Gefördert vom
bmb+f
Bundesministerium für
Bildung, Wissenschaft,
Forschung und Technologie

Verbmobil
Verbundvorhaben

# Spoken Language Processing in the Hybrid Connectionist Architecture SCREEN

Stefan Wermter, Volker Weber

Universität Hamburg

Vm
**Report 138**
Juli 1996

Juli 1996

Stefan Wermter, Volker Weber

Arbeitsbereich Natürlichsprachliche Systeme
Fachbereich Informatik
Universität Hamburg
Vogt-Kölln-Str. 30
22527 Hamburg

Tel.: (040) 5494 - 2530/2531
Fax: (040) 5494 - 2515
e-mail: wermter@informatik.uni-hamburg.de
weber@informatik.uni-hamburg.de

**Gehört zum Antragsabschnitt:** 15.2 Hybride konnektionistische Architekturen

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 101 A/O gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei dem Autor.

**Abstract**

In this paper[1] we describe a robust, learning approach to spoken language understanding. Since interactively spoken and computationally analyzed language often contains many errors, robust connectionist networks are used for providing a flat screening analysis. A screening analysis is a shallow flat analysis based on category sequences at various syntactic, semantic and dialog levels. Rather than using tree or graph representations a screening analysis uses category sequences in order to support robustness and learning. This flat screening analysis is examined in the context of the system SCREEN (Symbolic Connectionist Robust EnterprisE for Natural language). Starting with the word hypotheses generated by a speech recognizer, we give an overview of the architecture, and illustrate the flat robust processing at the levels of syntax, semantics, and dialog acts. While early connectionist models were often limited to a single network and a small task, the hybrid connectionist SCREEN system is an important step towards exploring connectionist techniques in larger hybrid symbolic/connectionist environments and for real-world problems. Based on our experience with SCREEN, hybrid connectionist techniques show a lot of potential for supporting robustness in interactive spoken language processing.

[1]This article has been origanly published in the IEEE Computer Journal. The original version appeared in the July 1996 issue.

# 1 Introduction

Recently there has been a renewed interest in interactive processing of spoken natural language, e.g., (Reilly and Sharkey 1992; Menzel 1994; Allen et al. 1995; Jurafsky et al. 1994; Geutner et al. 1996; Weber and Wermter 1996b; Wermter et al. 1996). If we want to process spoken language from human-to-human interactions or human-with-computer interactions we have to take into account the "interactive noise" in spontaneously spoken language. By interactive noise we refer to those phenomena which do occur in spontaneously spoken language but not in written language. Examples of this noise are interjections, pauses, repetitions, repairs, false starts, uncommon syntactic or semantic constructions, etc. Furthermore, there is noise based on incorrect word hypotheses from a speech recognizer which may lead to ungrammatical constructions.

Therefore the analysis of spoken language has to be fault-tolerant and robust. Traditional and recent symbolic methods have been used successfully for various aspects of text analysis. However, for dealing with faulty spoken language analysis more robust methods have to be used. Other approaches for spoken language analysis have used statistical methods in combination with symbolic methods (Allen et al. 1995; Charniak 1993; Jurafsky et al. 1994)[2]. In this article, we want to examine hybrid connectionist methods for spoken language since connectionist networks are known to be robust for unexpected but similar input. There has been some earlier work on language processing using connectionist networks (Waltz and Pollack 1985; McClelland and Elman 1986; McClelland 1991; Reilly and Sharkey 1992; Miikkulainen 1993). While this work demonstrated successfully how language processing can be achieved in connectionist networks, these early models had to be small in coverage. In contrast we describe the SCREEN[3] system which is based on a real speech recognizer, which learns a robust syntactic and semantic analysis, provides an interpretation up to the dialog act representation, and deals with uncommon syntactic and semantic language anomalies. We consider the hybrid connectionist SCREEN system as a step towards exploring connectionist techniques in larger hybrid symbolic/connectionist environments and for real-world spoken language problems.

There are three fundamental principles which are addressed in SCREEN. First, we want to examine hybrid connectionist *learning* techniques in a real-world speech/language system. Second, we want to explore to what extent hybrid connectionist techniques can provide the necessary *robustness*. Third, we want to examine a *screening approach* to spoken language analysis; that is, rather than an

[2] We will describe and compare some representative related approaches in the discussion section, after we have described SCREEN.

[3] **s**ymbolic **c**onnectionist **r**obust **e**nterprise for **n**atural language

in-depth understanding we aim at a flat understanding, but we want the understanding to be robust and learned. In general, our long-term perspective has been to examine the architectural consequences in hybrid connectionist systems based on these principles. In this paper, we will primarily focus on the second point, the robustness, since this is a key issue for interactive speech language systems.

Below we show some transcription examples which were taken from a German dialog corpus of business meeting arrangements. For easier readability these examples were translated in a *literal* and *improved* native version. The first sentence is a relatively simple request to make a suggestion. As a different example, the second contains a false start and an interjection ("afterix eh"). In the third sentence, utterance boundaries have to be identified. While these transcribed sentences just show the noise introduced by humans, later we will also see the noise introduced by the imperfect analysis of a speech recognizer which may insert, delete, and replace words.

- Gut machen Sie einen Vorschlag (Literal: Fine make you a suggestion; Improved: Fine please suggest a date)
- Vorschlagen am sechzehnten April nachix eh (Literal: Suggest on sixteenth April afterix eh; Improved: Suggest the sixteenth afterix eh)
- Das wird etwas knapp bei mir sagen wir lieber vierzehn Uhr fünfundvierzig (Literal: That is bit short for me say we rather 14 o'clock 45; Improved: That is a little short for me let's rather say 2:45 pm)

Such noise phenomena occur very frequently in spoken language and they have to be dealt with. However, systems for spoken language analysis cannot use directly the same technology which may have been proven useful for text processing. Rather, a much higher level of robustness and fault tolerance has to be integrated into interactive spoken language systems. We primarily consider two different ways to reach robustness.

First, robustness can be accomplished by mutually complementary modules which cooperate to make a common decision. If multiple modules are involved in one decision it is possible that an error by a single module may not influence the common overall decision. Since this robustness is based on the modularity we call this a "modularity-based robustness". A good example for modularity-based robustness would be the detection of a word repair, where the equality of lexical, syntactic, and semantic knowledge in different modules all contribute to the decision whether one word should be replaced by its subsequent word.

A second form of achieving robustness is by using *graded* representations for supporting similarity. While modularity-based robustness involves several modules, "similarity-based robustness" concentrates on the processing within a single module. Similarity-based robustness reflects the fact that graded analogous representations support a graceful degradation for potentially faulty input. Especially in real-world systems like speech/language systems there are gradual variations, for instance in the speech input signal. However, also at higher levels ambiguities and contradictory preferences are best dealt with using a gradual representation which can support similarity-based robustness.

In earlier work we have shown the potential of connectionist networks for learning a flat *text* analysis (Wermter 1995), referred to as a scanning understanding. Here we will describe work on our new system SCREEN for learning a flat *spoken language* analysis, referred to as a screening understanding. Both, a scanning understanding of text and a screening understanding of spoken language, are based on sequences of flat category representations. However, a screening understanding of spoken language has to deal much more with various forms of noise compared to a scanning understanding of written language.

Input to the SCREEN system is a stream of word hypotheses generated by a speech recognizer, output is a stream of syntax, semantics and dialog hypotheses for different utterance hypotheses. The overall processing is incremental so that analysis results can be used immediately after they have been produced and modules do not have to wait until the processing of a complete utterance has been finished. Furthermore, incremental processing allows SCREEN to let different modules interact early in the utterance, which is used for instance for repairing words or phrases. Incremental processing and robust processing are also relevant for computational models of spoken language, since they both exist in human interactive language processing.

## 2 Noisy Word Hypotheses from a Speech Recognizer

The domain we work with is the arrangement of business meetings between business partners. Currently we deal with 184 dialog turns and 2355 words. However, as we will point out in the discussion, our approach can be transferred easily to other spoken language domains. The amount of domain-dependent knowledge is mainly restricted to the semantic knowledge. Furthermore, this domain-dependent semantic knowledge can be learned automatically from examples.

Our analysis is based on a HMM[4] speech recognizer which incrementally provides a list of word hypotheses based on the spoken input signal. The interface between the speech recognizer and the language processing is based on these word

[4] Hidden Markov Model; see (Charniak 1993) for an introduction to HMM models.

hypotheses. Word hypotheses are illustrated using a word graph as shown in figure 1. In general and in practice, these word graphs can become much bigger with many more connections. The size of the word graphs depends on the quality of word recognition the HMM speech recognizer is able to provide.

Each word hypothesis is represented by the word itself (for readability with its literal translation) and a plausibility value for the acoustic confidence of the speech recognizer for this word hypothesis. Based on their start and end times these word hypotheses can be connected to (partial) sentence hypotheses.

Usually a single word hypothesis can have several subsequent word hypotheses. For instance, in figure 1, after the initial "pause", different word hypotheses have been computed by the speech recognizer: "ähm" (eh), "oh" (oh), "bis" (until), and "das" (that). As another example, after the word hypothesis "mir" (me), marked by "*" in figure 1, the word hypotheses "noch" (still), "auch" (also), "sollen" (should), and "sagen" (say) are all possible successors with different speech confidence. Through these different noisy word hypotheses, many undesired noisy sentence hypotheses can be built.

# 3 Screening analysis based on flat representation

In SCREEN we use rather flat representations which come from five basic knowledge sources: a basic syntactic and a basic semantic word description, an abstract syntactic and an abstract semantic phrase description, and a dialog act description. The flat analysis is based on category sequences at various syntactic, semantic and dialog act levels. A flat representation using the categories shown in table 1 supports robustness and learning within a modular hybrid-connectionist architecture. For instance, the syntactic, semantic, and dialog act analysis of an example utterance "das ist knapp bei mir" (that is short for me) produces the output shown in figure 2.

A basic category is assigned to each word of the utterance, an abstract category to each phrase, and a dialog act to each turn.

# 4 An Overview of the SCREEN Architecture

Now we will give a brief overview of our SCREEN system (see figure 3). In this section we summarize the most important architecture features and we will illustrate the processing in SCREEN with a more detailed example in section 5. SCREEN consists of six main parts each of which contains several modules. As

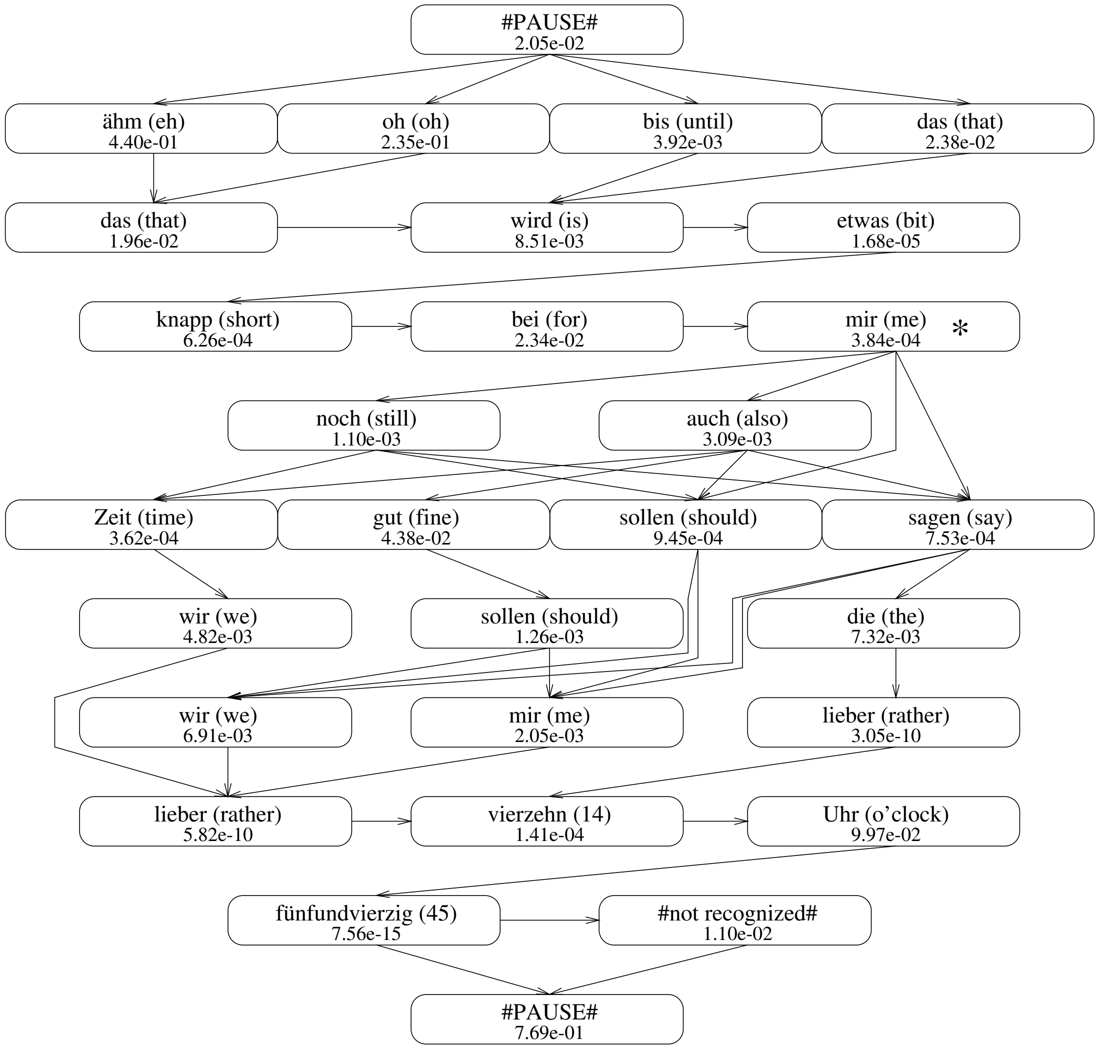


Figure 1: Word graph for the dialog turn: Das wird etwas knapp bei mir sagen wir lieber vierzehn Uhr fünfundvierzig (Literal: That is bit short for me say we rather 14 o'clock 45; Improved: That is a little short for me let's rather say 2:45 pm). Word hypotheses are shown with the word, its literal translation, and its acoustic plausibility value.

| basic syntax | basic semantics | abstract syntax | abstract semantics | dialog act |
|---|---|---|---|---|
| **n**oun | **sel**ect | **v**erb **g**roup | **act**ion | **acc**ept |
| ad**j**ective | **sugg**est | **n**oun **g**roup | **aux**-action | **query** |
| **v**erb | **meet** | **a**dverbial **g**roup | **agent** | **rej**ect |
| **a**dverb | **utter** | **p**repositional **g**roup | **obj**ect | **re**quest-**c**omment |
| p**r**eposition | **is** | **c**onjunction **g**roup | **recip**ient | **re**quest-**s**uggest |
| **c**onjunction | **have** | **mo**dus **g**roup | **instr**ument | **state** |
| prono**u**n | **move** | **s**pecial **g**roup | **manner** | **sugg**est |
| **d**eterminer | **aux**iliary | **i**nterjection **g**roup | **time**-**at** | **misc**ellaneous |
| nu**m**eral | **quest**ion | | **time**-**from** | |
| **i**nterjection | **phys**ical | | **time**-**to** | |
| **p**articiple | **anim**ate | | **loc**ation-**at** | |
| **o**ther | **abs**tract | | **loc**ation-**from** | |
| pause (/) | **here** | | **loc**ation-**to** | |
| | sou**rc**e | | **conf**irmation | |
| | **dest**ination | | **neg**ation | |
| | **loc**ation | | **quest**ion | |
| | **time** | | **misc**ellaneous | |
| | negative evaluation (**no**) | | | |
| | positive evaluation (**yes**) | | | |
| | unspecific (**nil**) | | | |

Table 1: Categories for the flat representation. Abbreviations are shown in bold face.

| Utterance<br>Translated | Das<br>That | ist<br>is | knapp<br>short | bei<br>for | mir<br>me |
|---|---|---|---|---|---|
| basic syntax | D | V | J | R | U |
| abstract syntax | NG | VG | AG | PG | |
| basic semantic | ABS | IS | NO | HERE | ANIM |
| abstract semantic | OBJ | ACT | NEG | RECIP | |
| Dialog act | REJ | | | | |

Figure 2: Flat analysis of a simple utterance (for abbreviations see table flat-categories).

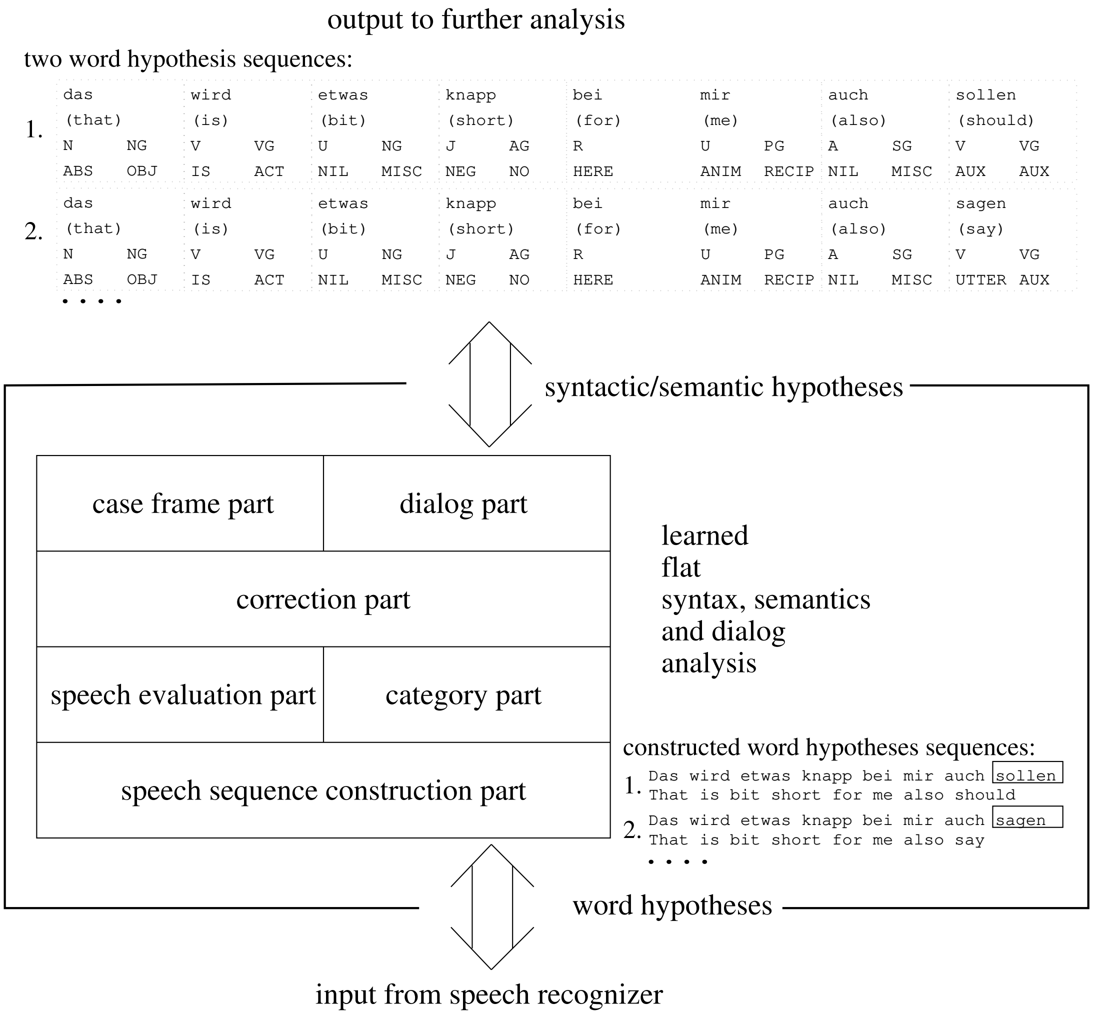


Figure 3: Overview of the SCREEN system

shown in figure 3 input for SCREEN are word hypotheses, output are flat representations of the screening analysis. The six parts are responsible for constructing sequences from words, evaluating the best sequences, providing syntactic and semantic categories, correcting mistakes, supplying a case frame and assigning dialog acts.

In figure 4 we will now focus on a more detailed description of the SCREEN system. The data flow is shown by arrows between modules, in some cases we have used numbers to avoid too complex arrow drawings. The *speech sequence construction part* (CON-SEQU-HYPS) at the bottom of figure 4 incrementally receives single word hypotheses from a speech recognizer and constructs possible partial sentence hypotheses. At the bottom of the figures 4 we see one of many possible sentence hypotheses "that is bit short for me also say".

The *speech evaluation part* contains modules for computing sequence plausibilities of individual partial sentence hypotheses. These plausibilities allow to choose better partial sentence hypotheses based on acoustic, syntactic, and semantic knowledge. The evaluation is done by comparing basic syntactic and semantic category predictions (BAS-SYN-PRE and BAS-SEM-PRE) with actually occurring categories.

Category knowledge is learned and generalized in the *category part*. This part contains several modules for a flat syntactic and semantic analysis of a current sentence hypothesis. The syntactic and semantic analysis is performed at two syntactic (basic and abstract syntactic categorization: BAS-SYN-DIS and ABS-SYN-CAT) and two semantic levels (basic and abstract semantic categorization: BAS-SEM-DIS and ABS-SEM-CAT). Furthermore, phrase starts are detected for identifying phrase boundaries (PHRASE-START). The categories assigned by these modules are shown in table 1.

The *correction part* contains modules for often occurring mistakes which have to be dealt with explicitly in spontaneous language. For instance, there are modules for detecting interjections and pauses (PAUSE-ERROR), word repairs (WORD-ERROR), and phrase repairs (PHRASE-ERROR). The principles of repair recovery and the correction part are described in more detail in (Weber and Wermter 1996a). The modules of the correction part provide an explicit machinery for robustness while individual connectionist modules also have an implicit similarity-based robustness.

The *case frame part* contains a segmentation parser for segmenting complete dialog turns into utterance segments and for filling the contents of a case frame with the utterance constituents. This parser uses the semantic and syntactic knowledge provided by the underlying flat syntactic and semantic analysis from the category part.

The *dialog part* is responsible for recognizing dialog acts of utterances and

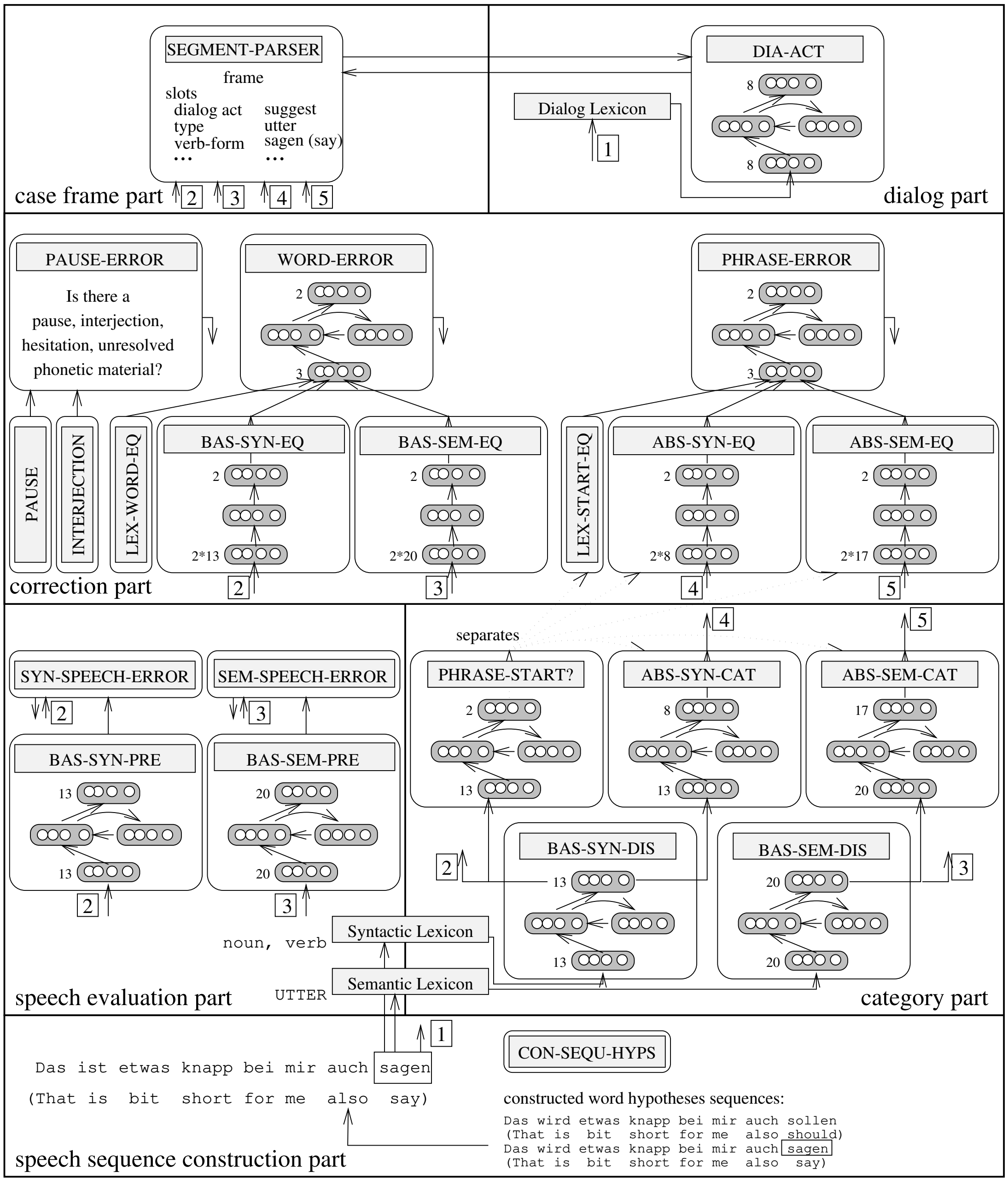


Figure 4: Architecture of the SCREEN system (see text for abbreviations).

interacts with the case frame part. The segmentation parser in the case frame part provides utterance boundaries for the dialog act part. In return, the dialog act part provides dialog acts (DIA-ACT) which are also stored in the case frame representation.

As shown in figure 4, we have chosen primarily connectionist feedforward networks and simple recurrent networks (Elman 1990) where the sequential context can be learned. Gradient descent is used to train these networks (Rumelhart et al. 1986). If a module does not contain a connectionist network it uses simple symbolic rules, for instance for a lexical comparison of two words.

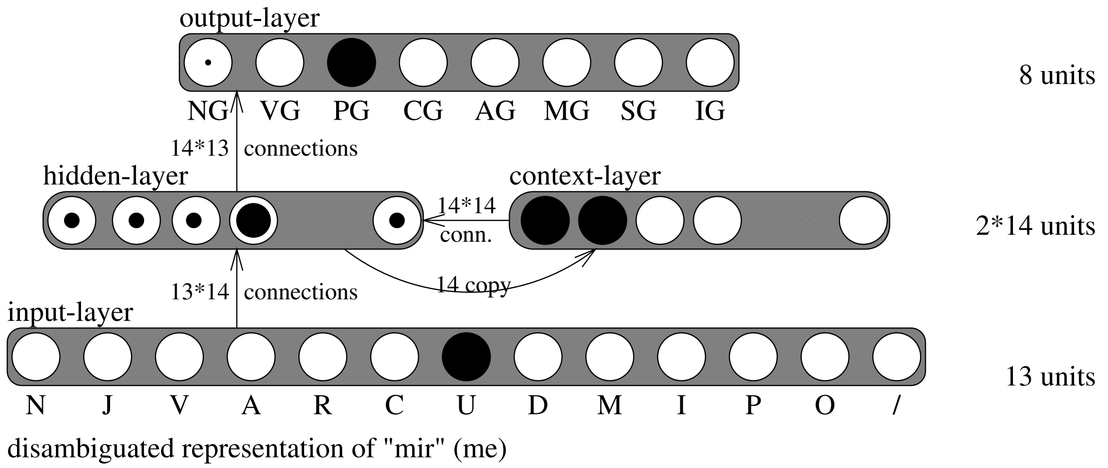


Figure 5: Abstract syntactic categorization in ABS-SYN-CAT: "(bei) mir", ((for) me) is found to be a prepositional group in this context.

As a detailed example of one network, we will describe the ABS-SYN-CAT module for abstract syntactic categorization. There are 13 input units (which is the output of the disambiguation of BAS-SYN-DIS) for the basic syntactic categories, and 8 output units for the abstract syntactic categories (for the abbreviations see table 1). The network in figure 5 shows the current network state when the representation of "mir" (me) has been presented to the network.

Although we cannot go into all the details of each module we have decided to illustrate the complete architecture, since this gives an impression of the similarity-based and modularity-based robustness. Similarity-based robustness is realized in particular by those modules which contain a connectionist network, for instance for abstract syntactic categorization (ABS-SYN-CAT). Modularity-based robustness is realized where several modules contribute to another module, for instance for the lexical equality of two words (LEX-WORD-EQ), syntactic equality of basic syntactic categories (BAS-SYN-EQ), and semantic equality of basic semantic categories (BAS-SEM-EQ) which all contribute to the detection of word repairs and word repetitions (WORD-ERROR).

# 5 Incremental Processing of Word Hypotheses

In order to describe the principle of incremental flat robust processing in SCREEN we show two snapshots of the running system. The example shown is: "Das wird etwas knapp bei mir sagen wir lieber vierzehn Uhr fünfundvierzig" (Literal: That is bit short for me say we rather 14 o'clock 45; Improved: That is a little short for me let's rather say 2:45 pm). The word hypotheses which were shown in figure 1 are processed incrementally, from left to right, and the results at various levels of syntax, semantics, and dialog acts are shown in figures 6 and 7.

Figure 6 shows the state of SCREEN at the middle of our sentence hypotheses. Horizontally, the first four current sentence hypotheses are shown at that time step (1.77 seconds). They have been built in an incremental manner after each incoming word hypothesis. At that time the top
four sentence hypotheses essentially represent the desired sentence and they only differ in their confidence values.

Each sentence hypothesis consists of word hypotheses and each word hypothesis is shown with its most favored categories. The illustrated boxes are graphical representations of the output activations of the respective connectionist networks with the highest values. In addition to these categories and the confidence representation, phrase boundaries are indicated by filled rectangles.

For relating the representation of a word hypothesis in this snapshot to our architecture we have shown the abbreviations from the architecture in figure 4 also at the bottom of figure 6 for the word "mir" (me). The module BAS-SYN-DIS assigned the basic syntactic category "pronoun" (U), the module ABS-SYN-CAT the abstract syntactic category "prepositional group" (PG), the module BAS-SEM-DIS the basic semantic category "animate" (ANIM), the module ABS-SEM-CAT the abstract semantic category "recipient" (RECIP), and the module DIA-ACT computed the dialog act category "reject" (REJ). The confidence category is an integration of the acoustic, syntactic, and semantic confidence values for that partial sentence hypothesis.

The confidence of each sentence hypothesis is another example how acoustics, syntax and semantics interact to identify the best sentence hypotheses. If acoustics, syntax and semantics provide relatively high single values, this will result in a high integrated value, but slight variations of the value of a single knowledge

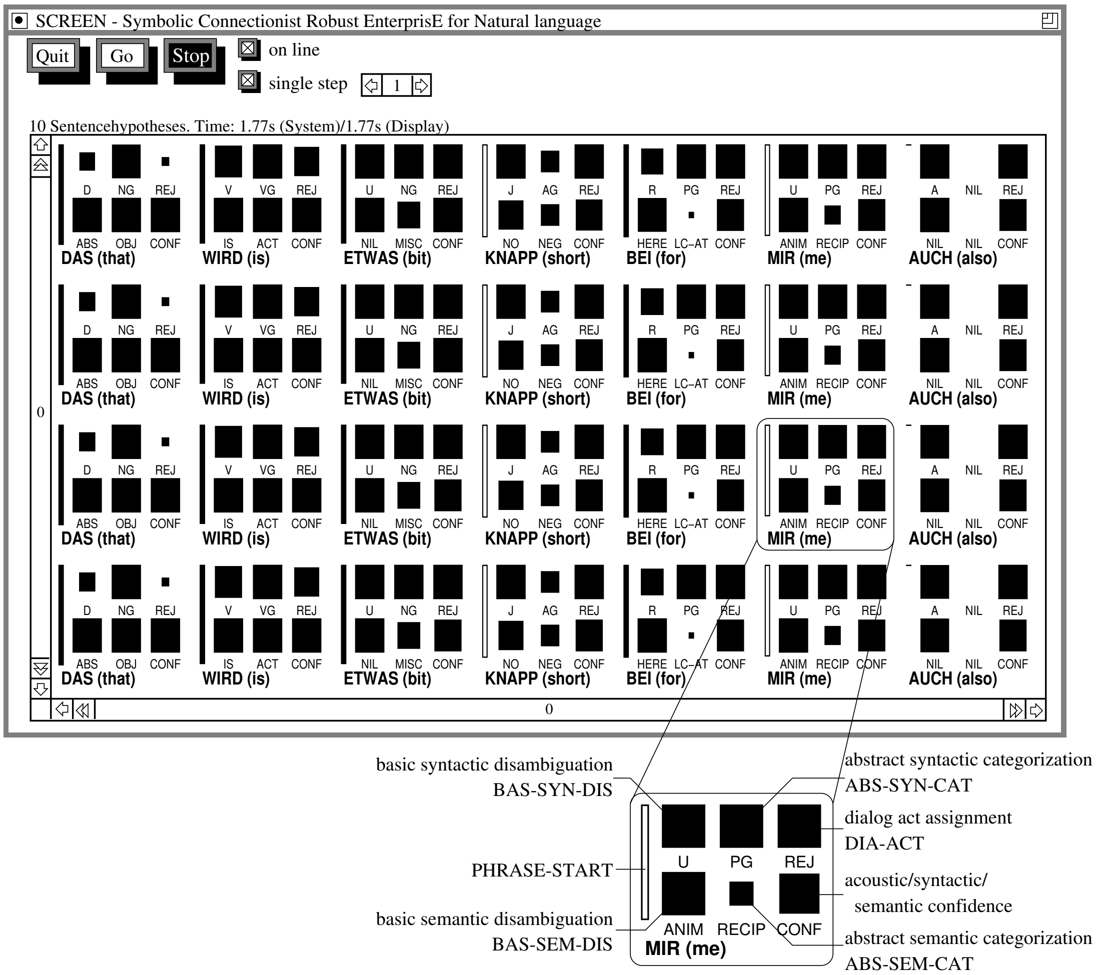


Figure 6: Snapshot 1 of four best found sentence hypotheses. Incremental processing proceeds from left to right. More details about abbreviations of syntactic and semantic categories can be found in tables 1 and 2. At the bottom of this figure we map the activation boxes for a single word hypothesis to the respective categorization tasks.

source do not necessarily result in a different order of the sentence hypotheses. So we have another example here how integration can provide a larger robustness in interactive speech language systems.

Figure 7 shows the state of SCREEN after all word hypotheses have been seen and processed. Since our environment allows the sentence hypotheses to scroll from left to right (and also from top best to bottom worst) we see only the right part of the top four sentence hypotheses due to the length of this utterance. Here we find the third sentence hypothesis as the desired sentence hypothesis with the exception of an additional "also". For clarity we focus primarily on this sentence hypothesis example.

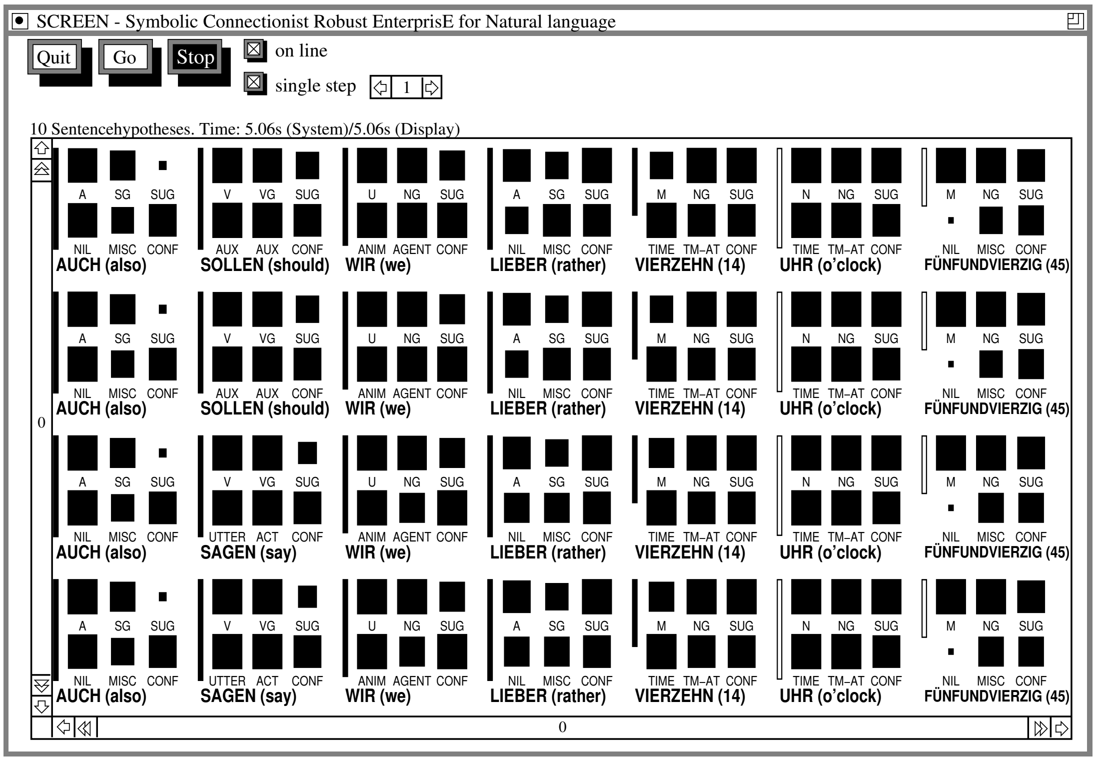


Figure 7: Snapshot 2 of four best found sentence hypotheses. More details about abbreviations of syntactic and semantic categories can be found in tables 1 and 2.

If we analyze SCREEN's output representations we notice an additional "also" which did not occur in the spoken sentence but which was introduced by the imperfect speech recognizer. In addition, the first two sentence hypotheses also differ from the desired "say" in the single word "should". In these examples, we have an insertion error ("also") and a replacement error ("should") from the speech recognizer. In this case, even additional syntactic and semantic knowledge cannot correct such errors since the sentences actually could have been spoken that way. On the other hand, as we can see from the similar sentence hypotheses, the best found sentence hypotheses are very close to the desired interpretation; the third sentence hypothesis is even the desired spoken sentence hypothesis with the exception of an additional "also".

Finally, for additional illustration, table 2 shows the syntactic and semantic categories which were assigned to our third sentence hypothesis. This table also shows the abbreviations in uppercase characters. As we can see in table 2 most words have been recognized (there is an incorrect additional word "also" which was not in the spoken sentence) and most categories have been assigned correctly. In addition to these syntactic and semantic categories there are also dialog categories which we will focus on in the next section. Together, syntactic, semantic, and dialog act representations provide a flat interpretation of sentence hypotheses.

| word-hypothesis | basic syntax | abstract syntax | basic semantics | abstract semantics |
|---|---|---|---|---|
| That (das) | Determiner | NounGroup | ABStract | OBJect |
| is (ist) | Verb | VerbGroup | IS | ACTion |
| bit (etwas) | pronoUn | NounGroup | NIL | MISCellaneous |
| short (knapp) | adJective | AdverbialGroup | NO | NEGation |
| for (bei) | pRepositon | PrepositionalGroup | HERE | LoCation-AT |
| me (mir) | pronoUn | PrepositionalGroup | ANIMate | RECIPient |
| also (auch) | Adverb | SpecialGroup | NIL | MISCellaneous |
| say (sagen) | Verb | VerbGroup | UTTER | ACTion |
| we (wir) | pronoUn | NounGroup | ANIMate | AGENT |
| rather (lieber) | Adverb | SpecialGroup | NIL | MISCellaneous |
| 14 (vierzehn) | nuMeral | NounGroup | TIME | TiMe-AT |
| o'clock (Uhr) | Noun | NounGroup | TIME | TiMe-AT |
| 45 (fünfundvierzig) | nuMeral | NounGroup | NIL | MISCellaneous |

Table 2: Another view on the syntactic and semantic categories assigned for the third sentence hypothesis of figures 6 and 7

# 6 Frames with Semantic and Dialog Knowledge

The current output of SCREEN is knowledge up to the semantics and dialog act level. Therefore, in this section we will give an outline of the case frame part and the dialog part. The case frame part contains a segmentation parser which receives the knowledge about the basic and abstract categories from the syntactic and semantic levels. Based on this knowledge the segmentation parser performs the tasks of segmenting a complete dialog turn in individual utterances and filling the constituents into appropriate frame slots. The segmentation is performed based on symbolic rules for possible sentence boundaries. For instance, conjunctions like "because" would start new utterance segments. Furthermore, new occurring full verbs like "say" could be an indicator for a new utterance segment. For each newly detected utterance segment a new frame is triggered. For our example turn with two utterance segments we show the incremental processing in two frames

| Slots | 1.–2 Phrase | 3. Phrase | 4. Phrase |
|---|---|---|---|
| dialog act | *rejection?* | *rejection?* | *rejection* |
| type | utter | utter | utter |
| verb-form | is | is | is |
| question | | | |
| auxiliary | | | |
| agent | | | |
| object | That | That | That |
| recipient | | | for me |
| time-at | | | |
| time-from | | | |
| time-to | | | |
| location-at | | | |
| location-from | | | |
| location-to | | | |
| confirm | | | |
| negation | | bit short | bit short |
| miscellaneous | | | |
| input | That is | That is bit short | That is bit short for me |

Table 3: Incremental slot filling in frame 1; for clarity of reading we have used the English literal translations: Das wird etwas knapp bei mir (Literal: That is bit short for me)

| Slots | 1. Phrase | 2. Phrase | 3. Phrase |
|---|---|---|---|
| dialog act | *suggestion?* | *suggestion?* | *suggestion* |
| type | is | is | is |
| verb-form | say | say | say |
| question | | | |
| auxiliary | | | |
| agent | | we | we |
| object | | | |
| recipient | | | |
| time-at | | | rather 14 o'clock 45 |
| time-from | | | |
| time-to | | | |
| location-at | | | |
| location-from | | | |
| location-to | | | |
| confirm | | | |
| negation | | | |
| miscellaneous | | | |
| input | say | say we | say we rather 14 o'clock 45 |

Table 4: Incremental slot filling in frame 2; Sagen wir lieber vierzehn Uhr fünfundvierzig (Literal: say we rather 14 o'clock 45

in the two tables 3 and 4.

On the left hand of table 3 there are the frame slots; the individual columns show the incremental slot filling after a phrase has been identified. In addition to the semantic slots, the frame type, and the frame input, we also show the dialog act of the current utterance segment. For instance, the dialog act of the utterance segment "That is bit short for me" is a rejection, the dialog act of the utterance segment "say we rather 14 o'clock 45" is a suggestion.

While the segmentation parser has been realized as a symbolic program due to the few segmentation rules, the dialog act assignment has been learned in a simple recurrent network. Since the utterances are processed incrementally the results of the analysis can be examined in the frames directly after each word or phrase. For more details of dialog act processing see (Wermter and Löchel 1996).

# 7 Discussion and Conclusion

We have described SCREEN, a new speech language system based on new hybrid connectionist technology. We have used connectionist modules wherever possible for learning and generalizing different functionalities in a robust and fault-tolerant manner. Since real spoken language contains much more noise compared to written language we have primarily used modularity-based and similarity-based principles from connectionist architectures to deal with such noisy environments. However, we also use symbolic techniques wherever necessary. For instance the communication and control of the individual modules has been realized symbolically since the communication paths were known and did not have to be learned.

The mode of processing in SCREEN is different from usual text processing techniques. First, we have developed techniques for dealing with interjections and pauses, word repairs and phrase repairs, as well as other ungrammatical constructions which do not appear in written language. Furthermore, the flat analysis in SCREEN supports the robust processing which is necessary for dealing with the output of a speech recognizer directly. Furthermore, we have designed SCREEN to be incremental left to right. Different syntactic and semantic modules can work in parallel rather than syntax before semantics as it is often done in traditional text processing. Our incremental processing of real spoken language also does not use techniques like multiple sequential parsings of the same sentence or processing techniques like backtracking which has been used often in symbolic text parsers.

We have evaluated the performance of our flat screening analysis in SCREEN in different ways. Since we are most interested in how we could reach robust language processing on new unknown real-world utterances, here we describe the results of using 184 dialog turns with 2355 words from the domain of business meeting arrangements. Among other tasks we trained simple recurrent networks on the tasks of basic syntactic, abstract syntactic, basic semantic, abstract semantic and dialog act categorization using 64 dialog turns as training material. After training we tested the performance of these networks on 120 previously unknown new and real-world test dialog turns. The performance on these new dialog turns was 89% correctness for basic syntactic analysis, 84% for abstract syntactic analysis, 86% for basic semantic analysis, 83% for abstract semantic analysis, and 79% for dialog act analysis.

We consider this performance as quite promising already, given that all utterances were real-word utterances without any preprocessing. In addition to these percentages it is interesting to note that this performance could be reached with a relatively small training set of 64 dialog turns. Therefore these connectionist techniques should be particularly useful for scaling up and bootstrapping in medium domains where the number of sentences is restricted. In fact, we have already examined the performance of SCREEN using the same architecture for the domain of interactions at a railway counter. In this railway domain people ask questions about train connections. The performance was comparable to the business meeting domain with 93% correctness for basic syntactic analysis, 85% for abstract syntactic analysis, 84% for basic semantic analysis and 77% for abstract semantic analysis (for more details see (Wermter and Weber 1996)).

The most related work to our approach are PARSEC (Jain 1992), BeRP (Jurafsky et al. 1994), and TRAINS (Allen et al. 1995). The hybrid connectionist system PARSEC is part of the larger speech translation effort JANUS (Waibel et al. 1992). PARSEC's input are sentences, its output are case role representations annotated with linguistic features. PARSEC contains several connectionist modules which trigger symbolic transformation rules. SCREEN's communication is done by a message-passing system which does not distinguish between connectionist and symbolic modules outside of the modules. In contrast to SCREEN, PARSEC uses prosodic knowledge while SCREEN also contains modules for learning dialog act assignment.

In the Berkeley Restaurant Project (BeRP) multiple different representations for speech/language analysis are employed (Jurafsky et al. 1994). The task is to provide guidance for choosing restaurants. Processing in BeRP is done as follows: A feature extractor receives acoustic data and extracts features which are used in the connectionist phonetic probability estimation. The output of this connectionist feedforward network is used in a Viterbi decoder whose output is transformed into database queries by a stochastic chart parser. A dialog manager is used to control the dialog with the user. Both, SCREEN and BeRP perform a flat analysis. However, while BeRP uses a probabilistic chart parser to compute all possible fragments, SCREEN uses multiple connectionist networks for flat parsing.

The goal of TRAINS is to build a general framework for spoken natural language processing and to plan train schedules (Allen et al. 1995). A person which interacts with the system is assumed to know more about the goals of scheduling while the system is supposed to know the details of the domain. In contrast to using the speech hypotheses as in SCREEN, transcripts of utterances are parsed by a syntactic and semantic parser. However, further linguistic reasoning is done by modules for scoping and reference resolution. Based on conversation acts of a dialog manager template-driven responses are generated. Processing in TRAINS focuses more on an in-depth planning level, based on a chart parser with a generalized phrase structure grammar, while SCREEN uses primarily a flat connectionist language analysis. Both systems are able to cope with performance phenomena like repairs and false starts (Allen et al. 1995).

In summary and conclusion, SCREEN is a robust system for speech/language analysis. Given that speech recognizers will continue to be suboptimal and will produce errors, robustness is a very important property for speech language systems. SCREEN has the ability to process input constructions even if they contain different errors. Based on the erroneous speech signal analysis it may not be possible to come up with the desired interpretation in all cases, but due to its robustness SCREEN does not break for any ungrammatical input and it attempts to provide the best possible interpretation at the syntactic, semantic, and dialog level. Due to this robustness hybrid connectionist architectures hold a lot of potential for further research in interactive speech language analysis.

## Acknowledgments

This research was funded by the German Federal Ministry for Research and Technology (BMBF) under Grant #01IV101A0 and by the German Research Association (DFG) under contract DFG Ha 1026/6-3. We would like to thank S. Haack, M. Löchel, M. Meurer, U. Sauerland, and M. Schrattenholzer for their work on SCREEN.

## References

Allen J. F., Schubert L. K., Ferguson G., Heeman P., Hwang C. H., Kato T., Light M., Martin N. G., Miller B. W., Poesio M., Traum D. R. (1995) The TRAINS Project: A case study in building a conversational planning agent. *Journal of Experimental and Theoretical AI 7*, pp. 7–48.

Charniak E. (1993) *Statistical Language Learning.* MIT Press, Cambridge, MA.

Elman J. L. (1990) Finding Structure in Time. *Cognitive Science 14*, pp. 179–211. *2*.

Geutner P., Suhm B., Buø F. D., Kemp T., Mayfield L., McNair A. E., Rogina I., Schultz T., Sloboda T., Ward W., Woszczyna M., Waibel A. (1996) Integrating Different Learning Approaches into a Multilingual Spoken Language Translation System. In: Wermter S., Riloff E., Scheler G. (Eds.) *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pp. 117–131. Springer Verlag, Heidelberg.

Jain A. N. (1992) Generalization performance in PARSEC - A structured connectionist parsing architecture. In: Moody J. E., Hanson S. J., Lippmann R. R. (Eds.) *Advances in Neural Information Processing Systems 4*, pp. 209–216. Morgan Kaufmann, San Mateo, CA.

Jurafsky D., Wooters C., Tajchman G., Segal J., Stolcke A., Morgan N. July/August 1994 Integrating experimental models of syntax, phonology, and accent/dialect in a speech recognizer. An investigation of tightly coupled time synchronous speech. *Proceedings of the 12$^{th}$ National Conference on Artificial Intelligence (AAAI-94) Workshop on the Integration of Natural Language and Speech Processing*, Seattle, Washington. AAAI Press/The MIT Press, Menlo Park.

McClelland J. L. (1991) *Towards a Theory of Information Processing in Graded, Random, Interactive Networks.* PDP-CNS-91-1, Parallel Distributed Processing and Cognitive Neuroscience Devision, Carnegie Mellon University, Pittsburgh, PA.

McClelland J. L., Elman J. L. (1986) Interactive Processes in Speech Perception: The TRACE Model. In: Rumelhart D. E., McClelland J. L., The PDP research group (Eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, 2., Psychological and Biological Models*, pp. 58–121. MIT Press, Bradford Books.

Menzel W. (1994) Parsing of Spoken Language under Time Constraints. Cohn A. G. (Ed.) *Proceedings of the 11$^{th}$ European Conference on Artificial Intelligence, ECAI-94*, pp. 561–564. (Amsterdam).

Miikkulainen R. (1993) *Subsymbolic Natural Language Processing. An integrated model of scripts, lexicon and memory.* MIT Press, Bradfor Book, Cambridge, MA.

Reilly R., Sharkey N. E. (Eds.) (1992) *Connectionist Approaches to Natural Language Processing.* Hove: Lawrence Erlbaum.

Rumelhart D. E., Hinton G. E., Williams R. J. (1986) Learning internal representations by error propagation. In: Rumelhart D. E., McClelland J. L., The PDP research group (Eds.) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, 1., Foundations*, pp. 318–362. MIT Press, Bradford Books, Cambridge, MA.

Waibel A., Jain A. N., McNair A., Tebelskis J., Osterholtz L., Saito H., Schmidbauer O., Sloboda T., Woszczyna M. (1992) JANUS: Speech-to-speech translation using connectionist and non-connectionist techniques. In: Moody J. E., Hanson S. J., Lippmann R. R. (Eds.) *Advances in Neural Information Processing Systems 4*, pp. 183–190. Morgan Kaufmann, San Mateo, CA.

Waltz D. L., Pollack J. B. (1985) Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science 9*, pp. 51–74. *1.*

Weber V., Wermter S. (1996) Artificial Neural Networks for Repairing Language. *Proceedings of the $8^{th}$ International Conference on Neural Networks and their Applications (NEURAP'95/96)*, pp. 117–123. (Marseille, FRA).

Weber V., Wermter S. (1996) Using hybrid connectionist learning for speech/language analysis. In: Wermter S., Riloff E., Scheler G. (Eds.) *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pp. 87–101. Springer-Verlag, Berlin.

Wermter S. (1995) *Hybrid Connectionist Natural Language Processing.* Chapman and Hall, Thompson International, London, UK.

Wermter S., Löchel M. (1996) Learning dialog act processing. *Proceedings of the $16^{th}$International Conference on Computational Linguistics.* (Kopenhagen, Denmark).

Wermter S., Riloff E., Scheler G. (Eds.) (1996) *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing.* Springer, Berlin.

Wermter S., Weber V. (1996) Artificial Neural Networks for Automatic Knowledge Acquisition in Multiple Real–World Language Domains. *Proceedings of the $8^{th}$ International Conference on Neural Networks and their Applications (NEURAP'95/96)*, pp. 289–296. (Marseille, FRA).