Speech Comprehension: Theoretical approaches and neural correlates

Bachelor Degree Project in Cognitive Neuroscience
15 ECTS
Spring term 2015

Magnus Roos

Supervisor: Daniel Broman
Examiner: Judith Annett
Declaration of authorship


Author name: Magnus Roos

The above noted work is submitted to the School of Bioscience at the University of Skövde, as a final year Bachelor project toward the degree of Bachelor of Science (B.Sc.) in Cognitive Neuroscience. The project has been supervised by Daniel Broman.

I, Magnus Roos, hereby declare that:

1. The above noted work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any other degree.

2. The above noted work is the result of my own investigations, except where otherwise stated. Where corrections services have been used, the extent and nature of the corrections have been clearly marked.

Magnus Roos 4/6-2015

Signature Date
Abstract

This review has examined the spatial and temporal neural activation of speech comprehension. Six theories on speech comprehension were selected and reviewed. The most fundamental structures for speech comprehension are the superior temporal gyrus, the fusiform gyrus, the temporal pole, the temporoparietal junction, and the inferior frontal gyrus. Considering temporal aspects of processes, the N400 ERP effect indicates semantic violations, and the P600 indicates re-evaluation of a word due to ambiguity or syntax error. The dual-route processing model provides the most accurate account of neural correlates and streams of activation necessary for speech comprehension, while also being compatible with both the reviewed studies and the reviewed theories. The integrated theory of language production and comprehension provides a contemporary theory of speech production and comprehension with roots in computational neuroscience, which in conjunction with the dual-route processing model could drive the fields of language and neuroscience even further forward.

*Keywords:* Language, speech comprehension, dual-route processing model, trace model, superior temporal gyrus, inferior frontal gyrus, N400, P600
## Table of Contents

Abstract............................................................................................................................................... 2
Introduction......................................................................................................................................... 4
Spoken Language................................................................................................................................ 7
    Aspects of Speech.................................................................................................................... 9
Neural Correlates of Speech Comprehension.................................................................................... 11
    Temporal Aspects of Neural Processes for Speech Comprehension........................................ 17
Theoretical Models of Speech Comprehension................................................................................ 22
    Dual-route processing model................................................................................................. 22
    A temporo-frontal network on sentence comprehension..................................................... 25
    MUC framework.................................................................................................................... 27
    TRACE model....................................................................................................................... 31
    Interactive alignment model.................................................................................................. 34
    An integrated theory on language production and comprehension....................................... 35
Discussion......................................................................................................................................... 37
References......................................................................................................................................... 49
Appendix........................................................................................................................................... 56
Introduction

Many species communicate in one way or another. Birds chirp, dogs bark, cows moo, and so on. Some may call such utterances language, and many humans may not hesitate when it comes to assigning those sounds as sophisticated use of language, as seen in all kinds of popular media. As Liberman and Whalen (2000) argue, however, language use is a skill exclusive to humans, or at best to humans and closely related primates. In a study of tamarin monkeys (Hauser, Newport, & Aslin, 2001) it was shown that the non-human animals were able to segment, process, and retain information from speech quickly in time, and Savage-Rumbaugh et al. (1993) found that bonobo monkeys (at least the two tested) showed sensitivity to semantics, syntax, and word order. Also, the monkeys were able to act correctly when presented with simple requests. In spite of the findings in non-human primates, though, both studies (Hauser et al., 2001; Savage-Rumbaugh et al., 1993) notes that human language yet involves more complexity, whereas the primates tested are only hypothesized to be able to grasp simple language systems.

Even though limited language understanding has been seen in non-human animals, humans will for the purposes here be assumed to be the only species yet capable of using language as means for sophisticated communication. Therefore, the discussion of whether the language domain should be extended to also include other species will have to be held elsewhere and language will henceforth be narrowed down to human usage only. Language has seemed to fascinate, and elude, for over a century. Broca and Wernicke are two of most famous pioneers in the field, active already in the 19th century. Broca is probably most widely known for study of the patient called Tan, who had lost the capability of producing coherent speech. The patient, however, could understand speech, and had many other intact mental functions. Wernicke, on the other hand, noticed patients with impaired speech comprehension, but intact speech production. Damage to the left inferior frontal areas investigated by Broca, or the posterior superior temporal areas that Wernicke examined
produce what came to be known as Broca's aphasia and Wernicke's aphasia, respectively. Therefore, the field is neither new, nor lacking relevant research. However, the complexity of the field, even if restricted to seeing language as something only accessible to humans, makes it a challenging area to approach in the first place. Even defining the relevant aspects to examine can be problematic. Some of these issues will be presented shortly, and in more detail in the discussion.

This review is aimed at taking a contemporary look at language, and more specifically speech comprehension. In the introduction a very broad presentation of the field will be given. This will then be narrowed down to spoken language perception. Various aspects defining and constructing language will be discussed, models of speech perception considered, and research findings indicating areas of involved neural activation will be reviewed. Prominent, classical, and current publications will be examined and discussed, with conclusions to be drawn from there.

To be able to fully explore the extremely broad and cross-disciplinary field of language some definitions are needed. Language is defined in the Oxford Dictionaries as how humans communicate through the sophisticated usage of words. The Merriam-Webster Dictionary defines language as how the usage of words are understood across a unified body of individuals. Therefore the fundamental assumption can be drawn here that conventional and sophisticated language is created by humans, and used by humans to allow communication between humans. If language is exclusively human, then, what does a language consist of? Based on the dictionary definitions above, language could roughly be explained as systematic usage of words in a way that is understood by a recipient. A related idea is the one between language and thought, and what the relation between the internal and external use of language is. One response can be that ones inner voice often may be represented by the same words that also could be spoken. Benjamin Lee Whorf came up with a theory for the relation between thought and language. As summarized in Eysenck and Keane (2010), the three Whorfian hypotheses suggests that language influences thinking on
different levels; the strongest hypothesis states that language in itself determines what can be thought, the weaker hypothesis states that language influences perception, but not thought processes, and the weakest hypothesis holds that language only affects memory. Casasanto (2008), however, discards the Whorfian hypothesis. By using Pinker's (1994) book The Language Instinct, Casasanto (2008) extracts strong arguments against the Whorfian approach. An example of how Eskimos might not necessarily think of snow in more detail just because of a large variety of words exist to define the phenomenon is given. Casasanto also explains the circularity of Whorf's hypothesis – “the only evidence that people who talk differently also think differently is that they talk differently” (p.67) – therein claiming that the hypothesis lacks, and subsequently needs, some kind of empirical testability.

Leaving further discussion of the Whorfian Hypothesis, and whether language does indeed shape thought, if language is not necessarily directly connected to the intentions of the thinker, where does the meaning of words lie and stem from? According to Communication Theory (2010), the Triangle of Reference was put forward by Ogden and Richards (1923) in the book The Meaning of Meaning. The triangular model suggests that a symbol (a stimulus) can trigger thought and this thought then can be represented by a referent (i.e. words). Whereas the thought evoked by the symbol may not directly correspond to the referent, the evoked thought is always correct, while the referent is no more than adequate. The referent is thus neither more than a representation of the symbol itself, nor more than the thought that the referent evokes.

To summarize, language will only be treated here as the pragmatic means of communication between humans; how, and if, language influences thought patterns, and whether intention or meaning use words as vessels, or if words carry meaning or intention in themselves is not examined further – even though there seemingly are such philosophical problems yet unresolved.
Spoken Language

So far language has been treated as if either spoken or written. From here on, mainly auditory language, i.e. speech and listening to speech, will be discussed. To start, though, a brief anatomical overview of the auditory system will be given, as extracted from various textbooks (Coren, Ward, & Enns, 2004; Gazzaniga, Mangun, & Ivry, 2009; Kandel, 2012):

Speech is like any other sound in that it consists purely of vibrations. When a medium creates vibrations (sound), these vibrations are transported in space by colliding with nearby molecules – causing the ones that were collided into to vibrate, which therefore moves the vibrations, i.e. the sound, further in space, until eventually decaying when the vibrations have diminished by friction.

When such a vibration reach the outer ear it is channeled through the auditory canal. The canal, leading to the eardrum, amplifies certain frequencies – therefore making them easier to distinguish. The eardrum itself is a sensitive membrane-like wall which vibrates in accordance to the sound frequency. When past the eardrum, the outer ear is also passed, which is followed by the middle ear. The eardrum's vibrations causes a small bone, the hammer, which is attached directly to the eardrum, to vibrate as the eardrum did. The vibrations from the hammer spread to another small bone, the anvil, and from there along to the third and final bone, the stirrup. The stirrup is connected to the oval window, to which the vibrations lead. As with the outer ear, the middle ear structures also amplify the sound signal. Since the oval window is very small (1/15th the size of the eardrum) some vibrations need amplification in order to reach the threshold of actually being able to be heard. If the pressure on the oval window is too great, however, the angle is switched as to decrease the force and therefore lessen the pressure to protect the inner ear from potential damage.

The oval window is a gateway to one of three tubes that together form a spiral formed structure called the cochlea, which is where the inner ear is located. The tube from the oval window
is called the vestibular canal, with another large tube called the tympanic canal located nearby. These canals are filled with a fluid which relays the vibrations from the oval window. The vestibular canal is where the vibrations are lead into the cochlea, and the tympanic canal is where the pressure created by vibrations is released (via a flexible membrane called the round window). Directly connected to, and lying in between the vestibular and tympanic canals is the cochlear duct, which is third canal. In the cochlear duct lies the organ of Corti, which contains over 15,000 hair cells. Of those hair cells, about 3,000 are located on the inner side, and these hairs are directly connected to the auditory nerve, i.e. making this the place where sound is converted from vibration to electric potentials. As vibrations are provided from the pulsating oval window, making the fluid in the vestibular canal flow in accordance, the vibrations transfers to the cochlear duct since the separating membrane of the canals is only a couple of cells thick. The fluid in the cochlear duct is less viscous (thus vibrates more easily) than in the other canals, making it easier for the hairs in the organ of Corti to bend in the direction the fluid flows. When the hairs bend, an electrical potential is generated, which is called transduction, i.e. the transformation from physical stimuli to neural activation.

The activity created in the cochlear duct then travels along the right and left auditory nerves to the dorsal cochlear nucleus and the ventral cochlear nucleus of each hemisphere. From the cochlear nuclei, the activity spreads symmetrically in two directions: from the dorsal cochlear nucleus directly to the inferior colliculus, and from the ventral cochlear nucleus to the superior olivary nucleus. The superior olivary nucleus of each hemisphere then carries both right and left auditory nerve information, which is then projected to the inferior colliculus. The activity is lead to the medial geniculate, and then finally arrives at the primary auditory cortex. Structurally, the primary auditory cortex is located by the superior temporal gyrus, and the previously mentioned areas necessary for the coding of audition is located more medial and inferior to the primary
auditory cortex.

It is from the primary auditory cortex that the speech stream can begin to be analyzed, with the meaning to be extracted. First, though, more structural aspects of language are needed.

**Aspects of Speech**

Below, the aim is to determine which elements to take into account regarding speech, i.e. what speech consists of. A start will be by looking at which sub-divisions have been constructed in scientific studies and in preexisting literature on the subject, and then perhaps at least temporary definitions can be derived. Mazoyer, Tzourio, Frak, Syrota, Murayama, Levrier, et al. (1993), e.g., chose to separate acoustics, phonology, lexicality, prosody, syntax, and semantics in their study, trying to determine the cortical representation of speech. Ingram's (2007) textbook on neurolinguistics suggested a model that features most of the Mazoyer, et al., (1993) divisions, save for an inclusion of lexical items, and in addition featuring a stage of morphology in speech processing. In the same model, the first stage of post auditory encoding is the acoustic phonetic feature extraction, followed by segmental phonology perception, leading to morphology, which is the basis for lexical-semantic word comprehension and syntax.

Assuming all aspects named by Mazoyer et al., (1993) and Ingram (2007) are relevant and correct, some definitions and explanations are natural to follow (as extracted and interpreted from Mazoyer et al., 1993): **Acoustics** refer to the stream of sound, from which several discriminating features can be extracted; **Phonology** is the speech in itself, which can be split into phonemes which are respective to each language. **Lexicality** refers to single word identification, as opposed to whole sentences. **Prosody** is the manner in which speech is performed – not the speech itself, but aspects like pitch or gestures. **Syntax** is roughly equivalent to grammar, or the set of rules deciding the word-order of a sentence. **Semantics** refers to the factual meaning of what is spoken.
The definitions in the above paragraph are derived from the original article, and are how each aspect were used in that same study. Hence, they are not dictionary, or universal, definitions. Because of that, the definitions of those features may have to be altered somewhat, as they will be featured throughout the present thesis. Although new definitions are given, they are still generally similar to the definitions above.

a) Phonology is still the acoustics of speech, which can be deconstructed from whole sentences into single words, and further into the syllables and phonemes constructing those words (e.g. Mazoyer et al., 1993).

b) Lexical definitions are, as the name might imply, what is retrieved from the internal lexicon; the literal definition of a concept or a sentence (Ingram, 2007).

c) Prosody refers to intonation, gestures, or other non-verbal ways to influence what is spoken (Mazoyer et al., 1993).

d) Syntax is the grammatical form of used words and what decides in which order words are to follow, i.e. the way that sentences are built (Dapretto & Bookheimer, 1999).

e) Semantics is the meaning of whole sentences, or interpretation of the meaning of multiple words, or a non-lexical use of word as intended by a speaker. Therefore, as opposed to the lexical definitions, semantics can include context-dependent speaker-specific definitions or meanings as in sarcasm, irony, or other non-literal usage of words (e.g. Mazoyer et al., 1993; Dapretto & Bookheimer, 1999; Ingram, 2007).

f) Morphology (as mentioned in Ingram, 2007) refers to the identification of language; words, sentences, syllables, prosody, meaning, etc. As the morphology is only the identification, no interpretation per se is regarded as being morphologic. The term morphology can be viewed as an umbrella term, including the acoustic aspect of phonology – where the latter is of greater relevance here.
To put these definitions to actual use, each account will be considered regarding the sentence “the young bird has now left the nest”. The morphological account is what makes the words separable, able to be read or heard in isolation, and not only as ink drops or some murmuring sound. If the sentence is spoken, the phonology is what makes different words possible to distinguish. Syntax decided the order of the words, and for example as seen in “left”, where “leaved” would make use of a regular inflection (-ed ending for past tense) but still be incorrect due to other syntactical grammatical rules. The lexical meaning is that a (specific) bird no longer is in a (specific) nest, which would differ from the semantic meaning, which could be interpreted as how a human parent laments over that their child has moved out.

**Neural Correlates of Speech Comprehension**

The neural correlates of speech comprehension will be presented in this section. As the theoretical models will be presented in a separate section, the necessary correlates for speech comprehension will be reviewed here. The figure in the appendix may provide visually guided help in showing some of the areas mentioned below.

In a meta-analysis Vigneau et al., 2006, analyzed 129 scientific studies, from which peaking clusters were found for phonological processing of words, semantic processing, and sentence processing. In another large-scale meta-review, Price (2010), reviewed 100 fMRI studies on the functional anatomy of speech, using only publications from January through October in 2009. As both these meta-studies reach similar conclusions, both will be presented alongside each other.

For (pre-lexical) phonology, Price (2010) noted increased activation in bilateral superior temporal gyri, with an emphasized left lateralization that is suggested to occur due to top-down predictions instead of bottom-up processing. This activation can then spread along multiple pathways, leading in ventral, posterior, or anterior directions. Vigneau et al. (2006) noted that, for
activation by phonology, four out of six peak clusters in the temporal lobe were along the superior temporal sulcus, with the two other being located by the middle temporal sulcus. The latter study also included frontal activation to phonology, which showed that activation clusters were present in the superior central sulcus, inferior precentral gyrus and dorsal parts of the inferior frontal gyrus, medial/orbital frontal areas, and along the Sylvian fissure.

D'Ausilio, Craighero, and Fadiga (2012) examined the part that the frontal lobe may play during speech perception, and also provided support for how (pre-)motor areas activate when listening to articulated sounds. Motor areas can be activated both visually by seeing lip/tongue movement and hand gestures. Hearing speech can also activate the same motor areas as if the listener itself were to articulate what was being heard. An rTMS study is presented that showed a double dissociation between where the rTMS was applied to a test subject, and which speech sounds that were perceptually impaired for that same subject. This, D'Ausilio et al. (2012) suggests, provides support for that motor and sensory processes interact during speech perception.

For the semantic parts of speech processing, Price (2010) concluded that semantic processing activation spreads from the superior temporal areas activated by phonology to more sylvian areas and on through multiple pathways to adjacent structures close to the temporal pole, inferior temporal gyrus, and in parietal areas around the temporoparietal junction. It is also suggested that Heschl's gyrus might be a hub of a semantic network that spreads throughout cortex. In line with those results are Vigneau et al. (2006), who found semantic activation to overlap with phonological activation, while also spreading anteriorly to the temporal pole, posteriorly towards the angular gyrus, and inferiorly to the fusiform gyrus. In addition to the temporal areas, Vigneau et al. (2006) also noted frontal activation around the central sulcus, from the most superior parts, to more inferior orbital frontal and medial frontal areas.

To further discuss the correlates of the semantics part of language, Binder, Desai, Graves,
and Conant (2009) made a meta-analysis of 120 fMRI studies investigating the neural correlates of the semantic system. It is argued that angular gyrus (temporoparietal junction) is connected to association areas in the parietal lobe, and there has a part in higher-level processes. The angular gyrus, more specifically, is suggested to be at the top of a hierarchy of concept integration and retrieval. Binder et al. (2009) also suggest that medial temporal, inferior temporal, and fusiform areas play a role in crossmodal sensory integration, and concept retrieval, while also stating that the superior temporal gyrus is likely to not be part of something else (relating to language comprehension) than the purely acoustic and phonological part of speech perception.

The Binder et al. (2009) study also discusses frontal involvement in semantic processing, proposing that dorsomedial prefrontal cortex is important for retrieval of semantic information, especially when goal-directed and self-guided. The same authors discuss that the ventromedial prefrontal cortex may be important for the affective part of processing concepts. Also mentioned by Binder et al. (2009) is the left inferior frontal gyrus (LIFG), and how that structure is suggested to be important for articulatory planning and various phonological and syntactic processes, and also that the structure may have a fundamental role for a normally functioning working memory. Inferior frontal gyrus is therefore suggested to not directly be linked to the semantic processing of words, but Binder et al. (2009) point out that for subjects performing semantic tasks longer response time is seen when LIFG is affected by transcranial magnetic stimulation, where it is then suggested that LIFG's role in working memory may be important for integrating phonological stimuli with semantic meaning. The final area to be discussed by Binder et al. (2009) is the posterior cingulate gyrus, which is proposed to be a kind of interface between the episodic encoding systems and the systems responsible for semantic retrieval. Therefore, the posterior cingulate gyrus provides episodic memory encoding structures (hippocampal areas) with the relevant semantic information given the current sensory stimulation.
Regarding whole-sentence comprehension, Price (2010) concluded there are reports of four primary activation sites: anterior and posterior parts of the left middle temporal gyrus, bilateral anterior temporal poles, left angular gyrus, and the posterior cingulate. Vigneau et al. (2006) also reported, almost identically as Price (2010), that sentence processing activates areas also important for processing semantics, with those areas being the (left) temporal pole, angular gyrus, and middle temporal gyrus. Peaking activation was also found in the posterior parts of the superior temporal sulcus, and posterior middle temporal gyrus.

In a study by Wilson, Molnar-Szakacs, and Iacoboni (2008) auditory and audiovisual narrative speech comprehension were examined by fMRI. The test subjects were separated into two groups and one group listened to a woman recapping the events of a cartoon, and the other group were shown the cartoon while listening to the same spoken recap as the other group. Therefore, the auditory stimuli were exactly the same, and the only difference between the groups was whether or not the visual stimulus accompanied the auditory stimulus. Intersubject correlational analyses were then performed. The results implicates the importance of superior and anterior temporal areas, bilateral angular gyri, inferior frontal areas, and premotor cortex for both the auditory and audiovisual stimulus. The main neural activation difference between the groups were seen in that visual areas showed activation where the audiovisual stimulus were given, and that auditory-only stimulation generated greater superior temporal activation.

Summarizing the temporal lobe involvement of phonology, semantics, and sentence comprehension, both Vigneau et al. (2006) and Price (2010) found most activation in bilateral superior temporal areas for (pre-lexical) phonology. For semantics in the same studies, activation is more left-lateralized, and in more spread out areas; going from the superior temporal sulcus/ gyrus towards the temporal pole, and towards more parietal regions like the angular gyrus/temporoparietal junction, and also more inferiorly, to the fusiform gyrus. Looking at whole-sentence
comprehension, the activation can be seen to fan out even further, though overlapping largely with semantic areas. The most activation was still found in superior/middle temporal areas, though notable activation was also found in more anterior temporal and posterior temporal (bordering on parietal) areas.

Since there are other components involved in effective comprehension, such as syllable recognition, and syntactical processes, and other dimensions such as prosody, precaution needs to be taken to correctly include and order the parts necessary for speech comprehension. What is not explicitly stated in Price (2010) and Vigneau et al. (2006), though, is whether all three of the examined aspects (phonology, semantics, whole sentence processing) are equally dependent on bilateral activation, and how one activation pattern's laterality differs from another. Therefore, unless otherwise stated, what is presented is (at least) applicable primarily to the left hemisphere. That being said, there is evidence of the right hemisphere being involved in the perception and comprehension of speech as well.

The bilaterality of language is a hotly discussed topic, and whereas many publications mentioned above explicitly states their findings involve mainly left hemispheric regions, there are indications that some processes may be bilateral, or more reliant on right hemispheric structures. Hickok and Poeppel (2007) proposed a bilateral view on speech comprehension with a slight left emphasis, where the right hemispheric areas are suggested to be able to process acoustic (speech) discrimination, lexical, and semantic processing. Bozic, Fonteneau, Su, and Marslen-Wilson (2015) found support for some degree of syntactical and lexical processing in both hemispheres. The study used fMRI as brain imaging method. Test subjects were exposed to different grammatical forms of spoken words; bare stems (sing, wash, cough, carrot), inflected forms of bare stems (ending with -s), and phrases (I, you, we, a an, the) combined with bare stems. Using different forms of inflections, the authors hypothesized, different neural signatures should be generated. What was
found, for example, was that a larger activation was seen, bilaterally, in superior temporal gyrus and medial temporal gyrus, when processing regularly inflected verbs vs. non-regularly inflected verbs (opened, walked vs. sat, slept). In contrast to this, though, more syntactically complex phrases are stated to be more reliant on left hemispheric regions. A difference between complex and non-complex is that the left hemisphere is suggested to be specialized in segmentation of speech. Therefore, the left hemisphere is seemingly able to carry out processing of longer and more syntactically complex sentences. Regularly inflected words, and non-complex syntactical phrases, only requires lexical access to a few items for discrimination, which it then is argued that the right hemisphere in isolation can carry out successfully. Sollmann et al. (2014) used rTMS to induce virtual lesions in combination with an object-naming task to determine the influence of right hemispheric areas in speech production and comprehension, and concluded that right areas are important for single-word naming, as seen by hesitation, errors, or lack of response. The study notes some major limitations, but states that the large number of test subjects made the results statistically reliable. Right hemispheric areas of importance for speech comprehension are suggested to be the posterior superior temporal gyrus, especially in females, and medial frontal gyrus for higher-order language task performance in both genders. Regarding a potential factor that could influence language laterality, Knecht et al. (2000) found that there is a relation between handedness and hemispheric language-dominance. In 4% of the right-handed participants a right-dominant language distribution could be seen, while 27% of the left-handed participants had right-dominant hemispheric language distribution. Although the language dominant laterality might not have a direct impact on speech comprehension, it can be important to note that the laterality can differ between individuals both as something to keep in mind when attempting to have experimental control, and when performing studies on patients with lesions.

In a study by Chang et al. (2010) it is noted that posterior superior temporal gyrus activation
indicates how that structure can distinguish between different phonetic sounds, and that the evoked activity patterns are dependent on transitory temporal and spatially non-overlapping neural representations. It is discussed that the posterior superior temporal lobe therefore has a specialized role in distinguishing phonemes, which is supported as the structure is reported to have direct anatomical connections to areas important for lexical and semantic aspects of speech processing. The neural inflexibility that could be seen, i.e. that distinct areas activate when processing e.g. /ba/ vs. /da/ sounds, may be caused by so called neural commitment (e.g. Kuhl, 2004; Zhang, Kuhl, Imada, Kotani, & Tokura, 2005). Neural commitment suggests that there is a critical/sensitive period for language, in the same way as for example vision. This causes language sounds that are based on, or similar to, the native language to be easier to learn, while foreign sounds can be more difficult to distinguish (e.g. native Japanese speakers cannot separate the sounds of /r/ and /l/, as there is no phonemic difference between those sounds in Japanese (Kuhl, 2004)). The role, as suggested by Chang et al. (2010) regarding the posterior superior temporal gyrus, is also in line with previous results, further strengthening the indication that superior temporal areas may be crucial for phonemic discrimination.

**Temporal Aspects of Neural Processes in Speech Comprehension**

In the above section, the neural correlates that may be most fundamental for the comprehension of spoken language was specified. In this section, the temporal intervals for neural activity during speech comprehension will be reviewed in order to provide information of when neural areas activate to generate speech comprehension. In knowing the spatial loci of language processing, the temporal aspects could provide necessary information that may be linked to the spatial counterpart in hopes of shedding further light on the neural activity that enables speech comprehension.
According to numerous findings (e.g. van Berkum, Zwitserlood, Hagoort, & Brown, 2003; Egorova, Shtyrov, & Pulvermüller, 2013; Kung, Chwilla, & Schriefers, 2014; Brunnelière & Soto-Faraco, 2015; Egidi & Nusbaum, 2012; Kolk, Chwilla, van Herten, & Oor, 2003; Groppe et al., 2010), there exists an ERP element called the N400. This wave has been shown to occur during language-related tasks, and then often in semantics-related anomalies. The N400 is a negative peak of the EEG occurring at around 400 ms (between 220-600 ms according to Groppe et al., 2010) from a critical word onset. To test the sensitivity and specificity of this wave, various studies have been performed. An ERP study by van Berkum et al. (2003) concluded that when a critical word was lexical-contextually coherent, i.e. in a sentence with a critical word that, when viewed as a sentence in isolation, were not violating the lexical-semantic context, and it did subsequently not generate an N400. On the other hand, when the semantic context as a whole was interfered with by an anomalous critical word, an N400 response was shown. This was tested by two experiments. In the first one, the test subjects heard a number of short stories. For example:

*As agreed upon, Jane was to wake her sister and her brother at five o’clock in the morning.*

*But the sister had already washed herself, and the brother had even got dressed.* Jane told the brother that he was exceptionally *quick/slow* (van Berkum et al., 2003, p.703)

The last word is the critical word, which does or does not fit into the semantic context, depending on which test group that had the coherent or anomalous critical word for that sentence. Eighty stories were used, where 40 used anomalous critical words and 40 used coherent critical words. The critical words were reversed between the test groups. In the second experiment the sentence containing the critical word was shown in isolation, and would therefore not violate a semantic context, regardless of which critical word was used. Kung et al. (2014) used intonation sensitive
words in Cantonese Chinese to test pitch and intonation in combination with asking a question, i.e. using Cantonese Chinese words with different meaning depending on pitch, and combining this with the rise of pitch used when asking a question. This generated an N400 response when there was a semantic context that the critical word deviated from. Brunnelière and Soto-Faraco's (2015) study suggest that the N400 is sensitive to semantic expectation of the following word, meaning that when a word differs from the expectation, an N400 is generated. In the experiment sentences with high or low semantic constraint were presented, i.e. more or less predictable final word, and with deviating or normal accent of that final word. The authors found that less constraining semantic context generated a greater phonological mismatch response at around 100 ms, which also lasted longer than a typical N400 response. On the other hand, greater semantic context generated an N400 response, even though the final word was only phonologically, and not semantically, deviating from the context. The occurrence of the N400 is suggested to occur due to top-down predictions clashing with bottom-up processes.

According to other studies, there are various factors to take into account when measuring the N400. A study by van den Brink et al. (2012) showed that scores on an empathizing questionnaire (EQ) correlated with the amplitude of an N400. Test subjects had their ERP recorded while listening to incongruent sentences (a child's voice expressing a wish to drink alcohol, for example). The authors hypothesize that higher EQ scorers would have more top-down processing generated from expectations about the speaker to result in a greater N400, while lower EQ scorers would process bottom-up information first, and then relate it to social information. Another potential influence on the N400 effect is mood, according to Egidi and Nusbaum (2012). The study the authors carried out found that the N400 effect was lesser when the listener was in a mood consistent with the semantic discourse; i.e. when in a positive mood, a negative discourse is suggested to be harder to integrate, thus generating a greater N400. A negative mood in combination with negative discourse therefore
elicits a lesser N400, as the content is suggested to be easier to integrate in the current mood state. It is explained that sad mood might increase readiness to a negative outcome, or induce reliance on an association network more tuned to negative information, and that mood therefore may affect discourse integration. van Alpen and van Berkum (2009) conducted a study aimed at examining whether words within words would elicit an N400 effect. They found that the /pi/ (pie) segment in the word pirate, when put in a semantically related context, still yields an N400. Even though the integration of the word segment is limited, it is explained that “listeners do still retrieve the meaning of the embedded word and try to relate it to the sentential context.” (p.2624) This is interpreted as a reflection of the constant integrating of the speech perception system, performing goodness-of-fit checks on-line, to enable rapid comprehension, and, by extension, potentially making communication effective and quick.

Another aspect that is important to keep in mind is that the N400 effect may be triggered by a wide variety of tasks or stimulation. The N400 can arguably be generated by language processing in some sense, since it has been reported to be sensitive to lexical-semantic anomalies (van Berkum et al., 2003), deviating pitch resulting in semantic ambiguity (Kung et al., 2014), semantic prediction (Brunnelière & Soto–Faraco, 2015), empathic capability, which in practice is suggested to mean more social predictions are made by the listener (van den Brink et al., 2012), and when integrating discourse (Egidi & Nusbaum, 2012). Kolk et al. (2003), on the other hand, reported that semantic anomalies need not generate a N400, where it is explained that selection restrictions did not elicit an N400. Also relating to the complexity of the N400 is the study by Egorova et al. (2013) that displayed how ERP waves around 100-200 ms also can affect the N400 response. The Egorova et al. (2013) study measured single word communication as can be used when naming or requesting an item, measuring the ERP during the naming/requesting task. The findings indicate that the ERP, and therefore the neural processing, is indifferent to what type of linguistic input is received (of the
pragmatic (contextual) and semantic (lexical) types of information tested), and the authors therefore suggest that all types of linguistic information are processed simultaneously. The study found neither significant N400 nor P600 (see below) effects, which is speculated to be due to how the methods used in the study did not make use of linguistic violations, which therefore lessened the need to integrate information or monitor information across the time spans of the test.

Another ERP element is called the P600. In Kolk et al. (2003) the P600 effect is stated to be sensitive to semantic and syntactical anomalies in speech perception. Both Kung et al. (2014) and Kolk et al. (2003) have come to the conclusion that the P600 seems to be important for on-line judgment of erroneous or ambiguous word usage. Kung et al. (2014) found that the P600 effect was present when the voice pitch made a single word utterance ambiguous in Cantonese Chinese, but the effect was removed when that same word, pronounced similarly, was put into context. Instead, an N400 effect was found in that instance. The Kolk et al. (2003) study suggests the P600 occurs when the brain reattends to an unexpected linguistic unit to evaluate its correctness. Suggesting that the P600 occurs when monitoring, reattending, or judging the grammaticality of a word also seems to be consistent with the lack of P600, as in the Egorova et al. (2013) study, since no linguistic violation or ambiguity were present in the test circumstances, which therefore would have no need for re-evaluation of the heard word.

In a recent study Brennan, Lignos, Embick, and Roberts (2014) reports that effects of semantic priming could be seen about 270 ms after word onset, even though word pronunciation often took about twice as long. This is suggested to indicate that 270 ms is about the time it takes for a primed word to be lexically identified and integrated with top-down expectations. Therefore, the time it takes for a word to be lexically activated is between 150-270 ms, which is around the same time when the N400 wave is seen to begin (Brennan et al., 2014; Groppe et al., 2010). The P600 wave may subsequently indicate how top-down processes attend to and evaluates the
veridicality of a single word, in contrast to the N400 which reacts to contextual semantic errors.

Attempts to integrate the temporal aspects of language processing presented in this section with the spatial findings presented in the previous section will be made in the discussion.

Theoretical Models of speech perception

Dual-route processing model

Hickok and Poeppel (2000) presented a framework that would become a very important theory on speech comprehension. Since the original article (2000), the theory became strengthened and more detailed in two additional publications (2004, 2007) by the same authors. Currently, the theory is one of the most prominent and complete approaches to the correlates of speech comprehension. The support the theory has gained can be seen both in how it is referred to in the literature, and in studies testing its fundamentals (e.g. Saur et al., 2008; Murakami, Kell, Restle, Ugawa, & Ziemann, 2015).

What Hickok and Poeppel (2000) presents is a dual auditory stream route framework of how and where spoken language is represented in the brain. The basis for this model is similar to the dual-route processing model as for vision (e.g. Goodale & Milner, 1992). That analogy can therefore give an idea of how this model of language is built.

The dual-route model of language comprehension (Hickok & Poeppel, 2000, 2004) suggests that there are (at least) two separate streams handling heard speech. One, the ventral stream, is suggested to stem from auditory cortices and move posteriorly towards the temporoparietal junction. The ventral pathway is what allows the understanding of semantic meaning. The ventral stream can hence be compared to the 'what'-stream of the dual-stream model in vision. The other stream, the dorsal one, is involved in sub-lexical tasks – such as distinguishing syllables, and creating sensorimotor representations of spoken word (Hickok & Poeppel, 2000).
When the speech signal is first encoded, the superior temporal gyrus is bilaterally activated to process acoustic-phonetic properties of what is spoken (Hickok & Poeppel, 2004). What is emphasized is the bilateral activation, which is given support by lesion studies, and how left unilateral damage to superior temporal areas does not generate as debilitating effects as to be expected if primarily left hemisphere alone would be activated when processing speech. From superior temporal gyrus (STG) the activation then fans out into the two predominantly left-hemispheric streams. It is suggested, however, that while the activation is similar across the bilateral portions of STG, which phonetic-acoustic features that are being processed in each hemisphere differs. Various suggestions, involving for example that speech is asymmetrically analyzed in time (Hickok & Poeppel, 2004, p.80), founded by other authors, are presented, though no conclusive stance is taken.

As the ventral stream is the stream that enables semantic comprehension, that stream is also the one requiring access to the mental lexicon. Hickok and Poeppel (2004) propose that this is done via a left-dominant network that involves posterior inferior temporal areas, possibly extending to temporoparieto-occipital structures. As pointed out by the authors, comprehension on a sentence-level, also taking the syntactical dimension of sentence processing into consideration, the ventral stream is likely to employ some more anterior temporal areas as well. It is discussed that the temporoparietal junction may serve as an interface between sound and meaning, and how a system of cortically distributed networks actually is the basis for semantic understanding. Such a system is suggested to link phonological and semantic information, thus allowing the linking of spoken utterances to the semantic meaning of those words.

The dorsal stream (Hickok & Poeppel, 2004) is suggested to have an especially important function during language acquisition, as the dorsal stream is where the acoustics of speech are stored. Hence, a child can compare it's own utterances to (properly pronounced) ones stored in
memory that has previously been heard. Those retrievals are collected from the dorsal stream, and via that a child can develop and practice it's own pronunciation by trial and error. The dorsal stream is also proposed to overlap with the verbal working memory, i.e. an aspect of the working memory that exclusively stores spoken words for brief periods of time.

When instead looking at the sound-based speech representations themselves, as absent in pure word deafness-patients that have damage to the ipsilateral superior temporal lobes, it is described (Hickok & Poeppel, 2000) how such patients can not hear speech – while still having a normal auditory tone threshold. Brain imaging data is also reviewed, providing support for the suggestion that the posterior-superior temporal lobe, bilaterally, may be crucial for the creation of sound-based representations of speech. Hickok and Poeppel (2000, 2004) turn to the consideration of other aphasics as well to help identify and gain support for the underlying correlates subserving their suggested framework. Conduction aphasia is caused by supramarginal gyrus lesions, i.e. damage located around the border of the temporal and parietal lobes. The aphasia is suggested to cause difficulty mainly when producing speech, as left superior temporal areas is particularly important for phonemic processes during speech production. Another aphasia is transcortical sensory aphasia, which is caused by inferior posterior lesions to the left temporal lobe. Damage to these areas is stated to interfere with semantic processes, as shown by functioning speech production and repetition tasks, suggested to mean that phonological and syntactical processes are largely intact. Wernicke's aphasia is proposed to be regarded as a combination of transcortical sensory aphasia and conduction aphasia, with both symptoms and lesion locales overlapping. This results in both phonological and semantic deficits, and lesions spreading from superior posterior temporal areas to inferior posterior temporal areas. In summary; regarding word deafness, conduction aphasia, transcortical sensory aphasia and Wernicke's aphasia, it is concluded that symptoms of each aphasia, and the respective locations lesioned to result in the aphasia, can be
accounted for and explained using the presented model (Hickok & Poeppel, 2004).

In a later publication by Hickok and Poeppel (2007), several parts of the theory had been slightly altered. The most central themes were still present, and the changes mostly lie in that the (2007) publication emphasizes the bilaterality of speech processing even further. Hickok and Poeppel (2007) also establish that the right hemisphere mainly deals with information integration over longer time timescales, subsequently suggesting that integrating information over shorter timescales instead are taking place bilaterally. What follows is a suggestion to leave behind the “traditional view of the left hemisphere being uniquely specialized for processing fast temporal information” (pp.396-397).

The dual-route processing model of speech (Hickok & Poeppel, 2000, 2004, 2007) therefore suggests that speech is processed using two streams. Initially, both hemispheres process the phonology and syntax of words, while the two left hemisphere- emphasized streams spread from posterior superior temporal cortex. The ventral stream spreads posteriorly to process the semantics of speech, and the dorsal stream spreads anterior-superiorly and is important in creating sensorimotor representations and maintaining the verbal working memory.

**A temporo-frontal network model of sentence comprehension**

Friederici (2002) published a review in which she reviewed current neuroimaging studies, using those studies as a basis to construct a model on spoken language processing, and more specifically sentence comprehension. The model introduces the fundamental areas and the corresponding time intervals, which in turn can indicate the order of processing, and thus provide clues to better understanding of sentence processing.

Two primary views are presented (Friederici, 2002) for the binding problem of sentences, i.e. how phonology, syntax, prosody, and semantics can be merged into coherent and fluent,
understandable speech, which also can be understood very quickly in time. The two views hold that either that syntax is processed prior to semantics, or that all information types are processed simultaneously, and both views are empirically supported.

In Friederici's (2002) model, speech is processed during three phases. Phase one, lasting the first 100-300 ms after stimulus exposure, is where the first syntactic details of the utterance are traced, based on each word's respective word category. Phase two, from 300-500 ms, is when semantic/lexical meaning is analyzed, and further syntactical processing takes place. Phase three, lasting 500-1000 ms, is where all information processed in phases one and two is integrated and merged, and thus conceived as comprehensible speech. This suggestion is therefore compatible with both views presented above; spoken information is both processed temporally separate and is also integrated simultaneously. It is pointed out, though, that the possible role of prosody-syntax interaction is not specified temporally (p.79).

Anatomically, Friederici (2002), suggests that sentence processing is based on a bilateral fronto-temporal network. A rough overview suggests that left temporal areas process lexical definitions and sentence structure, such as phonetic and syntactical elements. Left frontal areas integrate and analyze syntactic and lexical information to base semantic and thematic understanding. Right temporal areas is proposed to play a role in prosodic processing and interpreting, while the right frontal areas analyzes other details, like sentence melody. Areas contributing to sentence comprehension, as opposed to single-word understanding, are suggested to be the middle temporal, the angular, and the left inferior gyri. Areas specific for the semantics of sentence processing are indicated to include the left inferior frontal, and the right superior temporal gyri. Left posterior temporal regions are also proposed to be important for semantic processing of sentences. For executive and higher-order functions during sentence processing frontal cortical areas are proposed to be fundamental. Syntax is suggested involve activation of inferior frontal
cortex and anterior temporal regions. Unspecified as to what aspect of sentence processing, Friederici (2002) also reports that sentence processing can generate activation in both anterior and posterior areas of the temporal lobe. Summarized, sentence processing is suggested to be based primarily on left temporal and left frontal activation.

Temporally, Friederici (2002) argues that semantic processing takes place about 400 ms after stimulus onset, as seen in the N400, discussed above, which was seen to be elicited primarily during semantic violations. Preceding, and following, this semantic processing is the syntactical processing. Syntax is suggested to be shown by two waves; one negative wave (called left-anterior negativity, or LAN) that occurs between 100-500 ms after stimulus onset, and a late positive wave, the P600, which occurs 600-1000 ms after stimulus onset. Both the LAN and P600 have been seen to react to syntactical violations, which may indicate that syntactical processing takes place at those time intervals.

The prosodic account of processing is reported to be dependent on right temporal areas, such as superior temporal gyrus, but also right frontal areas, including the frontal operculum. Though a lot less empirically established, the conclusion is drawn that prosodic processing is based on right temporal and right frontal hemispheric activation (Friederici, 2002).

To conclude, Friederici (2002) suggests a model of sentence processing that is dependent on both right and left hemispheres, engaging mainly temporal and frontal areas. The aspects processed in each area, though, suggests that sentence processing is highly modular and temporally independent until all elements are integrated in the left frontal areas.

MUC framework

Broca's area, or left inferior frontal gyrus, has been shown (e.g. D'Ausilio et al., 2012; Smirnov et al., 2014) to play a pivotal role for language. Hagoort's (2005) publication provides a
proposal on the fundamental structure of language production and comprehension, and the related neural areas, with particular emphasis on the left inferior frontal gyrus.

The MUC model focuses on the binding, or unification, between different aspects of language. Hagoort (2005) suggests there are three components needed for functional language processing: memory, unification, and control – hence the name; MUC. It is argued that the memory is what allows for the storage and retrieval of language items from the mental lexicon. The author continues by stating that unification is the process of integrating multiple lexical items into a coherent utterance, and that control is the aspect of articulating the unified string of words. Control is also proposed to include the monitoring of speech, and to maintain the flexibility between speaking and listening when partaking in dialogue.

Focusing mainly on unification, Hagoort (2005) proposes there is unification taking place on different levels in a language. This three-layered structure requires unification on each level. The main level of unification is the syntactical one (with the other levels being the phonological and semantic/conceptual ones). It is suggested that all syntactical information and the syntactical, and grammatical, rules of sentence construction are originating from the mental lexicon. Hagoort (2005) claims that

[the] parsing account is 'lexicalist' in the sense that all syntactic nodes (e.g. S [subject], NP [noun phrase], VP [verb phrase], N [noun], V [verb], etc.) are retrieved from the mental lexicon. In other words, chunks of syntactic structure are stored in memory and there are no syntactic rules that introduce additional nodes. (p.417)

With each word instance, there are associated rules. Hagoort (2005) explains; the process of comprehension decodes, or unifies, every word of an utterance, adding each new syntax rule
instance to the previous rules and therefore narrowing the utterance down to fewer and fewer plausible semantic alternatives for each word that is processed. Since every word is associated with a limited set of nodes, there is also a limited amount of plausible outcome. In case of ambiguity the possible alternatives for the ambiguous node is kept pending until more information is available, where the then discarded option is then prevented from being selected.

Leaving the syntactical unification and instead moving to the semantic/conceptual and phonological unification, Hagoort (2005) proposes that the unification of semantics and phonology are individual and separate processes from the syntactical unification. Stating that computational models on semantic/phonological unification are less developed than their syntactical counterpart, unification is still equally as important for semantics and phonology as it is for syntax. The author suggests that semantic unification takes place in close relation to syntactical unification; as a sentence is being processed syntactically it then becomes processed semantically, directly following the syntactical processing. As the subsequent sentences are being processed syntactically, the following semantic processing narrows down alternatives and chooses a plausible, and/or contextually relevant, alternative that relates to the preceding interpretation to enable coherent understanding between utterances.

The phonological level of processing is argued (Hagoort, 2005) to be unified into acoustic, and more specifically intonational, expressions based on the underlying lexical elements. The then unified sounds are analyzed based on their characteristics (prosodic voice cues) to determine which utterance that should be focused on, or be processed in a different way from what syntax or semantics in isolation could determine.

Looking at the neural correlates of the MUC framework, Hagoort (2005) emphasizes frontal lobe processes. A point that Hagoort (2005) establishes is that the left inferior frontal gyrus (LIFG) shares the same functional level as other parts of the prefrontal cortex. The LIFG is also argued to
be the main language processing area in the frontal lobe. The prefrontal areas are then suggested to keep the spoken information (and possible interpretations) ready while also making a selection between the possible choices. This is argued to occur due to the ability of the prefrontal cortex to integrate multiple lines of information across varying time spans. The specific involvement of prefrontal areas during information selection is not mentioned, but LIFG is stated to be the primary unification area, i.e. the area where single word information is integrated into larger segments.

The unification is proposed (Hagoort, 2005) to work like this: Starting from the main hub, LIFG, syntactical information is retrieved from the superior temporal areas, post initial phonological processing, where the syntactical information is then analyzed to generate representations spanning across multiple words; thus allowing for semantic processing. Inside the LIFG, phonology is proposed to employ the superior areas (BA 44 & 6), semantic processing is suggested to take place more medially (BA 47 & 45), and syntactical information is argued to occur more inferiorly (BA 45 & 44).

The other two aspects of the MUC model, memory and control, are suggested to have their own respective neural areas important for maintaining the capability of processing said aspect. Memory, the author argues (Hagoort, 2005), stores phonological features, syntactical properties, and the associated semantic meaning, of words. Phonological processing is stated to activate areas from the posterior and central portions of the superior temporal gyrus to the adjacent areas in the superior temporal sulcus. Semantic processing of words, it is argued, activates more inferior areas, namely the left hemispheric middle and inferior temporal gyri. Syntactical processing is hypothesized to engage the left posterior superior temporal cortex. Hence, the memory aspect is proposed to take place mainly in the temporal lobe.

According to Hagoort (2005), control is the planning and execution of speaking and/or maintenance of functions related to communication (e.g. attentional control, or turn-taking during
dialogue). These higher-order functions are suggested to activate areas in the prefrontal cortex, and more specifically; anterior cingulate and dorsolateral prefrontal cortices are seen to activate during the planning of verbal actions and during attentional control tasks.

With all three MUC aspects discussed, it can be concluded that language acquisition employs the memory via associations between language concepts, syntactical features, and semantic meaning. Concepts, with their respective underlying associations, are therefore what needs to be unified in real time to make sense of on-line speaking. The final aspect, control, is what makes it possible to engage in effective communication between multiple parties. The neural areas involved are suggested to be temporal areas for initial acoustic and phonological processing. Activation can be seen to be more inferior whilst making out the segments that constitute words – and while connecting a word item to the corresponding semantic meaning. These temporal activations are then projected towards frontal areas, where the unification is suggested to occur, namely in the left inferior frontal gyrus. The projection of the temporal processing occurs in parallel as subsequent temporal lobe processes, as LIFG unifies multiple aspects; including phonology, syntax, and semantics. The ability to guide attention via focusing or ignoring, planning of an utterance, or self-monitoring makes up the control part, and is suggested to be based in prefrontal areas like anterior cingulate cortex, and dorsolateral prefrontal cortex.

**TRACE model of speech perception**

The TRACE model (McClelland & Elman, 1986) is a classic computational model of speech perception. It was developed from the basic assumptions of Marslen-Wilson and Tyler's (1980) cohort model on speech perception. In short, the cohort model states that spoken language is separated into segments consisting of each of the spoken words, where the segmented words in turn consists of the segments building those words (phonemes). When hearing a spoken word, each of
the uttered phonemes limits the availability of possible words that are being said. The spoken phonemes are then stored temporarily to inhibit the utterance from being heard as something else, while at the same time narrowing down the lexical possibilities by only making it possible to select one lexical item that matches the criteria, i.e. the non-inhibited phonemes. This, in turn, means the matching phonemes is paired to a specific lexical entry. Since there are more aspects to speech than what is being said, such as different contexts and prosody, the selection of a single word may actually be selected before phonemic information alone could have selected that same word. (Marslen-Wilson & Tyler, 1980; Eysenck & Keane, 2010)

The TRACE model by McClelland and Elman (1986) is similar to the Marslen-Wilson and Tyler's (1980) cohort model on at least one fundamental level, in that it assumes that speech processing systems do on-line narrowing to select a single corresponding entry by limiting the available options based on the acoustic information heard. The TRACE model divides speech into three major categories; the features, the phonemes, and the words.

The features are what defines, and builds, a phoneme. Phonemes, in turn, are what constructs words. A feature can be a pitch, or a defining characteristic of an utterance. Those segments, adding to the phonemes spoken, make up the individuality of words. What causes inhibition and selection to occur is because each feature, phoneme, and word, has a detector (what is directly associated with that feature, phoneme, or word). Activating those detectors excite word entries/phoneme combinations corresponding to what is heard, while the detectors also inhibit words that do not correspond, and thus making them unavailable. As the speech progresses, more and more detectors are activated, and more and more inhibitions are being made, until a single word is left and therefore selected as the word the speaker has uttered. The detectors on each level, feature, phoneme, and word, work independently, but can interact. This is a major change from, e.g. the Cohort model, which assumes a great importance of the initial part of a spoken word. Eysenck
and Keane (2010) states in relation to the (unrevised) cohort model that an unclear phoneme in the beginning of a word should cause the whole word to fail to become recognized. The TRACE model instead argues that detectors subsequent to the initial detectors still can identify a word, as long as enough detectors are activated to single out one word.

As explained in McClelland and Elman (1986), the levels (features, phonemes, words) each use

. . . competitive inhibitory interactions instead of bottom-up (or top-down) inhibition.

Competition allows the [TRACE] model to select the best interpretation available, settling for an imperfect one when no better one is available, but overriding poor ones when a good one is at hand. (McClelland & Elman, 1986, p.73)

This causes the various levels of bottom-up sensory input, and top-down predictions to each be a competitor for the selection of the most contextually relevant word, and in effect granting understanding of what is spoken. The (unrevised) cohort model (Marslen-Wilson & Tyler, 1980), for example, would argue that lack of input during the initial part of a spoken word would result in misunderstanding, while TRACE would argue that the best available option, in spite of lacking initial phonemes, is still better than none.

Thus, the trace part in the TRACE model comes from the so called imprint that words leave. A word's specific combination of features and phonemes is remembered by the system, which therefore helps to identify the word with subsequent uses, through inhibition of non-viable options as explained above. As this model is computational and has been tested (at least by the authors, and still holds up as a major and influential theory), it by that provides a (computationally) functional framework on speech perception.
Interactive alignment model

A publication by Pickering and Garrod (2004) proposed a novel approach to speech comprehension. Using so called situation models, first presented by Zwaan and Radvansky (1998), Pickering and Garrod (2004) developed a theory on speech comprehension. The interactive alignment theory holds that speech production and speech comprehension in fact should be looked at as a single integrated mechanism, and should hence be viewed as intimately intertwined aspects instead of independent phenomena. Therefore, as speech production and speech comprehension are suggested to be so closely related, Pickering and Garrod (2004) argue that dialogue is the basic form of verbal communication. Much research and many theories are explained to be primarily and almost exclusively reliant on monologue, where monologue is subsequently defined to be a non-basic aspect of language. In order to account for this previously unexplored idea, Pickering and Garrod's (2004) publication contains a proposal of a basic alignment model. Previously, the authors describe, dialogue was considered difficult to have experimental control over, and it is explained how it is hard to theorize the constant changes in, and individuality of, each dialogue. To make up for that the interactive alignment model takes the issues of experimental control and changing contexts into account.

In any dialogue (Pickering & Garrod, 2004), between the interlocutors, there is a situation model. This is the subjective representation of the situation, whatever it may be. That situation model may be represented semantically, and the situation model also has lexical and syntactical representations – since all these aspects are necessary for the production and comprehension of coherent language. These representations, in turn, corresponds to phonological representations, which can be verbalized to express the situation model. For each of the stages – semantic representation, syntactical representation, lexical representation, phonological representation,
phonetic representation – the listener (ideally) is in alignment to the speaker. When aligned on at least some level, dialogue is enabled. All the different alignment stages are independent of one another, but even though considered independent, they are still linked to each other. More of the alignment process is explained in the following paragraph.

An important aspect of the interactive alignment theory (Pickering & Garrod, 2004) is, as might be implied, the alignment. It is argued that, during normal conversation, interlocutors need to have both coordination and alignment. Coordination is explained to refer to when both parties mutually coordinates behavior accordingly, as explained by analogy to ballroom dancing (p.170). Thus, it is argued that the coordination is an important aspect of dialogue, to where both parties need to actively contribute and change accordingly, in order to be able to reach joint understanding. The alignment, on the other hand, is a largely automated process which is reached when the same representation is shared at the same time by both parties. When a level is aligned, it can causally influence other levels to also reach alignment, and this alignment is what enables successful dialogue. The automatic alignment is suggested to be a kind of primitive priming process. If a representation is erroneously made, and alignment is not reached, the authors suggest that another primitive mechanism can adjust the misalignment, to eventually reset the wrongfully set alignment level. There are other ways to reach alignment, and these are proposed to include theory of mind, and inferences of the other person's mental state. But, as such a strategy potentially is more costly than the automated process of alignment, it is rarely required other than when multiple ways to produce aligned representations have failed (Pickering & Garrod, 2004).

An integrated theory of language production and comprehension

The same authors as to the interactive alignment model recently published another proposal relating to language, though not with the focal point of the interactiveness of language (Pickering &
Garrod, 2004). Instead, their (2013) proposal is a theory in which language production and language comprehension are suggested to be highly integrated, as implied by the name of the article; an integrated theory of language production and comprehension.

In this model, speech production is made similar to an action, and speech comprehension made similar to perception, which in turn is drawn from Wolpert's (1997) proposal originating from computational neuroscience. In Wolpert's (1997) model, an action is stated to be accompanied by a feed-forward percept of the same action. Therefore, when performing the initiated action, that action is compared (by self-monitoring) to the forwarded, or predicted, model where the executed and predicted actions then should correspond to one another. Pickering and Garrod (2013) applies this model to language, suggesting that an utterance of speech is part action (speech production) and part perception (speech comprehension).

Pickering and Garrod (2013) suggest that a speaker, in planning the utterance, creates representations on semantics, syntax, and phonology levels. The semantics of an utterance is suggested to be constructed first, with syntax to follow, and the phonological representation is created last. These representations (semantics, syntax, phonology) are what is also verbalized, which, via on-line comprehension, is compared to the predicted utterance. This allows for the correction of errors, as a mismatch between action and perception implicates an alteration of the performed action from the predicted action.

One might ask how this relates to language comprehension, as in the understanding of another party's spoken words. The authors of the model (Pickering & Garrod, 2013) suggest that comprehension is done in ways resembling speech production. As a prediction of one's own utterance is created, a prediction of the utterance of another speaker is also created. Language is stated to often be very predictable, which therefore allows for feed-forward models of the respective semantic, syntactical, and phonological levels to be made by the listener, as the speaker speaks, thus
Speech Comprehension: Theories and Correlates

making it easier to comprehend dialogue in ongoing conversation. The way that the predictions are made available in the first place is by experiencing speech, where the speech is suggested to create associations and therefore facilitating prediction. The other way to ease the creation of relevant predictions is by producing own speech, which simulates possible routes of speech, which then in turn produces alternatives that can be used for the prediction of another speaker's utterances.

Simulation and association are hence two important aspects that helps to develop good predictions.

What is explained in the paragraph above is also where Pickering and Garrod (2013) tie in to their previous (Pickering & Garrod, 2004) publication containing the interactive alignment model. Since the proposed way to be able to carry out dialogue in the interactive alignment model was for the interlocutors to be aligned on various levels for mutual comprehension to be possible, this alignment is suggested in the (2013) publication to be possibly due to prediction. Pickering and Garrod (2013) exemplifies and summarizes as follows:

B’s prediction of what A is going to say is more likely to accord with what B would be likely to say if B spoke at that point. In other words, B’s prediction of B’s completion becomes a good proxy for B’s prediction of A’s completion . . . . In fact, linguistic joint action is more likely to be successful and well-coordinated than many other forms of joint action, precisely because the interlocutors communicate with each other and share the goal of mutual understanding. (p.345)

Discussion

This discussion section is aimed at widening the scope from earlier sections, making everything presented up to this point available for discussion. Limitations, comparisons, problematizations, and conclusions will be treated here.
At first, some methodological limitations will be addressed. The first major limitation worth pointing out is the complexity of the field. Neuroscientific findings can often be ambiguous and potentially hard to interpret. The field of language is equally (if not more) complex, as investigated variables may be arbitrary and uncontrolled factors could be influencing results. A merging of the two fields ought to be treated with great care and cautiousness. Since language is affected by a seemingly endless number of factors, which may or may not be of relevance, having experimental control can be difficult. Examples of possible influences can be prosody, semantic, or other contexts, test subjects’ relations or associations to words, syntactical complexity, and more. As these examples are general and broad, more pertinent and specific limitations are to follow. The difference between audiovisual speech is not addressed (Bernstein & Liebenthal, 2014). The usage of meta-reviews (Vigneau et al., 2006; Price, 2010) in establishing a literature-consistent basis may be viewed as improper, since the original studies are not treated or considered here. The potential difference between languages is not controlled for; studies performed in e.g. English (van Alpen & van Berkum, 2009), French (Mazoyer et al., 1993), German (Chang et al., 2010), Cantonese Chinese (Kung et al., 2014), and Dutch (van Berkum et al., 2003) are all used. Whether the topic selection of speech comprehension here in fact carries more relevance than if, for example, speech production or non-speech language comprehension would have been looked at is doubtful. Potential problems, such as determining if an inclusion of one dimension is more valid and relevant than another, perhaps could have been avoided if a wider selection had been made. Meanwhile, a more narrow selection could perhaps have given a better overview, with a more detailed look at one particular aspect. However, by keeping the approach at a more summarizing level, looking at the fields current status in regard to theories and broad activation patterns of speech comprehension, the attempted method is aimed at neither being too shallow, nor too deep.

A selection of which cognitive elements to properly examine can also be seen as a
limitation; as speech comprehension is made available by multimodal interplay, merely determining the involved systems might be seen as a challenge. Among others, the roles of attention, working memory, emotion, and social influences are seen to be important for speech comprehension. Perhaps some of those aspects ought to have been more deeply examined here, whereas none of those has been thoroughly considered instead. However, if too many elements affecting speech comprehension is included the result could have direct impact on the depth, and focus. The subject risks becoming scattered and lacking in depth if more aspects are included, rather than less.

A summary of the neural activation patterns put in effect when comprehending speech, see the appendix for reference, can perhaps be stated as such: bilateral primary auditory cortices receives auditory stimulus, the speech, where phonological decoding takes place in more posterior parts of the bilateral superior temporal gyrus. Sensorimotor representations corresponding to the heard speech are created in motor areas. For syntactical processes, activation was seen in left medial and temporal pole areas. Lexical and semantic processes use left medial, temporal pole and fusiform areas, and also more posterior areas around the temporoparietal junction. For integration of the different types of information the different areas project to the left inferior frontal lobe, with monitoring and attention guiding by dorsomedial and ventromedial prefrontal cortex.

Before moving on to the theories, a comment regarding the choice of theories is needed. Three of the theories (the dual-route processing model (Hickok & Poeppel, 2000, 2004, 2007), the MUC framework (Hagoort, 2005), and the temporo-frontal network model of sentence comprehension (Friederici, 2002)) provide a more emphasized correlate-basis in their frameworks. The three other theories (the TRACE model (McClelland & Elman, 1986), the interactive alignment model (Pickering & Garrod, 2004), and the integrated theory of speech production and comprehension (Pickering & Garrod, 2013)) provide more theoretical and computational frameworks, but does not present neural correlates on the same level as the other three. Very
broadly, each theory can be assigned to either explain *where* or *how* speech comprehension takes place. By selecting three of each kind, comparisons could be made both within each group, and also between the groups.

A brief note on how compatible the presented models are with each other is warranted. Similarities or differences may shed light on crucial aspects that perhaps are overlooked when the theories are seen one-by-one. The first major implication by all three of the presented where-models is that all propose solutions to unify/bind/integrate separate elements of speech into something coherent and comprehensible. Friederici (2002) suggests this is done by the processing resources available in the left inferior frontal areas. Hickok and Poeppel (2004) presents no locus specific for integrating all aspects of speech, though it may be derived that anterior temporal and inferior frontal regions are important, since those areas are suggested to be important for whole-sentence processing. This requires phonology, syntax, lexical access, and semantics to be integrated into one single coherent stream of information. Hagoort (2005) also emphasizes the importance of inferior frontal lobe, as that is suggested to be the primary location for the unification of his MUC framework. A key aspect all these models thus have in common, if not explicitly stated then implicitly indicated, is the importance of Broca's area, and its role. This is in line with what is stated in the section dealing with correlates where inferior frontal areas are also stated to be important for integrating processes (e.g. D'Ausilio et al., 2012).

The dual-stream model (Hickok & Poeppel, 2000, 2004, 2007) is perhaps the most well-developed of the three where-theories. In essence, the suggestion is that speech comprehension is based on two activation streams whose starting points is in the auditory cortex. Lexical comprehension makes use of the ventral stream, going posteriorly from the superior temporal areas towards the temporoparietal junction. The temporoparietal junction was suggested by Binder et al. (2009) to be the hub from which the association cortex was accessed, and lexical definitions are
retrieved. The dorsal stream is going anteriorly from the superior temporal regions to process syntax, which is mainly suggested to take place in areas around the temporal pole. The areas presented in the (2004) publication are consistent, or at least compatible, with both the MUC framework (Hagoort, 2005) and the frontotemporal network model (Friederici, 2002), and the basic assumption of the dual-stream model, the two streams, is compatible with the other models as well. The frontotemporal network model emphasizes the inferior frontal lobe-integration of speech elements, i.e. how prior processing ultimately leads up to the integration of different aspects of speech information. The dual-stream processing model, on the other hand, leaves the merging of the two streams less explained. The MUC framework does not explicitly suggest that there are two streams of information as Hickok and Poeppel (2000) does, though it can be derived from Hagoort's (2005) publication that the respective areas dealing with phonology, syntax, and lexical definitions in fact are compatible with the dual-stream suggestion's correlates. Both the MUC and the fronto-temporal network proposals assign the inferior frontal lobe the role of merging, or unifying, the various speech elements. Even though the MUC framework seems to emphasize a working memory aspect, in maintaining active word selection and inhibition processes, it still does not clash with the fronto-temporal network model. Because of this, it is plausible that all three of the presented models can be of relevance when further investigating the neural correlates of language, and not necessarily contradict each other even if one is proven correct. Perhaps the models should be viewed as validly based suggestions, where each has a different focal point.

Still focusing on the dual-route processing model, the framework has, over the years, become a very prominent and widely accepted theory on speech comprehension. Specht (2014), Saur et al. (2008), and Murakami et al. (2015) all investigate the ventral and/or dorsal pathways' contributions to speech processing, and all ended up supporting the assumptions of Hickok and Poeppel (2007). Perhaps the dual-route processing model is the best currently available
neurobiological framework for speech processing, which at least provides a view that is clear-cut and empirically testable. The MUC framework (Hagoort, 2005) suggests that unification is a key element for processing speech, and focuses more on the frontal areas involved in the making of unification than the entire chain from utterance to comprehension. Friederici's (2002) theory mainly suggests a possible solution to the binding problem of sentences. Neither Hagoort (2005), nor Friederici (2002) therefore addresses as many aspects as Hickok and Poeppel (2000, 2004, 2007), further strengthening that, at least of the theories reviewed here, the dual-route processing model provides a more complete framework for speech processing. That both the MUC framework (Hagoort, 2005) and the temporo-frontal network model (Friederici, 2002) are compatible with Hickok and Poeppel's (2007) model, can also indicate that the latter one is the more complete framework since it can envelop the other theories. The dual-route processing model may therefore be the most complete contemporary model available when focusing on speech comprehension.

Here, the where-models will be left for the moment. Instead, the how-models will be examined. At first, the theories' respective defining properties will be looked at. The interactive alignment model is based on how dialogue is the basis of language, which is, quite soundly, based on how language requires at least two parties in order to be functional. Dictionary definitions of communication (The Oxford Dictionaries' online dictionary, n.d.; Merriam-Webster's online dictionary, n.d.) state that a message is intended to be conveyed to an external party; i.e. to someone else. In order to able to successfully convey the intended message, then, Pickering and Garrod (2004) suggest that alignment between the speaker and the recipient is required. The TRACE model (McClelland & Elman, 1986) suggests that speech activates nodes corresponding to elements of the speech, where the activated nodes then inhibits other non-activated nodes. For comprehension, a selection is made from the non-inhibited nodes where the best available alternative is chosen, which is interpreted as what was heard. The integrated theory of language production and comprehension
(Pickering & Garrod, 2013) states that comprehension equals a kind of perception (and that speech production is an action), which then via feed-forward models and self-monitoring makes verbal communication possible. The action or speech production of a speaker is predicted by the recipient, where the prediction is compared to the on-line speech. If corresponding, comprehension of what was said is successful. If not corresponding, the perceived speech and the feed-forward model are compared and re-evaluated to reach an acceptable answer, or even result in a failure to comprehend altogether.

The interactive theory of language perception and production (Pickering & Garrod, 2013) and the TRACE model (McClelland & Elman, 1986) can be argued to be compatible. The feature/phoneme/word selection of TRACE can be roughly seen as equal to the integrated theory model's syntax/phoneme/semantics levels. As both theories also share the idea of how an activation of a node inhibits other nodes, and through that narrow the options available for selection, the main differences between the theories may be that the integrated theory on language perception and production also accounts for the production of speech and that it predicts on-line speech to a greater degree than TRACE. As TRACE is a computational model, and the integrated theory has roots in computational modeling the similarities may seem more justified, and perhaps it is not too much of a stretch to state that the TRACE model could be seen as a distant relative to the integrated theory of language perception and production. Assuming their compatibility, the interactive alignment model can also be argued to fit in, since the authors of the interactive theory state so themselves.

Pickering and Garrod's (2004, 2013) publications both propose that production and comprehension of language is highly integrated, and should be seen as two sides of the same phenomenon instead of separate from each other. While this may seem very appealing on paper, the question arises as to whether or not it is plausible in practice. Speech activates supplementary motor areas, as suggested in the dual-route stream model (e.g. Hickok & Poeppel, 2000) where the dorsal
Speech Comprehension: Theories and Correlates

stream goes upward to motor areas to create sensorimotor representations of the heard word. D'Ausilio et al. (2012), among other studies, also noted motor activation during speech perception. Motor representations may be important for correctly distinguishing and perceiving speech by having one's own motor areas simulate the motor usage of the interlocutor. To relate this to the interactive alignment theory, if alignment is as important as the authors suggest, then alignment on the phonetic level ought to be reached when both speaker and listener has aligned the corresponding neural areas, which in turn is argued to facilitate alignment on other levels. Another thing to note is that the action/perception model that Pickering and Garrod (2013) use also could translate well into neuroscientific findings. If the feed-forward percept is true, that would probably require fewer resources by not having every aspect in need of constant attention. Monitoring and higher-order tasks probably use prefrontal areas, while the perception and decoding of speech is likely to engage the previously presented language areas. Pickering and Garrod (2004, 2013) can be argued to have provided two applicable and relevant theories of speech comprehension, as their suggestions can be integrated with possible correlates handling each task.

Regarding speech and neural activation in time, the two primary EEG elements noted here are the N400 and the P600. As interpreted from the various studies, the P600 seems to be activated following anomalous bottom-up processing of syntax (Kung et al., 2014; Kolk et al., 2003), while the N400 seems to indicate a clash in top-down expectation and bottom-up processes of semantic information (Egorova et al., 2013). The initial processing takes place during the first couple of hundred milliseconds, after which at least lexical definitions seems to have begun to be integrated into a semantic context. According to Friederici's (2002) temporo-frontal theory syntax and semantics can be processed either in parallel, or syntax before semantics. Ultimately, though, the theory itself endorsed more of a parallel account, in contrast to e.g. the TRACE model (McClelland & Elman, 1986) which clearly states that the features are processed before the phonemes, which in
turn are constructing words. For the purposes of this discussion, however, the order in which syntax and semantics are processed are undetermined and left without a clear stance. With that said, though, the fact that the N400 precedes the P600, if indeed affected exclusively by semantics and syntax, respectively, could perhaps indicate parallel processing.

One thing that models on language ought to explain, whether implicit or explicit, is the very flexible and available working memory. To be able to follow a full conversation not only the on-line speech needs to be decoded, it also needs to be integrated into what has been stated earlier, and yet kept flexible enough to account for what will be said. If, as Pickering and Garrod (2013) hypothesize, and studies have indicated (e.g. Brunnelière & Soto–Faraco, 2015), language is predictive and very reliant on on-line predictions, that could indicate how priming takes place, perhaps to such an extent that the attention and other types of resource demand can be reduced. In the interactive alignment theory (Pickering & Garrod, 2004) alignment is suggested to be reached due to automatized priming mechanisms. If the same could be applied for language as a whole, not just in dialogue situations, such a proposal may explain why the potentially attention-heavy and monitoring-required activity of speech can happen with such fluency. Indirectly, the priming suggested in the TRACE model (McClelland & Elman, 1986) or the MUC model (Hagoort, 2005) could be relevant. The more a word is encountered, the more a word is primed. This corresponds well to the prediction element of speech discussed earlier, which stated that the more something is experienced, the greater the chance of predicting a future experience of a similar kind. Applied to language, it may be something like: the more speech that is heard, and the more speech that is created, the more primed words become, which means a certain string of words is more likely to follow, and makes a semantic context become more likely to include an utterance remarking something in that context. In short, Pickering and Garrod's (2013) suggested association and simulation practices for fostering language are perhaps very important and relevant for all types of
language usage, given that language is primed (e.g. McClelland & Elman, 1986) in the first place. Priming could take place as to decrease the resource demand by monitoring and integrating on-line speech, which then, in turn, paves the way for prediction making (Pickering & Garrod, 2013) as a viable strategy to engage in conversation, without constantly needing to attend or monitor every word.

If language use could be seen as some kind of extensive and constant priming, as mentioned above, it would probably tune expectations and predictions accordingly, which then could cause accent variations to speech, or even speech of a different language, to be unexpected and therefore harder to integrate, and by extension harder to comprehend. Although multilingualism is not dealt with here per se, Zhang et al. (2005) propose a neural commitment of neural areas to the native language, which would make processing of other languages less efficient. Zhang et al. (2005) suggest that acoustic cues could provide the necessary tuning to another language, and switching between expectations of different languages could therefore be a matter of priming. If priming indeed holds true for native language use, it could just as well apply for non-native language use.

A merging of the dual-route processing model (Hickok & Poeppel, 2000, 2004, 2007) and the computational models (McClelland & Elman, 1986; Pickering & Garrod, 2013) could potentially generate a novel perspective. Both standpoints carry relevant points which together could result greater understanding for both types of theories. The spatiotemporal details of empirical research combined with computational models could enhance the computational models with knowledge of neural activation clusters, activation hierarchy, or brain structure details. The models describing correlates could also gain, in a similar manner, from the computational models.

According to multiple studies (e.g. D'Ausilio et al., 2012; Smirnov et al., 2014), as well as models (e.g. Hagoort, 2005; Friederici, 2002), the left inferior frontal gyrus, or Broca's area, plays an integral role for merging multiple aspects of speech. With that in mind it seems reasonable to
assume a working memory is, at least in part, dependent on that structure given that the working memory needs access to all, or most, elements to be integrated in the left inferior frontal gyrus (LIFG). The role of prefrontal cortex in monitoring, attention, and higher-order tasks further indicates that frontal areas could be a good place, structurally, for a working memory dealing with language information. Caplan and Waters (1999) went as far as to hypothesize of a “separate language interpretation resource” (p.93) which would include all aspects of processed language. Kolk et al. (2003) dismisses the idea, however, as it is inconsistent with their test results. Instead, Kolk et al. (2003) suggests a working memory that is not specialized to a verbal aspect, but specialized for syntactic processing. This idea seems feasible; not dividing working memory into a verbal subdivision, and instead have the division aimed at syntax processing. If so, not only auditory stimuli would be enveloped in the suggestion. On the other hand, Baddeley's (2000) well established working memory model clearly suggests a phonological loop which has been shown to be sensitive to auditory, or phonological, stimuli. Rogalsky, Matchin, and Hickok (2008) noted the pars opercularis portion of the LIFG to be sensitive to syntactically complex sentences, and the pars triangularis part as potentially important, though in an undetermined manner. The precise fundamental basis for a language-related working memory will therefore have to be determined.

This attempt to summarize speech comprehension-related brain activation has made one thing clear: there are a multitude of aspects to take into account. From syntactical construction of sentences (Dapretto & Bookheimer, 1999), to usage of words (Vitello, Warren, Devlin, & Rodd, 2014; van Alpen & van Berkum, 2009), acoustic-only or audiovisual stimulus (Bernstein & Liebenthal, 2014), the relatively unexplored area of prosody (Meyer, Steinhauer, Alter, Friederici, & von Cramon, 2004), context influence on semantic processing (Smirnov et al., 2014), current mood (Egidi & Nusbaum, 2012), empathy levels (van den Brink et al., 2012), single-word processing or sentence comprehension (Mazoyer et al., 1993), or dialogue versus monologue (Pickering &
Garrod, 2004). In addition to that, the focal point here was only on speech comprehension (which even in itself can be divided further!). There are still, at least, the major areas of speech production, speech comprehension, reading, and writing that are in need of exploration, as well as thorough integration with the other fields, in order to properly account for most language aspects.

The field of combined neuroscience and linguistics has undoubtedly seen an explosive growth since the days of Broca and Wernicke. Relevant research is constantly being conducted, with substantial contributions made to the literature each year. The very capable scientists in the field have so far provided invaluable progress in narrowing down neural correlates for speech comprehension. With solidly founded theoretical bases regarding both how, where, and when different aspects of speech comprehension take place, the future seems bright for the field of neurolinguistics.
References


A brain with some of the most relevant areas for speech comprehension highlighted. Red shows the primary auditory cortex. Light blue displays the (posterior) superior temporal gyrus. Orange highlights the medial temporal gyrus. Teal shows the fusiform gyrus. Light green displays the temporal pole. Purple displays the temporoparietal junction, also called angular gyrus. Dark blue shows motor areas. Yellow highlights the inferior frontal gyrus, also called Broca's area. Dark green shows dorsomedial and ventromedial prefrontal areas.