What eye movements can tell us about spoken-language processing: a psycholinguistic survey
(Andrea Weber, Saarbrücken)

1 Introduction
Language is a central characteristic of human beings. Without language, neither human intelligence nor our existence as social beings is imaginable (e.g., Kintsch, 1999). It is therefore not surprising that the study of language has fascinated researchers for a long time. While the interest of a linguist lies in describing and analyzing the product language, the interest of a psycholinguist lies in the mental representations of language and in the processes by which humans access these representations to comprehend and produce language. Given the different goals, it is not surprising that linguists and psycholinguists draw on different types of linguistic evidence for theory building. In this paper, I will introduce eye-tracking as a psycholinguistic method to obtain linguistic evidence. Psycholinguists work mainly with quantitative data obtained from experiments. While controlled experiments certainly increase the validity of results, psycholinguistic paradigms often require a relatively unnatural testing situation. Eye-tracking, however, allows a more natural task and setting than most behavioral paradigms. In the last decade, the number of eye-tracking studies in psycholinguistics has literally exploded, and the paradigm is regularly used now in a variety of research areas including language comprehension and production on the word level, sentence level, and dialog level. I will describe in the following how eye-tracking can be used to address issues in spoken-language comprehension, concentrating on the two levels word recognition and sentence processing. The goal of this paper is to introduce readers who are unfamiliar with the paradigm to eye-tracking and to show how eye-tracking can effectively be used to address interesting research questions in psycholinguistics. To this end I have selected representative studies on word recognition and sentence processing which will serve the intention.

2 Eye-tracking
During everyday tasks that involve vision, people rapidly shift their eyes to bring task-relevant regions of interest into the central area of the fovea. Eye movements are necessary because visual sensitivity differs across the retina; acuity is the greatest in the central portion of the fovea.
and then markedly declines. Eye movements (saccades) are followed by fixations during which visual information can be taken in. Even though Cooper published a seminal study in 1974 on the close time-locking of eye movements and relevant information in speech (Cooper, 1974), it took another two decades before the psycholinguistic community started to fully embrace the paradigm. During a typical eye-tracking experiment (head-mounted SMI EyeLink system), participants’ eye movements to objects on a computer display are monitored while they are listening to recorded speech which is played over loud speakers (see Figure 1). The eye-tracking output contains locations of fixations with a sampling rate of 250 Hz, thereby informing us about which objects participants were looking at what point in time. Typical analyses of the output include the calculation of fixation proportions for different displayed objects, fixation latencies, or duration and number of inspections of objects.

![Figure 1: Participant wearing a head-mounted eye tracker.](image)

### 2.1 Word recognition

In order to understand spoken language, listeners must recover the meaning of an utterance from the acoustic speech signal. During an early stage of this comprehension process, they have to map the speech signal onto a set of discrete linguistic representations that are stored in our mental lexicon. It is generally assumed that the linguistic representations are sequences of sounds that form the words of a language. Main issues in spoken-word recognition research concern the mapping of the signal to stored lexical representations and the types of mechanisms listeners use to access words from their memory.
2.1.1 Competition between candidate words

A mechanism for lexical access that all current models agree upon is the multiple activation of lexical hypotheses. As the word *steak* is heard, words with similar sounds like *stay, stale, stain, take,* and *ache* will be considered in parallel with *steak.* How strongly a candidate is activated depends on the goodness of fit between the speech input and the stored representation. Words that no longer match the speech input as it unfolds are no longer activated. Concurrent activation of candidate words (e.g., Marslen-Wilson & Welsh, 1978; Zwitserlood, 1989) is followed by competition among those candidates. The more candidates are competing, the slower word recognition proceeds. In principle, the right choice between word candidates can be made purely on the goodness of fit between the speech input and the stored representation; the word which matches the speech input best is selected, and all other candidates are rejected. Competition, however, acts to sharpen the distinctions in lexical activation and facilitates the selection of the right candidate.

A seminal study which showed competition between activated candidate words was published by McQueen, Norris, and Cutler in 1994. In their study they used the word-spotting paradigm in which listeners hear nonsense sequences that have sometimes real words embedded. Whenever listeners spot an embedded word they press a button and say the word aloud; reaction times for the button press serve as the dependent variable. McQueen and colleagues found that it is harder for listeners (i.e., longer reaction times) to spot short words embedded at the beginning of longer words, such as *sack* in *sacrif* (the beginning of sacrifice), than to spot short words embedded at the beginning of a nonsense sequence such as *sack* in *sacrik* (no English word begins with *sacrik*; McQueen, Norris, & Cutler, 1994). This effect was assumed to reflect the competition between the short word sack and the long word sacrifice.

It can be argued, however, that listening to nonsense sequences and pressing a button is not what we do when we listen to speech outside the laboratory; also longer reaction times for embedded words with competitors in the lexicon, are only indirect evidence for the competition process. More direct evidence for competition was found in 1995 by Tanenhaus, Spivey-Knowlton, Eberhard, and Sedivy with the eye-tracking paradigm. Tanenhaus and colleagues were concerned with changes in the effective competitor population of words as a function of unfolding phonetic input. They had listeners wear a head-mounted camera which tracked the movements of their eyes as they carried out simple instructions, e.g. to click on one of several objects. The displayed objects included a target object on which listeners were instructed to
click, a competitor object with a name that was similar in onset with the target, and two distractor objects with unrelated names. They observed that a display containing, for example, both the target *candy* and the competitor *candle* attracts looks to both these objects as listeners hear *Click on the cand-* (see Figure 2). Listeners will look at the specific target object *candle* as later incoming phonetic information distinguishes between the two competitors. The interesting aspect of the paradigm is that it provides a window into the listeners’ processing before it is fully certain which word is being heard, when alternative word candidates still compete for recognition.

![Figure 2: Example of a display.](image)

Once the Tanenhaus et al. study was published in *Science*, it was important to prove the validity of the paradigm for psycholinguistic research. A number of studies showed that the results represent natural language processing and are not only generated by the visual display of objects. Dahan, Magnuson, and Tanenhaus (2001) showed, for example, that fixation proportions to competitor pictures are influenced by the lexical frequency of the picture names. Competitor pictures with high frequency names are fixated more often than competitor pictures with low frequency names (thereby replicating a frequency effect that has been found before with reaction time paradigms). Thus even though both competitors were displayed, they did not perceive an equal amount of looks; rather a linguistic feature of the names influenced fixation patterns.
2.1.2 Competition in a second language

For a specific domain in word recognition, the eye-tracking paradigm has demonstrated to be particularly useful; namely the recognition of words in second language. The processes involved in spoken-word recognition are presumably universal. Thus we assume that there will be competition during spoken-word recognition in a non-native language just as in the native language. The question at issue is thus not whether such competition occurs, but how much competition occurs in a second language. In principle, listeners might be faster at recognizing words in a second language. After all, they generally know fewer words in a second language and the fewer candidates they can activate, the less competition there will be and the faster recognition proceeds (e.g., Norris, McQueen, & Cutler, 1995; Vroomen & Gelder, 1995). This is, however, not what listeners subjectively experience. Listening to one’s native language is effortless; but listening to a second language can be distressingly hard. It is feasible that part of this effort is linked to a modulation of the competitor set.

Weber and Cutler (2004) investigated lexical competition in a second language using the eye-tracking paradigm. In their study, native Dutch participants, who were fluent in English, were listening to English instructions to click on pictures on a computer screen. A target picture was one of four displayed pictures. Eye fixation patterns revealed two potential sources of added competition for L2 listeners. First, the listeners’ native vocabulary apparently adds competition in non-native listening. When the Dutch name of a competitor picture was phonologically related to the name of the English target (kist, ‘box’, when hearing the English target kitten), Dutch listeners fixated the competitor picture more than they fixated unrelated distractor pictures (see Figure 3). Second, L2 listeners’ phonetic discrimination difficulties cause inappropriate competitor activation. When the English name of a competitor picture contained a vowel that is confusable for Dutch listeners (pencil, /ɛ/, when hearing the target panda, /æ/), Dutch listeners fixated this competitor picture longer than they fixated less confusable distractor pictures (beetle, /i/, when hearing the target bottle, /o/). Native English listeners did not consider the confusable competitor pictures as potential targets. This suggests that language-specific difficulties of the Dutch listeners with this phonological vowel contrast rather than acoustic similarity between /ɛ/ and /æ/ had caused the inappropriate competitor activation. In sum, even though the available vocabulary in the non-native language may be smaller, the extent of lexical competition is in fact larger. Monitoring eye movements to pictures made it possible to investigate word recognition in a second
language without the listeners’ native language being overtly used during the experiment.

![Figure 3: Average fixation proportions of Dutch listeners 300-800 ms after target onset for competitors and distractors.](image)

**2.2 Sentence processing**

Once individual words have been identified in an utterance, the parser begins to build the syntactic structure, that is the hierarchical organization of the utterance’s constituents. The words and structure of an utterance are essential to yield its interpretation. Building a structured, interpretable representation of an utterance is substantially different from the processes underlying word recognition. The sounds and words of a language presumably come from a finite set, and processing mechanisms mainly include mapping the speech signal onto stored representations. Understanding the meaning of a sentence as a whole is not a mapping process since there is an infinite number of sentences in a language and it is therefore basically impossible that full sentences are stored in the mental lexicon. Otherwise we would have difficulties understanding sentences we have never heard before. Research in sentence processing investigates the mechanisms responsible for the computation of structural representations and their interpretation.
2.2.1 Attachment ambiguity

A basic finding is that sentence processing is incremental. That is, we structure and interpret the words as they are perceived rather than store them as a list and combine them later. A seminal finding for incrementality was published in 1973 by Marslen-Wilson who showed in a speech-shadowing experiment that syntactic and semantic information is available to participants as they repeat the speech they hear; that is, even for close shadowers constructive errors were usually grammatically suitable with respect to the preceding context. This suggests that the shadowers’ performance was based on a syntactic analysis of the ongoing speech stream. There is, however, a potential cost to incremental processing. Sequences are often ambiguous; that is, they are compatible with more than one well-formed structural representation. For example, in the phrase Betty knew Monica’s date, Monica’s date could be the direct object of knew or could become the subject of a clausal complement (Betty knew Monica’s date had bought flowers). Disambiguating information may occur in later parts of the sentence, but due to incrementality, processing must proceed before such relevant information becomes available. A great deal of research has therefore focused on the processing of local ambiguities as a means for investigating the kinds of information and strategies listeners employ during the earliest stages of sentence processing.

A well studied ambiguity in the field of language comprehension is the PP-attachment ambiguity. In the utterance Put the apple on the towel, the phrase on the towel can be taken either as Location (VP-attachment) indicating where to put the apple, or it can be taken as Modifier (NP-attachment) indicating which apple to take. In 1986, Ferreira and Clifton used the self-paced reading paradigm to investigate which interpretation the parser initially chooses. In a self-paced reading experiment, participants see written sentence parts on a screen and press a button when they have read the sentence part, the next part is then displayed; longer reaction times for button presses are assumed to indicate processing difficulties (e.g., re-analysis of initial interpretation). Ferreira and Clifton compared temporarily ambiguous sentences such as Put the apple on the towel in the box with unambiguous sentences such as Put the apple that’s on the towel in the box and found longer reading times for in (the presumed point of re-analysis) in the ambiguous sentences. This was taken as evidence that participants initially interpreted on the towel as verb argument and had to revise their interpretation during the subsequent in.

However, longer reaction times are only a measurement for general processing difficulties, they can not inform us about what is being
processed and how it is being processed. Also having to press a button in order to be given the next part of a sentence is probably not very natural. Both pitfalls can be avoided with the eye-tracking paradigm. In 1995, Tanenhaus and colleagues investigated the same attachment ambiguity using the eye-tracker (Tanenhaus et al., 1995). On a display they showed an apple on a towel, an empty towel, a box, and a distractor object. They found that during the phrase on the towel, listeners looked at the empty towel rather than at the apple sitting on the towel, revealing a Location interpretation of the phrase. However, Tanenhaus and colleagues also found that the initially adopted interpretation is dependent upon the visual context. In a scene contrasting two apples of which only one is on a towel, people preferentially interpreted the phrase on the towel as Modifier of the apple.

2.2.2 Prosody in sentence processing

A further potential information source in spoken sentence processing (besides structural, lexical, thematic, and discourse information) is prosody. Prosody is the description of phrasing, stress, loudness, and the placement and nature of pitch accents in spoken language. It can express or aid a range of functions in communication: mark the difference between immediately relevant vs. background information, express contrast, contradiction, correction, or even indicate the intended syntax of ambiguous utterances. Prosody is different from the other information sources in that it is highly variable in its realization. There is, for instance, no clear simple and direct correspondence between syntactic and prosodic structures. Quite often, a speaker can choose between a number of different intonation contours to express a particular communicative function. Nevertheless, it has been shown that listeners rely on prosodic information in sentence processing. On a structural level, for example, evidence has been presented that prosody can guide listeners’ interpretation of attachment ambiguities (e.g., Kjeelgaard & Speer, 1999). Sentences with early closure (When Roger leaves the house it’s dark) were compared with late closure sentences (When Roger leaves the house is dark), and using a variety of experimental tasks it was shown that sentences with cooperating prosody (i.e. with a prosodic boundary after leaves in the early closure sentence) were processed more quickly than those with baseline prosody. Sentences with conflicting prosody were processed more slowly than those with baseline prosody.

Weber, Grice, and Crocker (2006) examined the role of prosody in a different ambiguity type, namely constitute order ambiguity in German. German uses morphological case to mark grammatical functions. Although four cases can be distinguished, the system often features
syncretism: In many NPs, nominative and accusative case share surface form (e.g., both the accusative form and the nominative form of the feminine definite article is *die*). Thus, case marking often fails to discriminate between two frequent grammatical functions, namely subject (nominative case) and object (accusative case). Furthermore, both Subject-verb-Object and Object-verb-Subject surface orders are possible in German. In an eye-tracking study, Weber et al. (2006) examined whether prosody can fill the functional gap arising from a combination of syncretism and free constituent order in German. Can prosody, in the absence of unambiguous morphological and configurational information, influence the assignment of grammatical function?

In the action-based variant of the eye-tracking paradigm used in the word-recognition studies described above, participants follow instructions to click on objects presented on a computer display. The timing and pattern of fixations to potential referents as they are being heard, are used to draw inferences about lexical access. But while processing sentences, listeners can also form expectations about upcoming arguments that they have not yet heard based on the currently available information in a sentence; empirical evidence shows that listeners make eye movements in anticipation that a picture in a display will become relevant. For example, upon hearing *the boy will eat*, listeners start looking at edible objects even before they are mentioned (Altmann & Kamide, 1999). Anticipatory eye movements can thus inform us about higher-level processes, such as the role of verb information in restricting the domain of subsequent reference. Weber et al. (2006) made use of anticipatory eye movements to explore the role of prosody in interpreting grammatical functions.

In their study, German listeners were presented with scenes depicting three potential referents while hearing ambiguous SVO and OVS sentence beginnings (*Die Katze jagt...*, ‘the cat[NOM,ACC] chases...’). While case marking on the first NP was ambiguous, clear case marking on the second NP disambiguated sentences towards SVO (*den Vogel, ‘the bird[ACC]’) or OVS (*der Hund, ‘the dog[NOM]’). Listeners interpreted case-ambiguous NPIs as Subject more often, and thus expected an Object as upcoming argument, but only when sentence beginnings carried an SVO-type intonation. This was revealed by more anticipatory eye movements to suitable Patients (Objects) than Agents (Subjects) in the visual scenes. No such preference was found when sentence beginnings had an OVS-type intonation. Prosodic cues were apparently interpreted rapidly enough to affect listeners’ interpretation of grammatical function before disambiguating case information was available.
3 Summary

In the last decade, the study of eye movements has become an important and productive way for investigating spoken-language comprehension and production in psycholinguistics. Indeed, a rapidly increasing number of publications attests to the significance of the paradigm. In this chapter, I want to introduce the reader to the basic principles of the eye-tracking methodology, using exemplary studies from the area of word recognition and sentence processing as showcases.

The fundamental idea behind eye-tracking is that fixations to objects on a screen can reveal insights about the processing of language. The tight link between eye movements and speech can thereby inform us about listeners’ interpretation of words or sentences while they are being heard. On the level of spoken-word recognition this implies, for example, that we can investigate which lexical candidates are activated before incoming acoustic input clearly distinguishes a target word from competing alternatives. Among other things, evidence for subcategorical effects on the competition process (e.g., McMurray, Tanenhaus, Aslin, & Spivey, 2003), semantic category effects (e.g., Yee & Sedivy, 2006), morphosyntactic effects (e.g., Dahan, Swingley, Tanenhaus, & Magnuson, 2000), prosodic effects (e.g., Dahan, Tanenhaus, & Chambers, 2002), and effects of the native lexicon in non-native listening (e.g., Weber & Cutler, 2004) has been found. On the level of sentence processing, the eye-tracking paradigm can inform us how the human parser initially interprets syntactic and semantic ambiguities. As for the word level, effects of morphosyntactic information (e.g., Kamide, Scheepers, & Altmann, 2003) and prosodic information (e.g., Weber et al., 2006) have been found, but inter alia effects of referential context (e.g., Tanenhaus et al., 1995) and constraining semantic verb information (e.g., Altmann & Kamide, 1999) were observed for the processing of spoken sentences.

A key advantage of eye-tracking over other behavioral on-line paradigms are certainly the almost continuous measurements as language processing carries on. The fact that eye-tracking can be used for a great variety of research topics using relatively natural tasks further make it a valuable paradigm in psycholinguistics from which we can expect exciting findings in the years to come.
3 References


