Distribution and Acoustic Features of Extra Linguistic Sounds in the Waxholm Corpus

Atelach Alemu Argaw, Eva Forsbom, Ebba Gustavii

December 22, 2004

1 Introduction

The presence of extra-linguistic sounds is claimed by many to contribute to a better performance of speech recognition systems. They can be included in language models and help to improve recognition accuracy in spoken language by providing clues about production processes. Natural language generation for dialogue systems could also possibly be improved by including extra-linguistic sounds, not only to make the machine-to-human interaction seem more natural, but also to help the listener to decode the information being propagated.

In this paper, we analyze the syntactic distribution of extra-linguistic sounds, given the acoustic feature duration and the type of sound, in a Swedish spoken dialogue corpus, the Waxholm corpus. We aim at revealing potential trends in machine-to-human interaction, although the data is sparse and the tools are rather crude.

The paper is outlined as follows: in the background the concept of extra-lingustic sounds is introduced and previous research on their function and usage in speech processing systems is reviewed. In section 3, we present the data and methods used; section 3.2 gives an analysis of the syntactic distribution, and section 3.3 an analysis of the duration. Finally, in section 4, some concluding remarks are given.

2 Background

2.1 Extra-Linguistic Sounds

Spontaneous speech consists of both lexical and non-lexical (extra-linguistic) elements. Sundaram and Naryanan [14] classify characteristics of spontaneous speech into 3 extra-linguistic groups: paralinguistic cues (falsetto, whisper, creak, laughter giggle, cry/sob etc), disfluency patterns (words such as *okay*, *oh*, *so* and *well*; repetitions and filled pauses such as *uh* and *um*) and reflexes (throat clearing, sniff/gulp, toungue clicking, lip smacking and breathing). These extra-linguistic features are considered to be a major factor in discriminating spontaneous speech from written text.

Historically, there have been three main ways of viewing extra-linguistic sounds. The oldest view is the one promoted by Chomsky [4], who claims extra-linguistic sounds to be errors that lie outside language proper. According to this view, extra-linguistic sounds should be excluded from linguistic theory. The second view, much related to the first, states that although these sounds are errors, they are worthy of study for what they reveal about performance. The third view is one that takes some of these extra-linguistic sounds to be genuine parts of a language [5]. This seems to be the current point of view, as will be discussed in Section 2.2.

2.2 Functions of Extra-Linguistic Sounds

Following the view that extra-linguistic sounds actually are part of language, a great deal of research has been made in order to establish what the function of these is. Mostly, these studies have dealt with filled pauses only. In the following, some of the findings will be presented.

According to Donzel and Koopmans [7], humans use extra-linguistic features, pauses, to structure the continuation of their discourse. Filled pauses are used to provide time for lexical choice-making, planning of the upcoming discourse segment, or to transmit some implicit message, such as a request for attention.

The more options there are, the more likely it is that the speaker uses filled pauses. These pauses can thus be indicators of the strength of association between sequential linguistic units, but can also, alternatively, be interpreted in more cognitive terms as time for choosing among word or phrase options, or for making decisions about the next thought [11]. It is empirically supported that extra-linguistic sounds precede unpredictable lexical items rather than predictable ones [1, 13].

The presence of filled pauses may thus be an indication of word-searching problems, leading to conclusions such as words following a hesitation have a low transition probability and thus a high information value. This in turn helps listeners detect upcoming important linguistic materials [16]. Clark and Fox [5] show that the filled pauses *uh* and *um* are conventional English words, and speakers tend to plan for, formulate, and produce them as they do for any other word. They also report results from experiments on the London-Lund corpus of 170,000 words from 50 face to face conversations and show that there is a difference in their usage, *um* being followed by much longer delays and pauses than *uh*, and claim that *um* is preferably used in connection with larger decision making processes, whereas *uh* is used when lexical choices are needed. This information can be used to segment discourse units in speech recognition systems.

Swerts et al [16] report similar results as those presented by Clark and Fox [5]. With 46.5 minutes long recorded data of two female Dutch speakers which is annotated by nineteen subjects, they also investigate the correlation between filled pauses and discourse boundaries. They show that nasalized filled pauses (e.g. *um*) may be more typical at the onset of major discourse units than other filled pauses. They make an interesting claim that although the presence of a

hesitation can not predict a discourse boundary, the presence of a boundary makes a hesitation highly likely. They also present results from prosodic and contextual investigations. In the results, it is shown that filled pauses show a gradual roughly linear F0 fall and that they tend to be lower than surrounding words. Moreover, they conclude that stronger breaks in the flow of information are more likely to co-occur with filled pauses than weaker once, and that these pauses at stronger breaks also tend to be segmentally and prosodically different from the others, and are usually preceded and followed by silent pauses.

Although it has not been empirically proven, there are suggestive evidence that extra-linguistic sounds may help listeners in their process of comprehension [6]. However, it has been proven that the presence of these sounds contributes to a better performance of speech recognition systems (c.f.[10, 15]).

Furthermore, there has been some experiments which suggest that the analysis of these extra-linguistic sounds can be used for language generation. Natural language generation for dialogue systems could include extra-linguistic fillers not only to make the machine-to-human interaction seem more natural but also to transfer information pertaining to the context in the synthesized speech implicitly and help the listener to decode the information being propagated [15].

3 Analyses

In the analyses, we first look at where the extra-linguistic sounds occur, as regards syntactic categories (chunk and PoS labels) of the preceding and succeeding token. We then look at the duration of the extra-linguistic sounds in relation to their syntactic distribution. We would have liked to look at difference in distribution of nasal and non-nasal filled pauses, but there were too few instances of nasal extra-linguistic sounds.

3.1 Data

The Waxholm Swedish spoken dialogue corpus is collected by a research group at the Royal Institute of Technology (KTH) in Wizard of Oz experiments in a spoken dialogue system, Waxholm, that provides information on boat traffic in the Stockholm archipelago. The collected data consist of utterance-sized speech files that are stored together with the text entered by the wizard and the corresponding phonetic labels. A complete log of the dialogue session is also stored. The acoustic-phonetic database includes phonetically rich reference sentences uttered by all subjects. A total of 66 different subjects (17 female, 49 male) in various age groups participated in the data collection. During the data collection, some 1900 dialogue utterances were recorded. The database contains 9200 words with a total recording time of 2 hours and 16 minutes, one third of which is labelled as pause. One fourth of the recording time pertains to the calibration and reference sentences. The calibration sentence, uttered by all subjects, was used to calibrate the hardware for the data collection. The 8 reference sentences included a rich variety of phonetical features to make

sure they were uttered by all subjects [2]. We only had access to the dialogue contributions of the subjects. In our analyses, therefore, we used the subjects' dialogue utterances for analysis of the data, and the phonetically rich reference sentences for normalization of interspeaker variation.

3.1.1 Extra-Linguistic Sounds in Waxholm

In the Waxholm spoken dialogue corpus, extra-linguistic sounds were labeled manually during the post editing of the data. The considered extra-linguistic categories in the corpus are interrupted words, inhalations, exhalations, clicks, laughter, lip smacks, hesitations and hawkings [2]. The authors also give a general picture of the distribution of these sounds according to their position in the utterances. It is summarized in Table 1, which is taken from their article.

Extra-Linguistic Sound	Number of occurrence	Most Common Position
Inserted vowel	230	Word final
Smack	152	In conjunction with place names
Inhalation	117	Utterance initial
Hesitation	67	Utterance or sentence initial
Exhalation	60	Utterance final
Interrupted word	32	In conjunction with place names
Click	7	In conjunction with place names
Laughter	4	In conjunction with place names
Sigh	1	In conjunction with place names
Hawking	1	In conjunction with place names

Table 1: Extra-linguistic sounds in the Waxholm corpus

3.2 Syntactic Distribution

To automatically approximate the syntactic distribution of the extra-linguistic sounds, we annotated the data with part-of-speech tags and phrasal chunks.[1] Though a functional description of the constituents would have been more interesting, we had to limit our definition of syntactic context due to the lack of a general-purpose Swedish parser and grammar. We used the TnT-tagger [3], trained on the Stockholm-Umeå Corpus [8] (with the PAROLE tag set[2]), and a rule-based Swedish chunker [9]. The chunker uses nine phrase categories: ADVP, AP, APMAX, NUMP, NP, NPMAX, PP, VC and INFP. We used all phrase categories except for NPMAX (maximal noun phrase projections incorporating post-attributes), since those were reported to cause a lot of errors. The phrasal

[1] In the syntactic distribution analysis interspreaker variation has not been taken into consideration.

[2] A listing with examples can be found at `http://spraakdata.gu.se/lb/parole/sgml2suc`.

labels are attached to each word and signals its position in the phrase structure hierarchy. Words at the beginning of a phrase are labeled XB, words occurring phrase-internally are labeled XI and words outside of any phrase (such as interjections) are labeled O. Since a word may be part of several phrases at different hierarchical levels, it may be tagged with more than one phrasal tag. The lowest node is always attached closest to the word, followed by the next-lowest node, and so on. The prepositional phrase *on the table*, would accordingly be labeled on_PB the_NP_PI table_NI_PI. [9]

The PoS tags and chunk labels at the positions immediately preceding and succeeding the extra-linguistic sounds, were manually verified and corrected.

Extra-Linguistic Sound	Number of occurrence
Smack	190
Inhalation	142
Hesitation	95
Exhalation	67
Interrupted word	43
Click	8
Laughter	3
Sigh	1
Hawking	3

Table 2: Frequency of extra-linguistic sounds in the Waxholm corpus (dialogue utterances)

317 out of the total 1747 utterances (18%) are initiated by one or more extra-linguistic sounds.[3] Most of these are succeeded by an NP which is mostly constituted by the first person nominative pronoun. This however, is nothing particular for utterances initiated by extra-linguistic sounds; the most frequent way of beginning utterances in the corpus in general, is to use an NP of this kind. A closer look at the phrase types and the part-of-speech tags appearing at the beginning of the utterances, reveals a close resemblance between those initiated by one or more extra-linguistic sound, and those that are not. The extra-linguistic sounds that predominantly occur at utterance initial position are hesitations, smacks, inhalations, hawkings and sighs.

An overview of the distribution of the extra-linguistic sounds is given in Table 3. Taken altogether, almost 70% of the extra-linguistic sounds appear at utterance initial position. Only 12% of the extra-linguistic sounds occur at utterance final position. These are mostly preceded by proper nouns, something which is true also for the utterances which are not ended with any extra-linguistic sounds. The extra-linguistic sounds that predominantly occur at utterance final position are, quite naturally, exhalations and laughters.

[3] In Table 3, each sound included in a complex of extra-linguistic sounds is counted, so the total is somewhat larger (383).

Type	Utterance position	Frequency	Most freq. prec. ph.	Most freq. succ. ph.	Most freq. prec. PoS	Most freq succ. PoS
Hesitation	Initial	$\frac{60}{95}$ = 63%	START	NPB (38%)	START	PF@USS@S (28%)
Hesitation	Final	$\frac{2}{95}$ = 2%	—	END	—	END
Hesitation	Internal	$\frac{34}{95}$ = 36%	PPB (41%)	NPB_PPI (41%)	SPS (47%)	NP00N@0S (32%)
Smack	Initial	$\frac{172}{190}$ = 91%	START	NPB (49%)	START	PF@USS@S (30%)
Smack	Final	$\frac{1}{190}$ = 1%	START	END	START	END
Smack	Internal	$\frac{18}{190}$ = 9%	PPB (39%)	NPB_PPI (39%)	SPS (39%)	NP00N@0S (28%)
Inhalation	Initial	$\frac{136}{142}$ = 96%	START	NPB (38%)	START	RH0S (24%)
Inhalation	Final	$\frac{0}{142}$ = 0%	NONE	END	NONE	END
Inhalation	Internal	$\frac{6}{142}$ = 4%	NPI (33%)	—	—	—
Exhalation	Initial	$\frac{2}{67}$ = 3%	START	ADVPB (100%)	START	RH0S (100%)
Exhalation	Final	$\frac{50}{67}$ = 75%	NPB_PPI (54%)	END	NP00N@0S (36%)	END
Exhalation	Internal	$\frac{15}{67}$ = 22%	PPB (27%)	NPB (33%)	SPS (27%)	—
Laughter	Initial	$\frac{0}{4}$ = 0%	START	NONE	START	NONE
Laughter	Final	$\frac{4}{4}$ = 100%	NPB	END	—	END
Laughter	Internal	$\frac{0}{4}$ = 0%	NONE	NONE	NONE	NONE
Hawking	Initial	$\frac{4}{4}$ = 100%	START	NPB (100%)	START	PF@USS@S (75%)
Hawking	Final	$\frac{0}{4}$ = 0%	NONE	END	NONE	END
Hawking	Internal	$\frac{0}{4}$ = 0%	NONE	NONE	NONE	NONE
Clicking	Initial	$\frac{3}{8}$ = 38%	START	NPB (100%)	START	—
Clicking	Final	$\frac{3}{8}$ = 38%	—	END	—	END
Clicking	Internal	$\frac{2}{8}$ =25%	—	—	—	—
Int. words	Initial	$\frac{6}{43}$ = 14%	START	—	START	—
Int. words	Final	$\frac{7}{43}$ = 16%	O (71%)	END	SPS (57%)	END
Int. words	Internal	$\frac{30}{43}$ = 70%	PPB (33%)	NPB_PPI (33%)	SPS (33%)	SPS (20%)
All	Initial	$\frac{383}{554}$ = 69%	START	NPB (43%)	START	PF@USS@S (27%)
All	Final	$\frac{67}{554}$ = 12%	NPB_PPI (42%)	END	NP00N@0S (28%)	END
All	Internal	$\frac{106}{554}$ = 19%	PPB (33%)	NPB_PPI (33%)	SPS (35%)	NP00N@0S (20%)

Table 3: Syntactic distribution of extra-linguistic sounds in the Waxholm corpus

19% of the extra-linguistic sounds occur somewhere in the middle of an utterance; mostly succeeding a preposition, and preceding a proper noun. When merging consecutive extra-linguistic sounds, there are 92 utterance internal occurrences. 33% of the instances occur inside prepositional phrases, immediately succeeding the preposition. 15% occur at the border between a noun phrase and a prepositional phrase, and 9% occur inside noun phrases. The rest of the positions are infrequent (5% or lower).

In utterance internal position, interrupted words and hesitations (i.e. the two disfluency patterns), show similar context. This underlines the suggestion that disfluency patterns occur more often in certain contexts.

But, maybe the most interesting pattern in these data, is the rather striking resemblance of smacks and hesitations. Although smacks occur more often in initial position, both smacks and hesitations are used in similar contexts in utterance initial and utterance internal position. In the literature, smacks are considered as reflexes, in line with inhalations and the like, whereas hesitations are considered as disfluency patterns, i.e. more voluntarily produced sounds. The similarity seems to indicate that smacks ought to be better understood as disfluencies.

3.3 Duration

Several of the cited experiments (e.g. [12, 5, 16]) imply that longer extra-linguistic sounds (including any following silent pauses) should be indicators of

more global decision-making processes, such as planning the next part of the utterance, whereas shorter sounds should be indicators of more local decision-making, i.e. lexical choice. If this holds, it is plausible to assume that duration of the extra-linguistic sound could give clues on the syntactic context. For instance, pauses preceding content words could indicate either lexical choice or planning. Pauses preceding function words, however, would almost always indicate more serious decision-making processes, since the set of function words is very limited.

In our analysis, we normalized the duration, in order to cater for speaker variation, using a simple weighting scheme based on average sentence duration. The phonetically rich sentences were used to calculate the average. For each sentence in this set, each speaker's sentence duration was extracted and an average was calculated. The weights were then obtained by taking the average duration to speaker sentence duration ratio.

We looked at the duration of extra-linguistic sounds (including any succeeding silent pauses) before function and content words to see if that revealed any patterns. To do that, we classified each PoS tag as either a content or a function word. All verbs, adjectives and nouns, and adverbs with the comparison feature were classified as content words, and the rest as function words. This may, of course, be a too coarse-grained division; auxiliary verbs, for instance, are better labelled as function words, but the PAROLE tag set does not make that distinction.

The results are shown in Figure 1. There does not seem to be any differences in duration before content and function words, respectively, except for interrupted words.[4] Rather, for the other sounds, the duration is more or less the same at positions preceding content and function words given the type of sound. Given that the succeeding silent pauses are included in the calculated duration, this is an interesting phenomenon.

Since many researchers have been working on hesitations (e.g. *um* and *uh*), we took a closer look at them (see Figures 2 and 3). For succeeding parts-of-speech, there is only one evident pattern: the duration was somewhat longer for adverbs (RG0S). For succeeding phrase labels, duration seemed to be shorter within a noun phrase than elsewhere. This might indicate that only lexical search, and no planning, occurs there.

4 Concluding Remarks

In this paper, we have analyzed the syntactic distribution of extra-linguistic sounds, given the acoustic feature duration and the type of sound in the Waxholm corpus. Most of the extra-linguistic sounds appear at utterance-initial position. Some of the extra-linguistic sounds tend to occur at utterance-initial position: hesitations, smacks, inhalations, hawkings, and sighs; while some tend to occur at utterance-final position: exhalations and laughter. Interrupted words tend to occur utterance-internally.

[4] This also holds for utterance internal and utterance initial occurences treated separately.

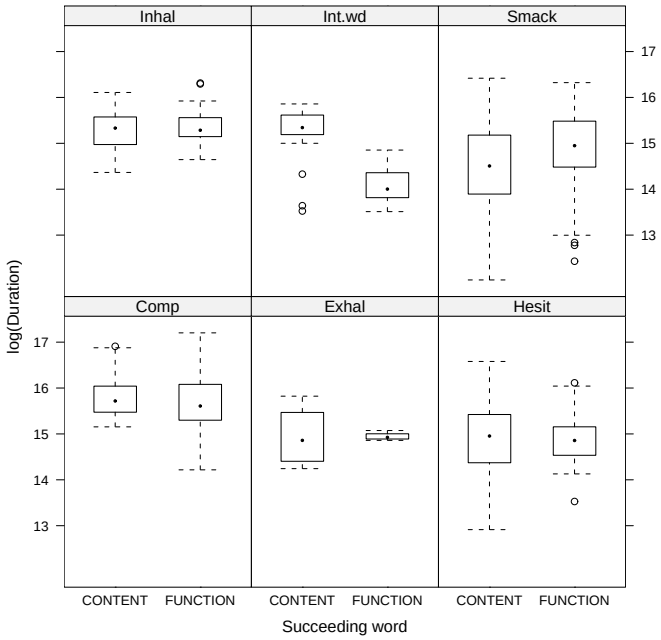

Figure 1: Duration (log) as a function of succeeding content or function word.

One third of the utterance-internal extra-linguistic sounds occurred within a prepositional phrase, immediately following the preposition.

Three of the extra-linguistic sounds show similar distribution in utterance-internal position, namely hesitations, smacks and interrupted words. Although both hesitations and interrupted words are classified as disfluencies, we have not seen any reports that they should have similar distribution, as was shown here in utterance-internal position. The similarity in distribution of hesitations and interrupted words, i.e. the two sounds that are previously classified as disfluencies (which are the more voluntary produced extra-linguistic sounds in our data[5]), might be an indication that these sounds are not produced randomly, but rather have a function.

Maybe the most interesting pattern in these data, is the rather striking resemblance of smacks and hesitations in relation to syntactic context. The similarity seems to indicate that smacks ought to be better understood as disfluencies rather than reflexes.

The duration analysis did not reveal much, except that duration seemed to be shorter within a noun phrase than elsewhere for utterance-internally occurring hesitations. This might indicate that only lexical search, and no planning, occurs there.

[5]Laughter excluded.

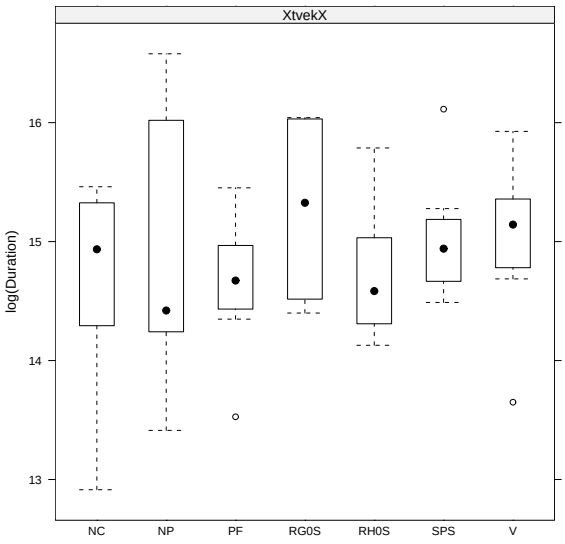

Figure 2: Duration (log) as a function of succeeding PoS for hesitations.

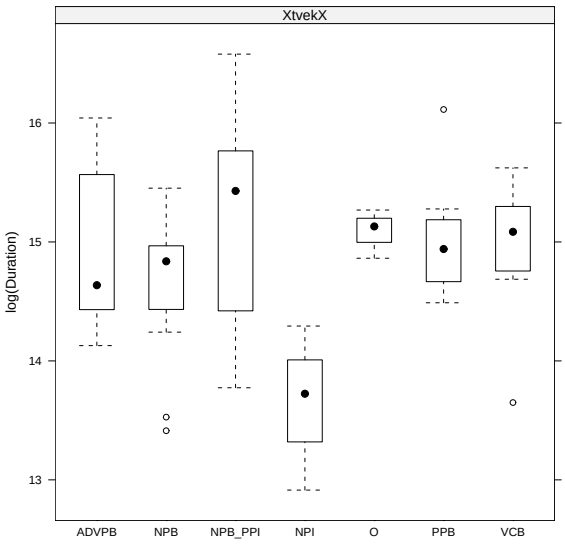

Figure 3: Duration (log) as a function of succeeding phrase for hesitations.

References

[1] G. Beattie and B. Butterworth. Contextual probability and word frequency as determinants of pauses in spontaneous speech': Effects of age, relationship, topic, role, and gender. In *Language and Speech*, 1979.

[2] J. Bertenstam, M. Blomberg, R. Carlson, K. Elenius, B. Granström, J. Gustafson, S. Hunnicutt, J. Högberg, R. Lindell, L. Neovius, L. Nord, A. de Serpa-Leitao, and N. Ström. Spoken dialogue data collected in the Waxholm project. Technical report.

[3] T. Brants. TnT - a statistical part-of-speech tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference*, Seattle, Washington, 2002.

[4] N. Chomsky. *Aspects of Theory of Syntax.* MIT Press, Cambridge, MA, 1965.

[5] H. H. Clark and J. E. Fox Tree. Using uh and um in spontaneous speaking. In *COGNITION*, Elsevier Science B.V., 2002.

[6] M. Corley and R. J. Hartsuiker. Hesitation in speech can ... um... help a listener understand. In *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*, 2003.

[7] M. E. v. Donzel and F. J. Koopmans-van Beinum. Pausing strategies in discourse in Dutch. In *IFA-Publications, ICSLP96*, 1996.

[8] E. Ejerhed, G. Källgren, O. Wennstedt, and M. Åström. The linguistic annotation system of the Stockholm-Umeå project. Technical report, Department of General Linguistics, University of Umeå, 1992.

[9] B. Megyesi. *Data-Driven Syntactic Analysis. Methods and Applications for Swedish.* TRITA-TMH 2002:7. Institution for Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Nov. 2002. PhD thesis.

[10] S. Pakhomov and G. Savova. Filled pause distribution and modelling in quasi-spontaneous speech. In *Proceedings of the ICPhS Satellite Meeting on Disfluency in Spontaneous Speech*, Berkeley, CA, 1999.

[11] S. Schachter, N. Christenfeld, B. Ravina, and F. Bilous. Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology*, 1991.

[12] E. Shriberg. Phonetic consequences of speech disfluency. In *Symposium on The Phonetics of Spontaneous Speech, Proc. International Congress of Phonetic Sciences*, volume 1, San Francisco, 1999.

[13] E. Shriberg, R. Bates, and A. Stolcke. A prosody-only decision-tree model for disfluency detection. In *Proceedings of Eurospeech*, volume 5, Rhodes, Greece, 1997.

[14] S. Sundaram and S. Narayanan. Experiments in the synthesis of spontaneous monologues. In *IEEE 2002 workshop in Text to Speech Synthesis*, Santa Monica, CA, 2002.

[15] S. Sundaram and S. Narayanan. An empirical text transformation method for spontaneous speech synthesizers. In *Proceedings of Eurospeech*, Geneva, 2003.

[16] M. Swerts, A. Wichmann, and R.-J. Beun. Filled pauses as markers of discourse structure. In *COGNITION*, Elsevier Science B.V., 2002.