a special input-output device, i.e., an instrument. The player provides an output by moving the fingers, controlling the air pressure, etc.[6] The instrument responds by creating vibrations as output.[7]

By analogy to the use of an instrument in music there is the use of a microphone or telephone for spoken language, and of a typewriter or standard computer for written language. An analogy to using several instruments unisono is a chorus, as in Greek drama. An analogy to polyphonic music is rare in spoken language because the blending of different words in simultaneous speech is an obstacle to understanding.

[5] When a singer or player improvises or composes, the abstract music originates in his or her head, created by the musical part of cognition. This source is analogous to a speaker inventing a content by inferencing, planning, or inventing fiction.

Another source consists in imitating a piece of music from direct experience, as when listening to a bird song, or retrieving a piece of music from memory. This is analogous to a speaker talking about present or past experience.

Finally, there is the source of singing or playing from sheet music. This is a case of inter-media transfer, analogous to reading aloud from print or typing in a handwritten letter. The origin of sheet music is due to the creation of abstract music by composers.

[6] This output serves as the input to the instrument, whereby different instruments provide different input devices: the strings of a violin activated by a bow, the mouthpiece of an oboe or flute, or the keys of a piano. Because of this variety, a player cannot master all instruments at the same time. Instead, a player usually specializes in one instrument or instrument type.

[7] It is also possible to combine several instruments, as in a trio, quartet, or orchestra. When a certain melody is played simultaneously by several instruments, the realization properties of the instruments combine. These are the different colors of their sound, the mono- versus polyphonic execution, and their conditions of input and output.

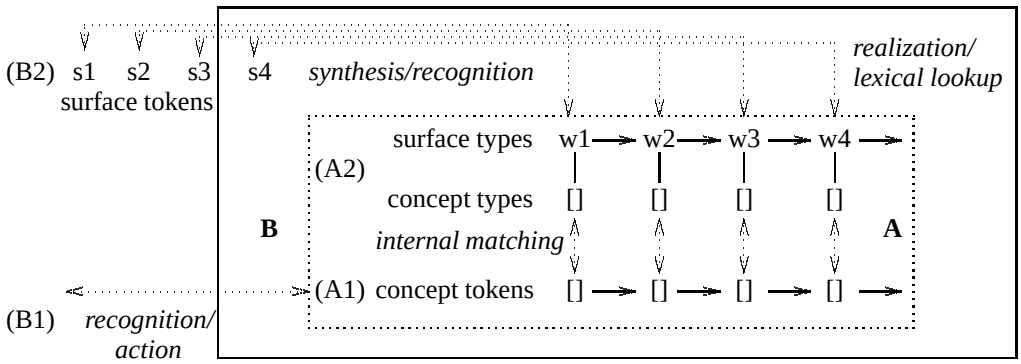
6 Consequences for computational linguistics

Of the four formats B1, A1, A2, and B2 of language communication, there are two natural groups:

- B1 and B2 for recognition and action at the levels of context (B1) and language (B2), and
- A1 and A2 for producing language from content (A1/A2) and content from language (A2/A1).

These groups may be characterized schematically as follows, indicated as **A** and **B**, respectively:

6.1 TWO DOMAINS OF COMPUTATIONAL LINGUISTICS



Domain **A** (inner box) comprises the data structure, the motor driving the navigation (A1), the internal matching mechanism of language with its rules of pragmatic interpretation, and (A2) the sign types with their two levels of concept types and surface types. This rule-based structure constitutes the cognitive core engine, whereby a suitable control structure is part of the mechanism of dabase navigation.

Domain **B** (outer box) comprises the procedures of recognition and action at both the levels of context (B1) and language (B2). This is where the most money is spent today, whereby speech recognition seems to have an edge over contextual recognition and action. The methods used, however, are mostly statistical, and thus limited in principle in that they are not functional.[8]

Why is there no speech recognition which is continuous, speaker-independent, domain-independent, with a realistic vocabulary of at least 100 000 word forms, and robust in the sense that one can use it to dictate email in a noisy environment, like while driving in an open car? The short answer is: Because current systems analyze natural language as if it were music.

In other words, current systems of speech recognition concentrate on the intermedia transfer B2, but completely neglect the processing of communication in the cognitive core engine **A**.[9] Without the free, meaningful production and understanding of natural language by artificial cognitive agents (machines), however, the search space of statistical methods will continue to be so large that it constitutes the biggest single obstacle to ever obtaining successful automatic speech recognition.

References

Hausser, R. (1999/2001) *Foundations of Computational Linguistics: Human-Computer Communication in Natural Language*, Springer-Verlag, Berlin-New York.

Hausser, R. (2001b) "Database Semantics for Natural Language," *Artificial Intelligence*, Vol. 130.1:27–74, Elsevier, Dordrecht.

Saussure, F. de (1972) *Cours de linguistique générale*, Édition critique préparée par Tullio de Mauro, Éditions Payot, Paris, France.

[8] A few years ago during a lunch at the University of Edinburgh with some highly qualified and argumentative people in speech recognition, I presented the following quip:

If the Martians came to earth and modelled cars statistically they would never run.

Despite the obvious analogy to speech recognition and language understanding, they had no response.

[9] The only solid linguistic knowledge used is first, lexica for helping decide what hypothesis in speech recognition could be a possible word form, and second, syntactic parsers for deciding what could be a grammatical sequence of word forms. While this is a step in the right direction, it does not amount to a modeling of understanding.