One path or two: Could differential grammatical class processing reflect human language evolution?

"A mysterious sound, the sound of creation; in the beginning was the word.

Was it a sound or a movement?"

(Christine Riley, Armidale Waldorf School, 2010)

1. Introduction.

There are in excess of 7 000 known modern human languages (Ladefoged, 2005). Of these, some share sufficient similarities to suggest a common ancestor. For example, modern English and Icelandic both appear direct descendants of the Germanic language, which itself was the descendant of a more general Indo-European language (Williams, 1986). What, if anything, emerged prior to Indo-European (or the other general historical language groups such as Afro-Asiatic) remains largely a mystery. Although it is possible that these historical languages (proto-Indo-European and proto-Afro-Asiatic) may reflect the emergence of different human language types via completely different evolutionary pathways, this is not accepted by all with some, such as Ruhlen (1994) and Durham (1990), arguing for a single proto-human or *proto-sapiens* language. Although Ruhlen and others base their argument on analyses such as lexicostatistics and glottochronology, there is relatively minimal individual variation in neural anatomy and general language processing suggesting common neural substrates (Rushton & Rushton, 2003). Even sign languages such as AUSLAN and American Sign Language (ASL) show a striking number of similarities with spoken languages (Petitto, Zatorre, Gauna, Nikelski, Dostie, & Evans, 2000; Rönnberg, Söderfeldt, Risberg, 2000); the method of effecting communication may use different sensory modalities, but many of the underlying cognitive processes appear identical to spoken languages such as English. As a result, independent evolutionary pathways appear unlikely; instead, a single evolutionary path from *proto-sapiens* through to today's rich linguistic diversity, is considered by many (e.g. Hauser, 1996; Cheney & Seyfarth, 2005) to be the most probable route for human language evolution.

1.1 Language Universals.

A single evolutionary progression is partially supported by the existence of (at least) two distinct *language universals*; that is, language properties displayed by every known modern human language, including sign languages (Kauschke, Lee, & Pae, 2007). The first of these is *prosody*; that is, those suprasegmental aspects of language such as intonation or stress that help to convey or modify the meaning of the actual physical utterance or gesture (Arciuli, & Slowiaczek, 2007). Although the way prosody is used can vary widely between languages, it is believed to convey both affective and linguistic information (Skwerer, Schofield, Verbalis, Faja & Tager-Flusberg, 2007). In the case of the former, this is posited as providing the emotional cues (e.g. communicative intent) necessary for successful social interactions (Skwerer et al., 2007). Linguistic prosody is, alternately, thought to provide more structural information, such as cues for accessing the *lexicon*, or word meaning memory (Cutler, Dahan & van Donselaar, 1997; Arciuli & Slowiaczek, 2006).

The second generally accepted language universal is the use of the formal structure known as syntax which constrains what word type can be used where in an utterance (Fromkin, Rodman, Hyams, Collins & Amberber, 2005). For example, English uses the general syntactic form of subject-verb-object (S-V-O), "*Cats hunt rats*", whereas Japanese is structured subject-object-verb (S-O-V), "*Cats rats hunt*". Despite different languages displaying such different syntactic structures, none is syntax-free and all use a variation on the same basic pattern; that is, an ordering of words or phrases representing *things* (subjects/objects) with words or phrases representing how such things interact (verbs). For the purposes of this thesis, subjects and objects will be collectively referred to as *nouns* while actions/interactions will be referred to as *verbs*. .

In a colloquial sense, the two grammatical word classes noun and verb are often described respectively as *naming words* and *doing words*. From a more functional perspective, nouns can be modified to show number and verbs to show tense (Fromkin et al., 2005). As with syntax, different languages can use different morphological methods to effect modifications, but the underlying function of the modification does not change (Fromkin et al., 2005). Although some posit that a noun/verb distinction is universal (Kauschke et al., 2007), others such as Gil (2000), dispute this. It should be noted, however, that there is considerable evidence to support differential neural processing between nouns and verbs. For example, lesion studies led Damasio and Tranel (1993) to conclude a double dissociation in noun and verb processing,

suggesting that the system underlying noun retrieval lies in the left hemisphere, particularly the left anterior and middle temporal lobes, while the system underlying verb retrieval is located primarily in the left premotor cortex. Similarly, Pulvermüller, Lutzenberger and Preissl (1999), used an event-related potential (ERP) paradigm to demonstrate a physiological noun/verb double dissociation within 200 milliseconds (ms) of stimulus presentation. The authors further concluded that the neural generators subserving noun perception are most likely located in the primary and/or higher order visual cortices within the occipital lobe while verb perception involves neural generators located in or near the primary motor cortex, the pre-motor cortex and the adjacent pre-frontal cortices.

Although the concept of differential grammatical class processing is disputed by some (e.g. Yokoyama et al., 2006; Siri et al., 2008; Tyler, Bright, Fletcher & Stamatakis, 2004), much of the previous literature is marred by methodological inconsistencies. For example, Pulvermüller et al. (1999) used written rather than spoken word presentation thus may have reported the workings of literacy rather than language processing, a related but fundamentally different skill (Olson, 1998). Similarly, Tyler et al. (2004) used functional magnetic resonance imaging (fMRI) data to refute such differential processing despite fMRI lacking the temporal resolution needed to capture millisecond changes in functional brain network connectivity (Aine, 1994). As a result, the existence (or otherwise) of differential noun/verb processing remains contentious.

1.2 Current Models of Language/Speech Evolution

Unlike other human developments such as walking upright and tool use, language by its very nature has left no archaeological record (Olson, 1998). Although some clues, such as vocal tract evolution, can be found in fossilised remains (Falk, 1990; Falk & Gage, 1997), the cognitive bases of language emergence and its subsequent evolution remain largely conjectural. It is thought by many, however, that some type of proto-language emerged approximately 200 000 years ago, around the time that *Homo sapiens* emerged as a species in its own right (e.g. Mithen, 2005). Although the term proto-language (i.e. *proto-sapiens*) has differing definitions in the literature (e.g. Gell-Mann & Ruhlen, 2011), it will be considered by this thesis to be different to simple communication (e.g. bird song) based on the three criteria proposed by Brown (1973). Specifically, to be considered *language*, a communication system must display the three criteria, of productivity, semanticity, and displaceability. The first of these criteria, also known as creativity, allows any user of a language system to recombine individual elements to create

original meaningful utterances; that is, true language is not restricted to a limited repertoire of memorised patterns, such as the waggle dances of bees (Su, Cai, Si, Zhang, Tautz & Chen, 2008). Semanticity, alternately, means that language elements must represent abstract ideas or symbols; thus, in certain varieties of English, the phrase "homicidal axe-wielding maniac" represents what PhD students threaten to become when attempting to re-write a thesis. Finally, displaceability allows an utterance to retain its original meaning without grounding that meaning in the immediate environment. For example, the meaning of the sentence "My dog smells bad" would be readily apparent to native speakers of English regardless of whether said dog, or odour, was physically present or not. The exact biological foundations for language are unknown, but, if language *evolved* rather than *emerged* as a complete, complex system, then this had to have occurred as a cumulative sequence of steps.

1.2.1. General Evolutionary Context.

As proposed by various theorists (e.g. Tan, 2005; Arbib, 2005; Coppens, 1994), the first precursor step for language evolution was moving to a bipedal locomotion system. As proposed by Coppens (1994), such bipedalism this was driven by the geological upheaval that created the Great Rift Valley which, in turn, divided parts of east and west Africa. Of the primate population of the time, those on the eastern side of the valley subsequently experienced changes to prevailing wind and rainfall patterns which gradually reduced the availability of food, water and shelter. Vast rainforests transformed into open grassland which required both speed and a wide visual angle to safely traverse, thus survival of the fittest meant fastest and tallest. Compared to walking on all fours, bipedalism has obvious height advantages, so a change in locomotion allowed for the increased visual field needed to identify danger while escape was still possible. Bipedalism also allowed for faster ground speeds, so that escape was not only possible but achievable.

As with any complex system, however, a change in one part of the system often results in changes to other parts of the system. In the case of humans' earliest ancestors, the changes in musculature needed to support bipedal locomotion also resulted in vascular system changes which, like an automotive radiator, allowed more effective cooling of the early hominid brain (Falk, 1990). This, in turn, allowed for the increased encephalisation that is believed necessary for many of modern humans' cognitive abilities (Jerison, 1973/2012). The environmental changes resulting from the formation of the Great Rift Valley also changed the type of food available to early primates (Coppens, 1994). By necessity, the eastern primates were forced to

adapt to a diet of grasses and roots, supplemented by protein scavenged from carnivore left-overs (Blumebschine & Cavallo, 1992). To adjust to these dietary changes, the larynx descended, elongating the vocal tract, serendipitously allowing for the production of a much broader range of vocalisations, including consonant sounds (MacNeilage, 1998). This, combined with increased encephalisation, meant that the early hominid brain was now equipped with the neural substrates necessary for a proto-language to emerge. This emergence need not, however, have been the giant cognitive leap forward that some theories, such as Universal Grammar (Chomsky & Halle, 1968) suggest; rather, a more basic physiological change such as the development of more plastic mirror neuron networks may have been the tipping point from communication to language. To explain: if a relatively minor genetic mutation in some type of protein production resulted in an improved ability to form some types of functional neural networks and these networks were better able to form/re-form in real-time, then this improved neural network plasticity may have provided the foundations for recursion, one of the fundamental pre-requisites for language (Brown, 1973). Although the FOX-P2 gene has previously been implicated in the development of language (Enard et al., 2002), it is not species specific; that is, mutations of the gene have been directly linked to altered ultrasonic vocalisations in mice (Shu et al., 2005). The FOX-P2 gene has also been connected to vocal learning in zebra finches (Haesler et al., 2004) while Teramitsu, Kudo, London, Geschwind and White (2004) report similar patterns of expression of the gene in both humans and songbirds. As a result, despite mutations of this gene being associated with human language impairments such as verbal dyspraxia (Feuk et al., 2006), it appears to be more indicative of communication than language as such.

1.2.2. The Proto-Speech Model of Language Evolution.

Of the various models proposed to represent human language evolution, the proto-speech model (Steklis & Raleigh, 1979; although for review, see Hewes, 1977) may appear the most obvious given modern humans' reliance on vocal speech; that is, normally-developed adult modern humans speak rather than gesture therefore modern language must be the direct evolutionary descendant of sound based systems such as primate vocalisations. In much the same way that modern human infants acquire language (see Section 1.3.1), early hominid communications evolved from reactive to intentional vocalisations. These vocalisations eventually became conventionalised, such that meaning no longer required contextual grounding (Fromkin et al., 2005). As the range of speech sounds expanded, so too did the communicative power of the sounds until proto-speech diverged into the various regional varieties (e.g. Germanic, Asiatic) that, in turn, evolved into the multitude of languages known today.

Although possible, the proto-speech model fails to adequately account for the advanced language capabilities displayed by congenitally deaf modern humans. As demonstrated by a vast pool of literature (e.g. Johnston, 2001; Johnston & Schembri, 2007; Stokoe, 2005), AUSLAN, ASL and the various other sign languages satisfy every formal requirement needed to be considered true languages (Brown, 1973). Furthermore, the linguistic complexity shown by such languages is at least equivalent to any spoken language (Bernstein Ratner, 2005). Native signers of these languages also show near identical activation of language general areas traditionally associated with speech (Hickok, Bellugi & Klima, 1996) and a number of lesion studies have demonstrated that impairments to specific areas (such as Wernicke's area) result in equivalent aphasic symptoms; that is, a signer's ability to sign is impaired in an equivalent manner to a speaker's ability to speak (Hickok, Bellugi & Klima, 1998). Congenitally deaf infants also progress through a finger babbling stage at around the same time that hearing infants babble vocally irrespective of exposure to sign language (y Moores, 1980). Even prosody, conveyed in spoken language by variations in the sound, is conveyed via variations in the gesture in signed languages (Dachkovsky & Sandler, 2009). Taken together, this evidence clearly refutes a need for sound in advanced language acquisition, development and use, instead suggesting that the neural substrates of language (as opposed to speech) perception are not *exclusively* sound based; if they were, deaf infants would not be able to acquire advanced language ability. This, in turn, suggests that although speech may be the auditory expression of language, it is not the underlying cognitive mechanism that subserves human language. As a result, the ability to exercise sufficient vocal control to speak may have been necessary for the evolution of speech, but it was not sufficient by itself for the evolution of language.

1.2.3. The Proto-Sign Model of Language Evolution.

An alternate model for language evolution is that of proto-sign. As suggested by numerous authors such as Zlatev (2008), Corballis (2003), Wacewicz & Żywiczyński (2015), Arbib (2002), language first emerged in hominids via gesture. Under this model, mimetic rather than vocal gestures became conventionalised. For example, rather than panting or hooting, "Stop! There's a snake down there!", a proto-sign user would simply sign *stop* (possibly in much the same way as modern humans do) then point down, then wiggle their hand/arm to indicate a snake's movement.
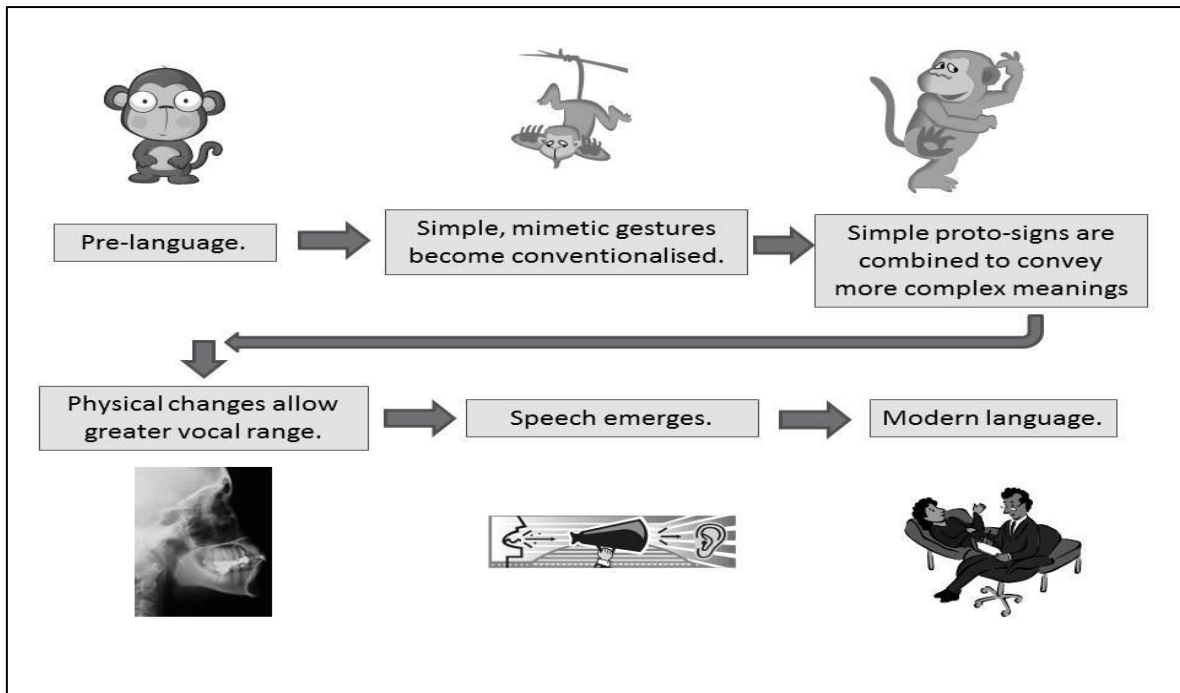
Figure 1.1: The proto-sign model of language evolution.

Silent communication would, in numerous situations, be highly advantageous. For example, if stalking prey, a group of early hominids would likely be more successful if a coordinated yet silent approach was taken so as to not spook the target. Conversely, if a large predator was observed nearby, escape would be more likely if the group was alerted to the danger without the predator being made aware of the group; that is, gesturing "Run! Really big lion!" would be more protective that panting/hooting/grunting/screaming the same message. Vocalisations also require specific shared knowledge between speaker and audience of the abstract concepts associated with those sounds; that is, sounds are arbitrarily strung together to form bigger segments (such as words) based on each specific language/communication system and without knowledge of that specific system, the sounds are merely sounds . Thus, modern English and Icelandic may both be vocal languages but they are by and large mutually unintelligible. Mimetic gesture, alternately, requires only a shared desire to communicate rather than a shared knowledge of a specific language's phonology, morphology and lexicon. For example, it is possible for an English-only speaking Brisbane girl to have a memorable night on the town with a French-only speaking Parisian girl by simply using gesture. Allowing then that mimetic gesture requires less specific shared knowledge, the idea of intentional rather than reactive communication may have been more easily grasped by other hominids. The general concept of language would then spread quickly and start diverging based on regional or familial forces (Fromkin et al., 2005).

Like the proto-speech model, however, the proto-sign model cannot fully explain language evolution. To again use the analogy of human infant language acquisition, if language development was exclusively based on gesture, then blind infants would not be able to acquire advanced language ability. As reported by numerous authors (e.g. Landau, Gleitman & Landau, 2009; Barrett, Lindquist & Gendron, 2007; Gervain & Mehler, 2010), this is simply not the case; blind infants acquire spoken language in a near identical manner to normally developed infants. This, in turn, suggests that although gesture may be a manual expression of language, like speech, it is not the underlying cognitive mechanism that subserves human language. Again, ability to gesture may have been necessary for the evolution of language, but it was not sufficient by itself for the evolution of language.

1.2.4. Arbib's Model of Co-Evolution.

As neither the proto-speech nor proto-sign model adequately explains language evolution, a combined model was subsequently proposed. As the name suggests, this model proposes a co-evolution of language via both proto-speech and proto-sign (Arbib, 2002, 2005). From this perspective, language did not originate in either vocalisations or gestures; rather, an ability to parse action sequences into ever-increasingly smaller constituent parts pre-empted both. The concurrent emergence of the ability to imitate, when combined with a sufficiently fine-grained action parsing ability would thus lay the basic framework for the conventionalisation of purposeful communication, using whatever physical means were available. If, as suggested by Mithen (2005) and others (e.g. Corballis, 2003), language first emerged in early hominids approximately 200 000 years ago, then both gross (gesture) and fine (speech) motor skill development would have been sufficiently advanced to allow both proto-speech and proto-sign to emerge at around the same time. A merging of modalities, or the mutual innervation of vocal and gestural control areas, would then allow the emergence of *proto-sapiens* (Ruhlen, 1994). Such a merging may have been driven by physical environmental pressures (such as needing to hunt, and thus communicate, during both the day and the night) and/or the evolution of more advanced cognitive processes (such as developing time and space concepts) which required the ability to convey more complicated meanings and messages.

1.2.5. Mathematical Modelling of Language Evolution.

The advent of super-computers and the development of advanced signal processing techniques led to the proposal that "…neurobiologically realistic computational models…" may provide a more usable framework for combining disparate data sets into a "…conceptually

coherent…" whole (Horwitz & Tagamets, 1999, p. 138). Such modelling also facilitated the development of synthetic brain imaging (SBI) techniques specifically designed to test computational models. By applying such techniques, computer modelling has successfully simulated a PET investigation of a saccade generation task (Arbib, Bischoff, Fagg & Grafton, 1995) as well as both simulated PET and fMRI examinations of a delayed match-to-sample shape task (Horwitz & Tagamets, 1999).

Cangelosi and Parisi (2004) also applied SBI to an evolutionary plausible computational language model to suggest that nouns activate posterior brain areas related to sensory and associative processing. Verbs, alternatively, activated anterior motor areas, especially those relating to the integration of sensory and proprioceptive information. Cangelosi and Parisi (2004, p. 2) further suggested that by the "…grounding of symbols in sensory-motor experience…" as well as addressing "…neural plausibility…" and "…network architecture and functioning…" that their model incorporated both evolutionary and connectionist techniques that could reliably simulate some cognitive/linguistic interactions. Application of this model (Cangelosi & Riga, 2006) subsequently suggested that nouns evolved from verbs; that is, proto-language initially emerges in verb (or verb phrase) form only with nouns only appearing after this verb-centric language emergence.

1.3. Using Modern Language Development as an Evolutionary Template.

Although ontogeny does not necessarily recapitulate phylogeny (Gould, 1977), until the development of modern brain imaging techniques, theories of language evolution were primarily based on observations of language acquisition in human infants. Although all except the most radical behaviourist theories posited some innate component, the predominant guiding theory for much of the latter twentieth century was *universal grammar* (Chomsky, 1969, 1986). As argued by Chomsky (1969, 1986), there is insufficient environmental input for language acquisition to be based on stimulus-response learning; that is, a *poverty of stimulus* (PoS) precludes associative learning such as classical or instrumental conditioning. In more simplistic terms, the basic PoS argument suggests that exposure to language does not provide sufficient clues, by itself, for an infant to acquire language; that is, mere exposure to the sounds (or signs) and vocabulary of a language does not provide the infant with the information needed to map object and action concepts onto meaningful sound or sign strings. Given, however, that infants *acquire* rather than are formally *taught* speech/sign, this thus infers that some language faculties must be innate, therefore human infants must be born with dedicated functional language faculties pre-encoded

in brain anatomy. As broadly explained by universal grammar, this innate neural module is primarily related to syntax with environmental exposure driving the acquisition of the syntactic rules specific to the infant's native language environment; thus, an infant born into an English speaking community would learn the syntactic rules subject-verb-object (SVO), whereas an infant born into a Japanese speaking family would learn subject-object-verb (SOV). Universal grammar, and thus the premise of an innate language model, is also used to explain a variety of other language phenomena including the universal tendency for young children to overgeneralise morphological rules relating to tense and number (e.g. pluralising *man* to *mans*; Pinker, 1994).

Evidence for universal grammar was also reported by Senghas and Coppola (2001). By studying deaf Nicaraguan children, Senghas and Coppola (2001) detailed the emergence and refinement of a unique yet systematic sign language in <20 years. Following analysis, age of exposure displayed a significant effect with children under ten driving the development and formalization of the language. The authors concluded that this was evidence of language innateness due to the "…markedly impoverished…" stimulus input from the language environment (Senghas & Coppola, 2001, p. 323). Similarly, Pinker (2000) describes the case of "Isabelle", a neglected child who was completely isolated from language exposure until her rescue at age 6. Despite this lack of exposure, "Isabelle" was fully language proficient within eighteen months of her release, capable of grammatical constructions and language usage far more complex than could be explained by the remedial instruction she received. Taken together, this evidence clearly supports an underlying innate component to human language, but precisely what that is and how it evolved remains conjectural.

1.3.1. Normal Language Development in a Modern Infant.

As described by the U.S. Department of Health and Human Services (2010), normal "language" development is marked by a number of *milestones*, each representing an average rate of development. See Figure 1.2.

Age	Milestones
0 – 3 Months	Reacts to loud sounds Calms down or smiles when spoken to Recognizes your voice and calms down if crying When feeding, starts or stops sucking in response to sound Coos and makes pleasure sounds Has a special way of crying for different needs Smiles when he or she sees you
4 – 6 Months	Follows sounds with his or her eyes Responds to changes in the tone of your voice Notices toys that make sounds Pays attention to music Babbles in a speech-like way and uses many different sounds, including p, b, and m Laughs Babbles when excited or unhappy Makes gurgling sounds when alone or playing with you
7 – 12 Months	Enjoys playing peek-a-boo and pat-a-cake Turns and looks in the direction of sounds Listens when spoken to Understands words for common items Responds to requests Babbles using long and short groups of sounds ("tata, upup, bibibi") Babbles to get and keep attention Communicates using gestures such as waving or holding up arms Imitates different speech sounds Has one or two words by first birthday

Figure 1.2: Taken from U.S. Department of Health and Human Services (2010) developmental guide.

Of particular note, these milestones are primarily sound based; that is, language development is gauged wholly and solely on speech development. Although such milestones can facilitate the early diagnosis of a hearing impairment, they do not necessarily reflect true language development; profoundly deaf babies will also smile when they see you, will often pay attention to music (due to the vibration not actual sound), and will communicate via gesture. If, however, more generalised milestones such as motor development are considered (see Figure 1.3), then it becomes apparent that infant speech development is largely constrained by the motor system; that is, until fine motor skills have become sufficiently refined to master the intricate coordination of the vocal apparatus (i.e., lips, tongue, diaphragm), speech is physically impossible. This does not, however, naturally confer identical constraints on language development; instead, language development needs to be gauged within the context of actual motoric ability. For example, a congenitally deaf infant who finger babbles (y Moores, 1980) at four months may have more advanced language skills than the infant who babbles vocally at six months. Similarly, the ability to follow eye-gaze at two months might be a more reliable indicator of language development than the ability to "Follow (s) sounds with his or her eyes" (Young, Merin, Rogers & Ozonoff, 2009) at four to six months.

Age	Milestones
0 – 2 Months	Begins to smile at people Tries to look at parent Turns head toward sounds Begins to follow things with eyes and recognize people at a distance Can hold head up and begins to push up when lying on tummy Makes smoother movements with arms and legs
2 – 4 Months	Smiles spontaneously Copies some movements and facial expressions Reaches for toy with one hand Uses hands and eyes together, such as seeing a toy and reaching for it Holds head steady, unsupported Pushes down on legs when feet are on a hard surface May be able to roll over from tummy to back Can hold a toy and shake it and swing at dangling toys Brings hands to mouth When lying on stomach, pushes up to elbows
4 – 6 Months	Likes to play with others, especially parents Strings vowels together when babbling Responds to own name Begins to say consonant sounds (jabbering with "m," "b") Brings things to mouth Begins to pass things from one hand to the other Rolls over in both directions (front to back, back to front) Begins to sit without support When standing, supports weight on legs and might bounce Rocks back and forth, sometimes crawling backward before moving forward
6 – 9 Months	Makes a lot of different sounds like "mamamama" and "babababababa" Copies sounds and gestures of others Uses fingers to point at things Plays peek-a-boo Moves things smoothly from one hand to the other Develops pincer grip (thumb and index finger) Stands, holding on Can get into sitting position Sits without support Pulls to stand Crawls
9 – 12 Months	Repeats sounds or actions to get attention Puts out arm or leg to help with dressing Responds to simple spoken requests Uses simple gestures, like shaking head "no" or waving "bye-bye" Explores things in different ways, like shaking, banging, throwing Copies gestures Starts to use things correctly; for example, drinks from a cup, brushes hair Pokes with index (pointer) finger Follows simple directions like "pick up the toy" Gets to a sitting position without help Pulls up to stand, walks holding on to furniture ("cruising") May take a few steps without holding on May stand alone

Figure 1.3: More generalised language milestones incorporating motor development.

1.3.2. Developmental Similarities with Existing Models of Language Evolution

If language development milestones are reconceptualised to include motoric constraints, then normal language development could be represented as per Figure 1.4.
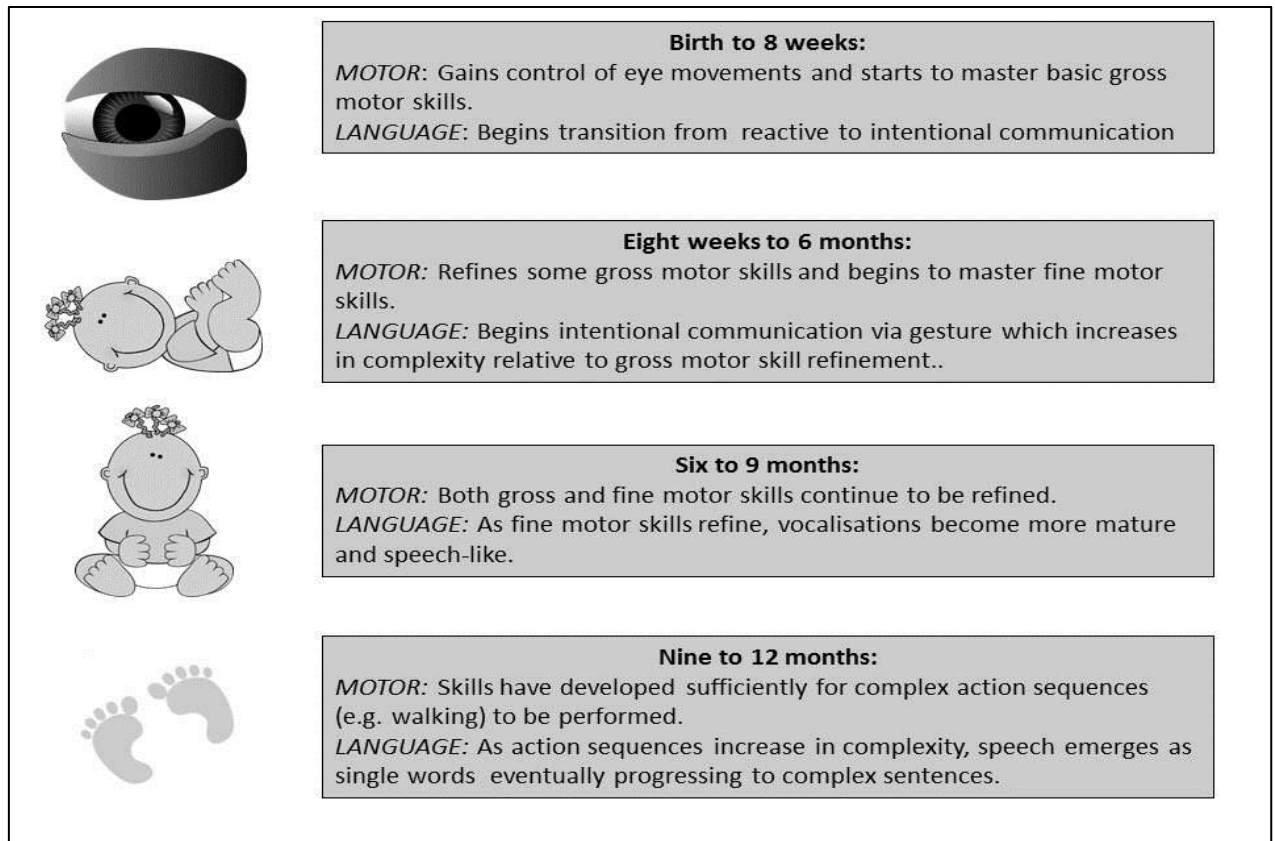
Birth to 8 weeks:
MOTOR: Gains control of eye movements and starts to master basic gross motor skills.
LANGUAGE: Begins transition from reactive to intentional communication

Eight weeks to 6 months:
MOTOR: Refines some gross motor skills and begins to master fine motor skills.
LANGUAGE: Begins intentional communication via gesture which increases in complexity relative to gross motor skill refinement..

Six to 9 months:
MOTOR: Both gross and fine motor skills continue to be refined.
LANGUAGE: As fine motor skills refine, vocalisations become more mature and speech-like.

Nine to 12 months:
MOTOR: Skills have developed sufficiently for complex action sequences (e.g. walking) to be performed.
LANGUAGE: As action sequences increase in complexity, speech emerges as single words eventually progressing to complex sentences.

Figure 1.4: Combined speech and motor developmental guide for normal language development.

The similarity to the co-evolution model (see Section 1.2.4) is obvious; specifically, that motor development, both gross and fine, constrains the ability to produce informative, intentional language and that as motor development progresses (evolves) and more refined control is achieved, intentional language can be expressed in an increasingly more refined and controlled manner. It has further been argued (e.g. Capirci & Volterra, 2008) that such motor development is the necessary neural substrate upon which spoken language develops; that is, gesture acts as a scaffold for speech. Such motoric constraints also, potentially, suggest different developmental trajectories for different word classes, specifically nouns and verbs, which, in turn, potentially suggest different evolutionary paths. For example, intentional communication by human infants begins with movements such as following eye-gaze and smiling at their primary caregiver (Young, Merin, Rogers & Ozonoff, 2009). Although they may also use vocalisations (crying) to

communicate states (such as tired or hungry) this tends to be reactive more than intentional, and directly related to need satisfaction. As the infant matures, movement continues to dominate *intentional* communication. For example, the raising of the arms to indicate a desire to be picked up occurs long before the child can vocally express this desire in an intentional manner. This suggests then that intentional language first emerges via purposeful actions rather than more abstract representations of objects or things. Although this may be purely due to the lack of fine motor skill development, it could also represent different underlying cognitive development trajectories. For example, action infers movement which occurs in the *time* domain while objects/things occur in the *space* domain, thus the development of time and space concepts and perception may constrain the development of language. Allowing that verbs are functionally representative of movement/time (action) while nouns are functionally representative of objects/things (subject/object), it may then be that the two general grammatical classes of words are physical manifestations (speech or sign) of more basic perceptual processes relating to time and space. It is the merging of these two processes that generates language which is then primarily expressed via nouns and verbs, with the process underscoring verb production/comprehension developing slightly faster than that underscoring noun production/comprehension. Such a developmental trajectory in a modern human infant may or may not recapitulate language's evolutionary trajectory, however, it does not definitively exclude it.

1.4. Using Contemporary Primate Communications as an Evolutionary Comparison.

Given the lack of hard evidence available to support the empirical investigation of human language evolution, numerous researchers (e.g. Newport, Hauser, Spaepen & Aslin, 2004; Patterson & Cohn, 1990; Tyack, 2000; Van Cantfort & Rimpau, 1982) have turned to studying animal communications; in effect, separating *us* from *them*. In the case of our nearest relative (the common chimpanzee), the basic differences are not great with humans and chimps sharing 98.77% of the same genes (Mikkelsen et al., 2005). From a social perspective the difference between human and ape is also relatively small, with bonobos, chimpanzees and gorillas displaying behaviours ranging from advanced tool use to homicide, behaviours all once thought purely the domain of humans (van Lawick-Goodall, 1968; Whiten 1999; Matsuzawa, 1985). In experimental situations, monkeys and other apes have also displayed a vast range of cognitive abilities ranging from the intentional use of numbers to voice–face matching-to-sample tasks (Matsuzawa, 2006). Allowing for these similarities, primate communications systems appear to provide near ideal human language comparisons.

1.4.1. Physiological Differences – Vowels versus Consonant Production.

Of the research into ape vocal communication, one striking difference between *them* and *us* is the use of consonants (MacNeilage, 1998). Although all primate types use vocal communication systems of various complexities, none except humans display consonants as part of their phoneme inventory as the monkey vocal tract is physically incapable of intentionally producing the vast majority of consonant sounds (MacNeilage, 1998). Instead, monkeys must use vowel sounds, or modifications thereof, as the basis of vocal communication (Bonatti, Peña, Nespor & Mehler, 2005). Vowels, as defined by Ladefoged (2005, p.26) are "any sound…that is produced without any kind of obstruction of the outgoing breath". The number of possible vowel sounds capable of being produced by humans is considered finite, currently estimated at approximately 200, and although vowel sounds are relatively easy to produce in isolation, they are challenging to produce consecutively in connected speech (Ladefoged, 2005). As a result, any communication system based purely on vowel sounds would be physically constrained. Consonants, on the other hand are defined as "Sounds that have some obstruction to the breath stream, such as the bringing together of the lips" (Ladefoged, 2005, p. 26). Although English uses only 24 consonantal sounds (Crystal, 1995), it is thought that in excess of 600 exist across all language types (Ladefoged, 2005). Thus the evolutionary lengthening of the human vocal tract to allow the production of consonant sounds greatly expanded the capacity of human vocalisations to convey differences in meaning.

This capacity is further expanded through the ability to combine consonant and vowel sounds to produce syllables (Ladefoged, 2005). Unlike pure vowel combinations, the consonant/vowel combinations found in syllables are relatively easy and fast to articulate (Ladefoged, 2005). Although still restricted in number, the ease and speed with which syllables can be produced allows for the recombination of syllables to form larger meaning units (Ladefoged, 2005). This building or recombination of syllables into words then forms the basis of combining words into phrases and phrases into sentences, creating, in effect, a communication system without limit.

The ability to use consonants does, however, do more than just allow the combination of consonants with vowels to form syllables; it also allows vowels to multi-task (Bonatti et al., 2005). With a vowel-only communication system constrained by a limited number of possible vocalisations, concepts able to be communicated by the system would also be constrained; thus, while possible that monkey communications could have developed ways to express concepts

such as *parturiphobia*, *pseudoephedrine,* or *paedomorphism*, this would be unlikely in a physically limited communication system; instead, meaningful sound strings would be more beneficial being associated with concepts of survival. Perhaps unsurprisingly, this is what has been found in studies of monkey communications to date, with vocalisations revolving around satisfying the basic needs of food, water, safety and shelter (Macedonia & Evans, 1993; Clark & Wrangham, 1993; Crockford & Boesch, 2005). In other words, monkey vocalisations convey affective states such as hunger, fear and separation, thus vowels function as conveyors of affect.

Human language, on the other hand, allows vowels to perform multiple functions. While vowels convey prosodic information in human language, and as such are still carriers of affective information, there is evidence that vowels also convey grammatical class information. For example, Bonatti et al. (2005, p.451) suggested that the human brain uses transitional probabilities between "adjacent and non-adjacent syllables to segment "words" from a continuous stream of artificial speech". Based on the results of two separate experiments, the authors further concluded that consonants are used to help identify words while vowels convey grammatical class information. This is in keeping with Ladefoged (2005) who suggested that consonants are start and stop points for vowels within a syllable; that is, consonants modify the action and meaning of vowel sounds. It is also in keeping with Caramazza, Chialant, Capasso and Micell (2000, p. 428) who used case studies to demonstrate separable consonant/vowel neural processing; that is, that "consonants and vowels are processed by distinct neural mechanisms, thereby providing evidence for their independent status in language production".

1.4.2. Animal Gesture as a Means of Communication.

As previously mentioned, language does not need to be sound based. Although few studies have examined primate gestures, those that have report such communication as being closer to human language than vocalisations. For example, Tomasello and Carpenter (2007) suggest that great apes have the capacity to demonstrate flexibility within their own communication system in that specific individual gestures may be used in different contexts while similar contexts can result in a variety of different gestures. Similarly, wild Ugandan chimpanzees adapt their use of native communication gestures depending upon the state of their intended audience; that is, if their intended target is attending to them, the gesturer uses smaller, distal gestures whereas a non-attentive audience evokes larger contact signs (Hobaiter & Byrne, 2011). These chimpanzees also display obvious turn taking in conversational settings, such as waiting for a response to a gesture and/or adapting their gesture if the response was inappropriate

(Hobaiter & Byrne, 2011). Although not necessarily a reflection of true language, the gestural communication systems of primates appear more advanced than their vocal communication systems.

1.5. The Basic Unit of Modern Human Speech Perception.

Some argue that the phoneme is the basic unit of speech perception (e.g. Humphries, Sabri, Lewis, & Liebenthal, 2014), others argue that the syllable (or sometimes mora) is the basic unit (Ladefoged, 2005; MacNeilage, 1998; Nakayama & Saito, 2014), while still others opt for sub-syllabic segments (that is, onset, nucleus, coda) forming the basis for speech perception. Of these, the phoneme appears, intuitively, to represent the basic functional unit.

1.5.1 The Phoneme.

As defined by Reber (1995, p.566), a phoneme is a"…minimal unit of speech in a given language that 'makes a difference' to the fluent speaker of that language." For example, changing the English phoneme /k/ to /r/ can turn a "cat" into a "rat" (Ladefoged, 2005). It may be, however, that phonemes are not directly related to speech perception; instead, they may reflect *literacy* processing, or the ability to read and write. This is neither a pedantic nor minor distinction to make; humans do not listen and speak or sign in the same way that they read and write. This is easily demonstrated by physically examining simple voice recordings. By looking at the wave forms of spoken language it is possible, sometimes, to identify syllables and words, but impossible to accurately pinpoint the different sound units that equate to individual phonemes (Ladefoged, 2005). Even spectrogram analysis is not infallible in discerning individual phonemes in words, nor differences between these phonemes; it relies on probabilities rather than actualities (Ladefoged, 2005).
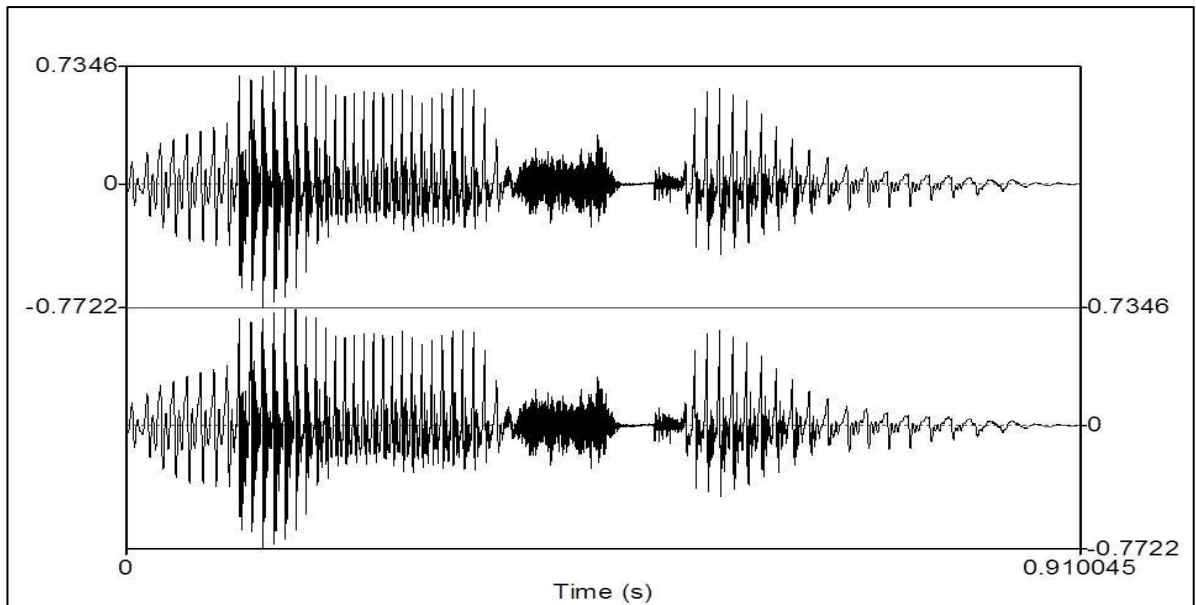
Figure 1.5: The stereo sound wave representing the word Maelstrom – upper wave form = left channel, lower wave form = right channel.

This has led some, such as Olsen (1998), to propose that literacy acquisition permanently alters some aspects of language cognition, a proposal that has a growing body of supportive evidence. For example, Treiman, Bowey and Bourassa (2002) demonstrated that as literacy levels increase, so too does the tendency to assign syllabic breaks according to formal spelling rules rather than naturalistic phonological factors. Similarly, Byrne (1998) suggests that many of the problems children encounter when acquiring literacy stem directly from a basic, fundamental difference between language and literacy; that is, literacy is based on artificial alphabetic units corresponding to phonemes, whereas language is based on larger units, possibly corresponding to syllables or syllabic segments such as onsets or rimes. When considered collectively, data such as these, suggest that phonemes are artificial constructs relating to literacy usage rather than natural constructs related to language perception. Thus, by imposing these artificial constructs on small humans while their brains are still plastic enough to be modified by environmental factors, literacy instruction forces the development of the belief that language is constructed from phonemes, when it is arguably not.

1.5.2. The Syllable and Sub-syllabic Segments.

A linguistic unit conceptualised as (generally) being smaller than a word but larger than a phoneme, the syllable is vexingly difficult to accurately define. From a classical perspective, the syllable is viewed purely as a speech property representing the intonation curves of language (Chomsky & Halle, 1968). Alternatively, some contemporary researchers have suggested that

the syllable is an active language process, with Levelt, Roelofs and Meyer (1999) positing a mental syllabary that acts in a similar manner to the lexicon, as well as proposing that syllabification equates to phonological syntax. More generally, however, the syllable tends to be defined structurally, comprising an onset and rime, with the rime composed of a nucleus and coda.
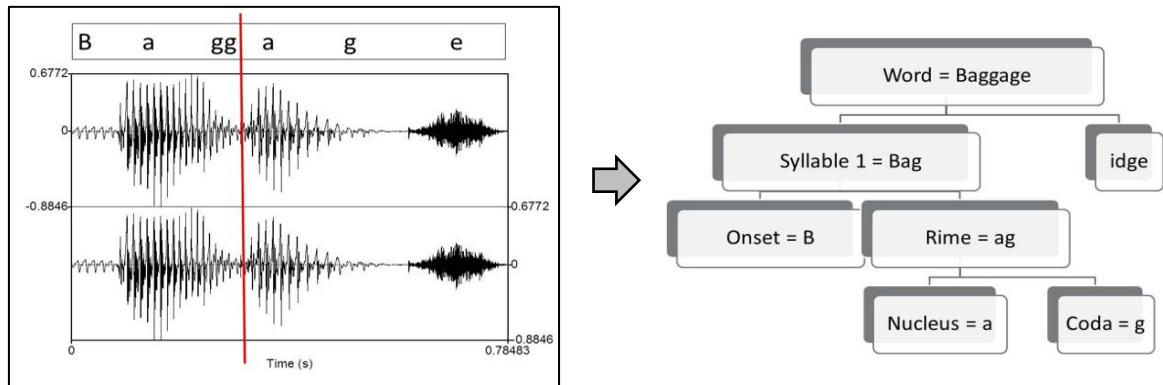

Figure 1.6: The word baggage - wave (sound) file (left) and its syllabic and sub-syllabic constituents (right).

The nucleus, the only obligatory constituent, is always a vowel (V). The remaining constituents are consonants (C) or consonant clusters, with the phonotactic rules of a given language governing which sounds can be sequenced within a syllable (Fromkin et al., 2005). Some languages, such as Japanese, allow only simple CV syllable structures; others, such as English, permit more complicated syllable constructions, such as CCCVCCCC (Fromkin et al., 2005). Although some languages are proposed as allowing syllabic constructions that are nucleus-free (i.e., lacking a vowel) such constructions are rare (Ladefoged, 2005). Furthermore, although typical consonants and vowels (e.g. /k/ as in *cat* and /ɛ/ as in *bed*), conform to the definitions described in Section 1.4.1, some approximant and lateral sounds (e.g. /j/ as in *yacht* and /l/ as in *lemon*) do not, and sometimes appear to behave more as consonantal vowels (Ladefoged, 2005). It is also possible, by drawing from classical generative syntactic analyses (e.g. Haegeman, 1997; Radford, 1986), to suggest that a lack of a nucleus is not a lack per se; rather, it is a *null* vowel. See Figure 1.7.
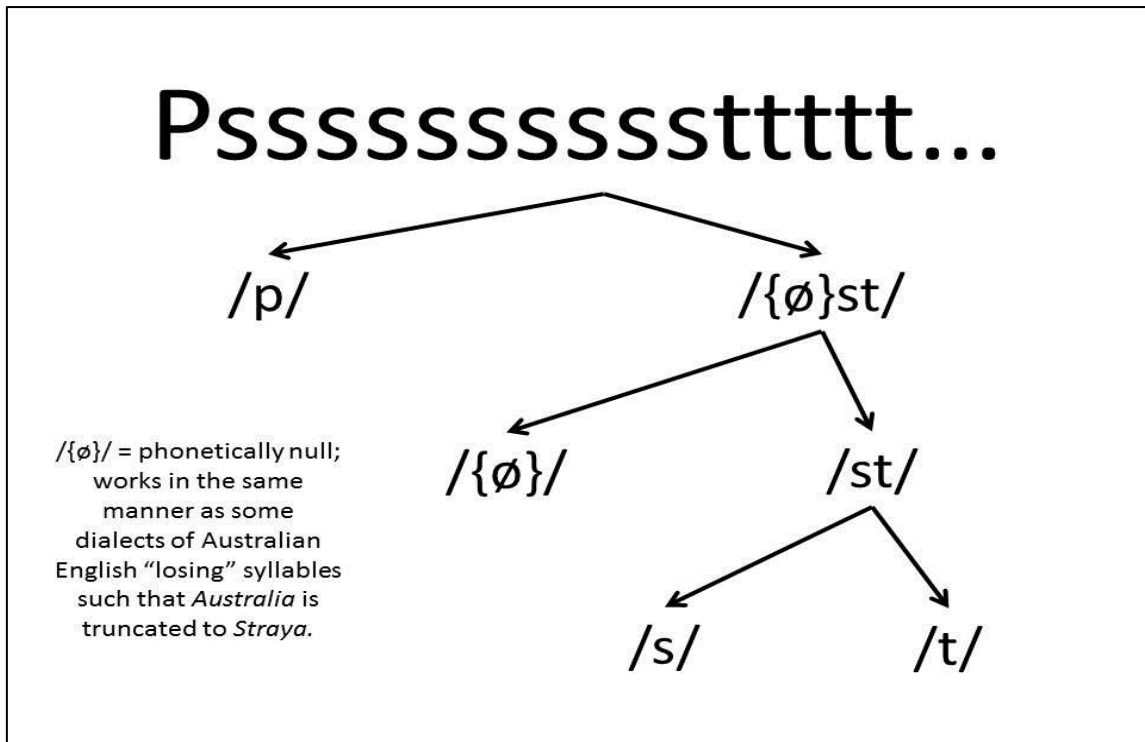
Figure 1.7: A generative syntax model of apparently vowel-less utterances.

1.6. Mirror Neurons

As defined by Rizzolatti (2006, p. 419), mirror neurons (MNs) are "…neurons that discharge when an individual performs an action, as well as when he/she observes a similar action done by another individual". As a result, MNs are generally considered to be motor neurons, although this is not accepted by all. For example, Bickerton (2005) asserts that the fundamental mirroring property of some neurons was first described by Perrett, Rolls and Caan (1982). By implanting microelectrodes into Rhesus monkey brains, recordings of 497 individual neurons in the superior temporal sulcus (STS) were obtained. Of these, at least 48 were selectively activated in response to facial stimuli, whether real or projected, human or monkey. These neurons did not, however, respond to geometric shapes, 3D shapes, or either arousing or aversive stimuli. Furthermore, when faces were transformed, such as presenting only eyes or mouths in isolation, there were differential responses within this sub-population of neurons, with some responding to eyes, others to mouths. Of particular note, these neuronal responses generally occurred without any accompanying behavioural manifestation; that is, monkey see face, monkey brain cells react in a mirror-like fashion. This thus led Perrett et al. (1982) to conclude that the Rhesus monkey brain had a specialist face recognition system located primarily in the STS.

Later recordings of single neurons in Macaque monkey brains further identified that neurons in the temporal lobe react in a mirror-like fashion to examples of hand movement (e.g. an experimenter reaching for a piece of paper), with differential responses according to the type of hand movement (Perrett et al., 1989). It was also noted that these hand-sensitive neurons were in an anatomically distinct area of the STS compared to the face-sensitive neurons identified in 1982. It was subsequently concluded that these neurons were part of a system involved in agent/object identification, and agent-object interaction intent; that is, goal orientated activity including the attribution of cause.

It was not until three years later, however, that true mirror neurons were discovered; that is, motor neurons that fired both when an action was performed as well as when observing another biological agent performing a similar action (di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992). Again using Macaque monkeys, di Pellegrino et al. clearly demonstrated that selected neurons in the inferior premotor cortex of the Macaque (i.e., area F5 located near the arcuate sulcus) responded selectively to goal-directed hand movements of the monkey. Notably, some of these neurons discharged only when the monkey gripped with its thumb and forefinger, others during finger prehension, while a third class discharged only for whole hand prehension. Unexpectedly, these same neurons discharged in an identical manner when watching an experimenter perform the various movements even without overt movement by the monkey. Di Pellegrino et al. (1992, p. 176) thus concluded that "…this surprising new class of premotor neurons…" was able to recreate movements based not only on stimulus characteristics but on the meaning of the observed actions; in effect, monkey see, monkey premotor brain cells prepare monkey to do even if monkey does not actually do it.

Following on from this, Gallese, Fadiga, Fogassi and Rizzolatti (1996), more clearly established the stimulus properties that evoke such neuronal responses. By recording activity from 532 neurons in area F5 of two Macaque monkeys, Gallese et al. delineated that the relationship between agent and object formed the basis for mirrored neuronal responses. Specifically, it was noted that neither an agent nor object was capable of causing such a reaction; rather, it was the interaction between mouth or hand agents and goal objects such as food. The strict discharge congruence between monkey seeing and monkey doing thus prompted Gallese et al. (1996, p.593) to label these neurons as "…mirror neurons…". Perhaps more importantly, Gallese et al. theorized that because monkey area F5 is homologous with Broca's area in humans, not only are MNs (or more specifically a mirror neuron system) involved in action or

movement recognition by humans, but that MNs are also involved in phonetic gesture identification and recognition; that is, language.

1.6.1. Do MNs Exist in Humans?

Until the confirmation via single cell recording of the existence of MNs in humans (Mukamel, Ekstrom, Kaplan, Iacoboni & Fried, 2010), empirical support relied on the inferential method of measuring changes in brain wave patterns.

1.6.2. Measuring MN Activity via Mu Variation.

Neural activity, as measured via EEG, can be transformed via Fast Fourier Transform (FFT) into five, functionally relevant frequency band spectra: delta Δ (0 - 4Hz), theta θ (4 - 8Hz), alpha α (8 - 12 Hz), beta β (12 - 24 Hz) and gamma γ (24 - 70Hz). Each of these spectra has been associated with specific types of cognitive state (e.g. Δ with sleep states, Richards et al., 2013), with α proposed as a reliable measure of attention and a real-time marker of processing flow (Klimesch, 2012; Klimesch, 2011). It has been further proposed that α can be divided into as many as three functionally discrete sub-components: lower (8 - 10Hz), upper (10 - 12Hz; Pfurtscheller, Neuper & Krausz, 2000) and, occasionally, middle (9 - 11Hz; Bhattacharya, Coyle & Maguire, 2011). Of these sub-components, it is a distinctive rhythm (*mu*) found within the upper alpha range that is proposed as reflecting MN activity (Cheng, Lee, Yang, Lin, Hung & Decety, 2008); more specifically, function-related changes in the oscillatory activity of groups of MNs are thought to be reflected by changes in the oscillatory activity of μ (Nyström, Ljunghammar, Rosander & von Hofsten, 2011). This is also known as event related synchronisation/desynchronisation (ERS/ERD; Pfurtscheller, 2001), which can itself be quantified through frequency power increases and/or decreases. From there, various techniques can be used to visualise ERS/ERD activity within the brain, including LORETA/sLORETA source localisation (Pascual-Marqui, 2002), Kolmogorov entropy changes (Aftanas, Lotova, Koshkarov, Pokrovskaja, Popov & Makhnev, 1997), normalised transfer entropy (Shovon, Nandagopal, Vijayalakshmi, Du & Cocks, 2015) and time-varying autoregressive pole displacement analysis (Dahal, Nandagopal, Cocks, Vijayalakshmi, Dasari & Gaertner, 2014).

1.6.3. Event Related Synchronisation/Desynchronisation (ERS/ERD)

The average human brain contains approximately 86 billion individual neurons (Herculano-Houzel, 2012). Although each of these neurons is an independent entity, neural function is effected by the coordinated activity of (probably) millions of neurons. Although

much is known about inter-neuronal communication at the cellular level (i.e. transmission of neurochemicals across the synaptic cleft), little is understood of network level communication due, in part, to the dynamics of the system; neural networks form, dissolve, then reform within a few hundred milliseconds. Such dynamics can, however, be measured, most notably by event-related synchronisation/desynchronisation (ERS/ERD; Pfurtscheller et al., 2001).

As neurons form into functional groups or *ensembles*, their individual electrical activity synchronises with the activity of other neurons in the ensemble to create synchronous oscillatory activity. By analogy, a pod of dolphins cruising off the South African coast is a discernible entity albeit with fluid boundaries; that is, it is a number of individuals forming a single functional group (Mann, 2000). To do that, each individual dolphin must (to a degree) synchronise its movements with those of the other dolphins in the group; if it did not, the pod would quickly fall apart. This is very much what happens in the first few moments when a Great White shark crashes the party; the pod disintegrates and dolphins swim in all directions. In other words, each individual dolphin desynchronises its movement with other dolphins in the pod. Depending on a variety of factors such as size of the pod and the presence of infants (Mann, 2000), the dolphins may remain in a chaotic, desynchronised state with at least a few ending up as whale food. Alternately, the pod may reform and through highly organised, coherent activity drive the hunter away.

At a neural network level, globally synchronised activity is thought representative of a state of readiness; that is, the network is connected but is processing little or no information (Klimesch, Sauseng & Hanslmayr, 2007). Desynchronised activity, alternately, is thought to represent active processing by the system. When an event requires processing, those networks associated with that specific processing desynchronise. This can be measured by comparing pre-stimulus with post-stimulus power fluctuations (increases/decreases) in the frequency band of interest (Pfurtscheller & Da Silva, 1999). If the difference is significantly higher post-stimulus, it is believed that this represents desynchronisation or active processing. In the specific case of upper α, if that desynchronisation was observed over sensorimotor areas, it would qualify as evidence of active MN processing directly related to stimulus or task demands.

1.6.5. Evidence to Support the Function of MNs and MN Networks.

Although inferring MN activity indirectly from functional and/or behavioural methods rather than through single cell recording is not globally accepted, it has nonetheless formed the basis for a number of investigations displaying compelling evidence to support the function of

MNs in humans. For example, Fadiga, Fogassi, Pavesi and Rizzolatti (1995) used transcranial magnetic stimulation (TMS) to demonstrate that animate action observation excites motor activity in human brains and that muscle activation evoked by TMS during action observation is near identical to that displayed when the action is actually performed; that is, human brains and monkey brains react to movement, whether performed or observed, in the same manner. Similarly, Grafton, Arbib, Fadiga and Rizzolatti (1996) used positron emission tomography (PET) to suggest that humans have a specialised circuit for *recognising* hand-object interactions, as well as a second specialised circuit for *imagining* hand-object interactions. This latter *imagination* circuit was posited by Grafton et al. (1996) as being homologous to the MN system in monkeys due to the anatomical similarities in activation between human and Macaque.

Extending this research, Iacoboni, Woods, Brass, Bekkering, Mazziotta and Rizzolatti (1999, p. 2516) used fMRI to examine the *direct matching hypothesis* (DMH) whereby imitation is proposed as being the direct result of a neural mechanism that maps an observed action with an "…internal motor representation of that action…". By comparing cortical areas activated in normal human participants in both observation and action conditions, two areas were identified that were similarly activated whether an action was performed or simply observed – the left inferior frontal cortex and the rostral-most region of the right superior parietal lobule.

1.6.6. Mirror Neurons and Language

Allowing that area F5 in the monkey brain is anatomically homologous to Broca's area in humans, it is not surprising that MNs have been proposed as underscoring human language evolution, acquisition and processing (Arbib, 2002). Although initial attempts to match monkey area F5 with Broca's area drew criticism (e.g. Turella, Pierno, Tubaldi & Castiello, 2009), there is a variety of evidence to support that there are more similarities than there are differences. For example, Small, Buccino and Solodkin (2010) review various studies, concluding that Macaque area F5 is a dysgranular area comprised of three distinct regions - F5c, F5a and F5p. Based on a comparison between species of the cytoarchitectonic feature of granularity in cortical layer IV, Small et al. (2010) suggest that F5c equates to the ventral premotor (PMv) cortex in humans; F5a equates to the human pars opercularis; and F5p to those areas in between the pars opercularis and the PMv.

This is well in keeping with Rizzolatti and Fabbri-Destro's (2009) assertion that humans possess an expanded MN system that includes the inferior parietal lobule as well as F5 areas, most notably the PMv and pars opercularis. Known as the parieto-frontal mirror system

(*PFMS*), Rizzolatti and Fabbri-Destro suggest that certain nodes in the PFMS display somatotopic organisation. The pars opercularis and precentral gyrus, for example, show a "…classical homunculus…" when watching motor acts using the leg, hand or mouth (Rizzolatti & Fabbri-Destro, 2009, p. 224). In the inferior parietal lobule, however, mouth movement observations appear to activate rostral areas, hand/arm movements in more caudal areas, and leg movements even more caudal, also extending into the superior parietal lobule. That similar somatotopic organisation was reported in the inferior parietal lobule of macaque monkeys (Rozzi, Ferrari, Bonini, Rizzolatti & Fogassi, 2008) adds further credence to the F5/Broca's homology.

1.6.7. A Possible Role for MNs in Language Learning/Acquisition.

If, as has been argued by a number of authors (e.g. Chomsky, 1969, 1986; Senghas & Coppola, 2001), there is insufficient environmental information for an infant to acquire language by mere exposure (i.e. PoS), then language must have an innate or instinctual (Pinker, 1994) component; precisely, what this component is remains conjectural, however, a specific *language acquisition device* (LAD; Chomsky, 1969, 1986) has been proposed. Mirror neurons, or more precisely mirror neuron networks (MNNs), could potentially form the substrates of such a *device*, which would not only allow an infant to acquire language with minimal exposure but potentially allow mature language users to learn new words or languages. Furthermore, given the specific role of MNs in action and movement imitation (Di Pellegrino et al., 1992; Fadiga et al., 1995; Rizzolatti et al., 1996), MNNs may respond differentially to the learning of new words representing action concepts (i.e. verbs) when compared to the learning of new words representing object/thing concepts (i.e. nouns).

Although highly conjectural, such a proposal is not necessarily impossible. For example, sound waves represent the vibration of molecules through air or liquid therefore represent a type of movement. These movements, if reflective of speech, are strongly constrained by the intonation patterns and rhythms of the specific language type (Jackendoff, 2002). If human foetuses develop a basic MN network prior to, or at around the same time as, basic movement perception, then repetitive *in utero* exposure to the vibrations or movements of speech may guide the basic neural circuitry that, after birth, gives the infant a basic ability to parse complex sequences (or rhythms) into smaller parts. This is at least circumstantially supported by the observation that although newborn human infants can respond to every known human language phoneme (Menn & Stoel-Gammon, 1995) they *prefer* the sound of their mother's voice to those

of strangers (DeCasper & Fifer, 1980); that is, they prefer the movement of the air molecules by their mother's voice. Newborns also prefer the sound of their mother's language than that of an unknown language (Mehler, Jusczyk, Lambertz, Halstead, Bertoncini & Amiel-Tisson, 1988); for this to occur strongly suggests that human infants are familiar with the rhythms of the language environment into which they are being born *prior* to actually being born. Language rhythms, in isolation, are not syntax; they are, however, movement based, and form identifiable, highly-repetitive patterns or sequences. As a result, the mirroring of these sequences by the foetal brain over hundreds or perhaps thousands of exposures may provide the scaffold from which syntax, or the ordering of subjects, verbs and objects, can later develop.

This basic circuitry, in conjunction with normal infant brain plasticity (Kolb, Gibb & Robinson, 2003) should be sufficient to allow this network to expand in an exponential manner in response to neonatal exposure to various environmental inputs (such as sound, vision or touch). Once cognitive, motor and sensory processes had developed and expanded sufficiently, MNNs could then, through their ability to imitate/simulate actions, provide a further scaffold for the vicarious learning of higher aspects of language (e.g. prosody and complex syntax) which are neither modality (speech versus sign) nor lexicon dependent; that is, they are more universal general purpose language features. For example, an infant born into an English-language speaking community would, through pre-birth exposure, be attuned to the intonation patterns of English. These patterns, which are in large part syllable based (Ladefoged, 2005), would provide both the substrate and basic motivation for learning after birth; every time the pattern was re-encountered the MN network would activate. MNs would also activate in response to observed movement, so, for example, an infant which "smiles" at a caregiver will generally evoke a return smile, often associated with exclamations of delight and/or hand claps and debate as to whether it was a real smile or simply wind. This would then, in turn, simultaneously activate MNs in the infant brain related to intonation as well as mouth and/or hand movements. This dual firing would then form the first neural association between the patterns *of* language and the patterns of movement associated *with* that language.

At the same time, the syllabic nature of intonation would provide a reference point for parsing complex sound sequences into increasingly smaller parts; a first step into language. As the infant develops, associations between MN network activation, movement (both observed and performed) and communication would increase in both size and complexity until the broad concepts of *subjects, objects* and *actions* emerge. Once these concepts emerge, then simple language begins to emerge with its development primarily constrained by motor development

(e.g. the ability to control vocal apparatus) and experience (for lexical development). It is only when the higher-order concept of *interactions* between basic concepts (e.g. subject-actions-object) develops, that a need to differentiate between these basic concepts becomes necessary. This, in turn, leads to the development of nouns and verbs as distinct but inter-related means of conveying different types of information (objects/things versus movement/action).

1.6.8. A Specific Role for MNs in Verb Processing?

Given the association of MNs with movement, it may be that MNs specifically support only action (or verb) learning and/or processing in relation to language. For example, Damasio and Tranel (1993) reported a neural double dissociation in the retrieval of concrete nouns and verbs. By comparing lesion sites, Damasio and Tranel (1993) suggested that there are two distinct although intimately inter-related neural systems which mediate lexical retrieval of the two syntactic classes. For nouns, it was suggested that this system lies within the left hemisphere, particularly the left anterior and middle temporal lobes, while the system underlying verb retrieval is located primarily in the left premotor cortex. These areas, the authors note, lie outside of the neural areas traditionally associated with language comprehension and production, so it was suggested that the systems underlying at least some areas of language processing, specifically the processing of nouns as opposed to verbs, are not language-specific; rather, that they form part of a widely distributed neural network associated with concept retrieval that is proposed to be hierarchical, starting with the visual cortices in the occipital lobes, progressing ultimately to the entorhinal cortex. Verbs, alternatively, are proposed by Damasio and Tranel (1993) to be less stratified, involving dorsal aspects of the temporal, occipital and parietal cortices, with projections to both the premotor and prefrontal areas. Allowing that some areas within this suggested verb processing network have been nominated as likely sites of MNs in humans (e.g. the pre-motor cortex and inferior parietal cortex; Molenberghs, Cunnington & Mattingley, 2009), this suggests a basic, fundamental processing difference between the two syntactic types; nouns, as proposed by Damasio and Tranel (1993), are processed with reference to concept memory systems, while verbs are computed on-the-fly in reference to motor related patterns generated internally (i.e. modelled or simulated) by the actions of mirror neurons.

1.7. Current Grammatical Class Perception Models

Grammatical class perception currently has two mutually exclusive working models. The first, the one-path model (1PM), posits that, in isolation, a word cannot be determined as noun or verb prior to lexical access; that is, grammatical class is tagged and retrieved via linguistic

memory processes (e.g. Siri et al., 2008) thus cannot be processed early (<300ms) by different pathways prior to such access. In modern English this is undoubtedly true for numerous words when heard in isolation. For example, the word *row* could mean something that you do gently to your boat down the stream, or it could mean how pretty maids grow in Mary's garden. In this case, it is purely a matter of context. Modern English is, however, a hybrid rather than pure language so must accommodate different ancestral language rules including lexical replication (e.g. *row* the verb and *row* the noun) (Ladefoged, 2005). As a result, contextual grounding by either a functional linguistic marker such as an article (e.g. *the*) or the word's syntactic position within a sentence is sometimes required for meaning to be conveyed, although this is not always the case; for example, a word such as *differentiate* is able to stand alone. The complexities of modern English do not, however, necessarily mean that language in a more general sense is processed via a singular path thus the alternate two-path model (2PM) has been proposed in which it is suggested that grammatical class differentiation occurs at an unconscious neural level early in perception (some suggest <100ms, but up to 300ms has been reported e.g. Pulvermüller et al., 1999). Furthermore, the 2PM posits that this early differentiation between nouns and verbs is fundamental to the processing of the word, including different processing paths and lexicons. Unfortunately, despite an apparent wealth of evidence to support both the 1PM and 2PM, much of the previous research in this area is contradictory and has had methodological questions raised (see Chapter 2), thus it remains unknown which, if either, model is correct.

1.7.1. Two-Path Model

Support for the 2PM is extensive. For example, case studies were used by Shapiro and Caramazza (2003) to suggest that the morphological impairments shown by agrammatic aphasics are not due to morphological disruption; rather, that morphology is disrupted due to syntactic constraints on the type of morphological change specific word classes can undergo. Just as nouns cannot be transformed to display tense, verbs cannot display number; thus it is damage to the neural substrates underlying the distinction between nouns and verbs which manifests as morphological impairment. Based on lesion location, it was proposed that both the conceptual and linguistic processing of nouns involves the left posterior frontal lobe, the anterior prefrontal cortex, and anatomically adjacent parts of the insula, putamen and pallidus. Additionally, some involvement by the parietal operculum and the superior-anterior temporal lobe was implicated. Verb processing, although also suggested to lie in the left frontal cortex, was proposed as recruiting resources from nearby but distinct areas, including the left midfrontal gyrus (approx.

BA9), the parietal cortex and the paraventricular white matter subserving the supramarginal and posterior temporal regions.

Pulvermüller et al. (1999) also reported a physiological noun/verb double dissociation within 200 milliseconds (ms) of stimulus presentation. Using an event-related potential (ERP) paradigm, current source density (CSD) analysis was employed in conjunction with ANOVA; results suggested significant activation of central compared to occipital sites at 60 - 90, 120 - 140 and 200 - 230ms post-stimulus presentation, with the latter time window showing a significant regions-of-interest (ROI) x syntactic class interaction between C3/4 and O1/2, and FC5/6 and O1/2. T-test comparison of noun and verb evoked potentials subsequently showed that nouns produced significantly more ingoing current flows at occipital sites while verbs displayed significantly more ingoing current flows over premotor and motor cortex areas. These results, the authors concluded, provide reasonable evidence that the neural generators subserving noun perception are most likely located in the primary and/or higher order visual cortices within the occipital lobe. Verbs, alternately, appear to have their neural generators located in or near the primary motor cortex, the pre-motor cortex and the adjacent pre-frontal cortices.

1.7.2. One-Path Model.

As with the 2PM, support for the 1PM is extensive. For example, Tyler, Bright, Fletcher & Stamatakis, 2004) used fMRI to propose that inflectional morphology (e.g. inflecting nouns with the suffix "-s" to denote plural) alters syntactic processing. Results showed significant activation of the left hemisphere frontal/temporal region regardless of syntactic class which was proposed as being consistent with the findings of Vandenberghe, Price, Wise, Josephs and Frackowiack (1996) that this area is involved in lexical-semantic processing. When noun stimuli were specifically compared to verb stimuli, the only significant difference was found in the left inferior frontal gyrus (LIFG) with verbs showing increased activation relative to nouns in the left inferior frontal cortex (BA 44, 45 and 47). By drawing comparisons with previous studies (e.g. Tyler, Russell, Fadili & Moss, 2001), Tyler et al. (2004, p. 520) concluded that nouns and verbs are "…represented in the same unitary distributed fronto-temporal network…" as opposed to different "categorically organised neural system(s)".

This position is also supported by Siri et al. (2008) and Yokoyama et al. (2006) but despite the evidence, it is not entirely convincing. For example, Tyler et al. (2004) used visual rather than auditory stimulus presentation. This could therefore mean that results were attributable to *literacy* rather than *language* processing. Also, Tyler et al. (2004) used activation

analysis which, as noted by McIntosh (1999), detects functional differences rather than functional interactions and thus may not detect the connectivity associated with large scale functional networks. From a more linguistic perspective, it is also possible that inflecting nouns, is not the same fundamental process as inflecting verbs. Unlike verbs, nouns can only be directly inflected in English one way; that is, to show plural. Verbs, on the other hand, can be inflected in a variety of ways to denote tense, including adding "-ing" and "-ed", with or without the use of additional auxiliary verbs such as "to be". If nouns and verbs are at least in part processed early via cognitive functions such as the ability to understand number, the use of pre-pluralised nouns as stimuli may have expedited noun perception to such a degree that it was invisible to Tyler et al.'s (2004) fMRI paradigm. If so, the enhanced processing of verbs in the LIFG may indeed be related to morphology, but *after* basic syntactic processing had occurred. This same criticism can also be applied to Siri et al. (2008) and Yokoyama et al. (2006), who both used inflected versus uninflected stimuli. As with Tyler et al. (2004) the choice of stimuli may have expedited perceptual processes such that fMRI could not detect any early, automatic noun/verb dissociation.
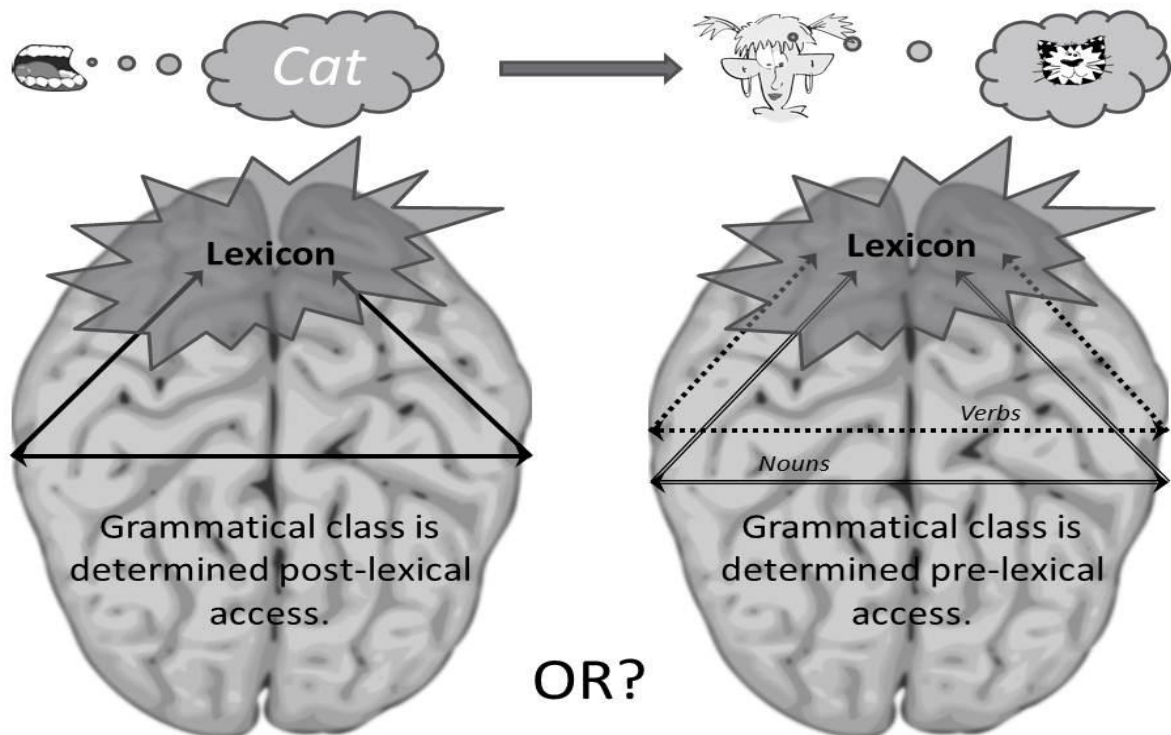

Figure 1.8: 1PM (left) versus 2PM (right).

1.7.3. Acoustically Comparing the One and Two-Path Models.

It is, perhaps, surprising that the debate over grammatical class processing continues because if grammatical class differentiation does indeed start within 300ms of a word's acoustic onset, then that differentiation must occur at an unconscious level. If so, then significant variations in physical acoustic properties between nouns and verbs must also be measurable within this time frame. Although there is some evidence to support this, a systematic, physical investigation across stimulus word sets and different voices does not appear to have been conducted. This may, in part, be due to the complexity and sheer volume of data such an investigation would require, however, modern signal analysis techniques may help to overcome this problem. Alternately, it may be that such a differentiation does not exist; at this point, it remains unknown. Given the mutual exclusivity of the two models, such an investigation would clearly support one or the other: if a differentiation occurs in <300ms of a word's onset then the 2PM is supported; if differentiation does not occur in <300ms, then the 1PM is supported, even if only by default.

1.8. Purpose of the Current Thesis.

The primary purpose of the current thesis is to compare the 1PM and 2PM from a new perspective and interpret experimental results within an evolutionary context; that is, if evidence is found to support the 2PM, it will be proposed that this is more in keeping with the co-evolution model of language than either the proto-speech or proto-sign models. Similarly, if evidence is found to support the 1PM, it will be proposed that this supports language evolving from a singular lineage such as proto-speech or proto-sign. Comparison of the two models (1PM and 2PM) will be undertaken by: (i) addressing previous methodological issues in the research literature; (ii) establishing if, at an unconscious neural level, human speech is processed differently to other types of communicative sound strings and thus has language-specific early components; (iii) examining early neural responses to nouns and verbs to identify the presence or otherwise of differential processing; and (iv) examining the role, if any, of mirror neurons in mature language processing and/or learning.

1.8.1. Overcoming Previous Methodological Concerns

As previously mentioned, there is a large pool of empirical data examining grammatical class (nouns versus verbs) processing. There is an even larger pool of neurolinguistic data

examining the neural substrates of human language. Unfortunately, many of these previous investigations have contained small but important methodological issues which could potentially have biased both results and interpretations. For example, numerous studies (e.g. Pulvermüller et al., 1999; Longe et al., 2007) have used written rather than spoken stimuli. Although *literacy* (reading and writing) is a means of recording speech it is not necessarily processed within the human brain in an identical manner to language (Olson, 1998) and there is some neural evidence emerging to support differential processing of literacy and language (Schlagger & McCandliss, 2007; Preston et al., 2010). Furthermore, although congenitally deaf individuals easily acquire language (sign) through simple exposure, their ability to become literate is seriously compromised by a physical inability to map phonemes onto graphemes (LaSasso & Mobley, 1997). This suggests then, that using written rather than spoken (or signed) language may be capturing different processes. To address this potential confound, the current thesis uses only auditory stimuli in a sample of normal hearing participants.

A second potential issue with previous research is the use of normative databases to control stimuli across cognitive dimensions such as concreteness and familiarity. Although there are a number of databases available (e.g. CELEX, SUBTLEX, the MRC Psycholinguistic Database), the published values have primarily been obtained from written sources, or structured speech formats such as movies or lectures. Furthermore, most of the published values are based on data gathered from the Northern Hemisphere (most notably the USA) which may or may not be the same as data gathered in the Southern Hemisphere. For example, although both Americans and Australians speak English, there are a number of dialectal differences including differences in vocabulary (e.g. the meaning of the word *root)* and stress patterning (e.g. or-e-**ga**-no versus o-**reg**-a-no) which could potentially influence values assigned to different cognitive constructs. Some of the existing databases (e.g. the MRC; Coltheart, 1981) are also potentially approaching their use-by-date, given that the published values are based on data gathered over thirty years ago; as shown by Senghas and Coppolla (2000), languages can changes dramatically during such a time-frame. To this end, the current thesis will investigate the validity of using normative database values by comparing data values obtained directly from a prospective sample population, with those obtained from the MRC, British National Corpus (BNC; Aston & Burnard, 1998) and CELEX (Baayen, Piepenbrock, & Gulikers, 1995).

A larger problem with speech research in general is the neglect of the acoustic properties of speech stimuli. As will be examined in Chapter 3, sound waves (such as single spoken words) contain a large amount of physical, measurable information such as pitch, amplitude, shimmer,

jitter and voicing. Such information can be extracted using software such as PRATT (Boersma & Weenink, 2001; see section 1.8.1.1 following) or KayPENTAX (KayPENTAX Inc., Lincoln Park, NJ) and is used extensively within clinical areas such as speech pathology (e.g. Albertini, Giaquinto & Mignano, 2009). This information is not, however, regularly (if ever) reported in more research based areas such as neurolinguistics or psycholinguistics. As will be demonstrated in Chapter 3, acoustic measures can vary significantly between constructs (e.g. noun versus verb) but these variations are not consistent; rather they are voice and/or corpus specific. This thus suggests that previous research that does not take acoustic considerations into account, may be reporting sound feature rather than language processing. In the current thesis, all stimuli have been compared across acoustic measures using PRATT.

From a slightly different perspective, the choice of baseline measures in previous research could possibly be improved. For example, computer generated tones or white noise are often used as baseline comparisons in language studies (e.g. Iacoboni, 2008). Although such baseline measures are designed to allow the separation of sound processing from language processing, artificial sounds or sound strings do not have a communicative function and do not contain any inherent meaning. As a consequence, results obtained using such baseline measures can separate sound from communicative intent but not necessarily from human language processing. In the current thesis, baseline measures were chosen which were both complex in their acoustic nature as well as communicative in intent; that is, human speech was compared with a variety of animal vocalisations thereby giving any potential claims to language specificity greater credence,

Finally, there is potentially an issue with previous research in the technology chosen to examine language processing. Although fMRI has excellent spatial resolution (mm^3), its temporal resolution can be in the order of several seconds due to its reliance on changes in oxygen levels in the cerebral blood flow. Language, or in the case of the current thesis spoken language, is perceived and processed over the course of just a few hundred milliseconds in the case of single words (e.g. *paddle* = 500ms). As a result, fMRI provides excellent insight into the location of cortical hubs involved with, but not necessarily specific to, language processing but cannot identify more transient areas of activity which, from a temporal perspective, are more likely to be language-specific. By comparison, electroencephalography (EEG) by itself provides millisecond temporal resolution but relatively poor spatial resolution (cm^3). The latter problem has, to a large extent, been overcome by the development of LORETA software (Pascual-Marqui, 2002; see Section 1.8.1.2 following) which in its latest release suggests zero source

localisation errors when used to analyse EEG data. Therefore, if data is gathered via EEG and then analysed using LORETA, an EEG based paradigm should be able to provide both the temporal and spatial information needed to more adequately tease apart language processing. In the current thesis, EEG data is gathered via 64 channels from 28 adult, human volunteers as they undergo two different tasks. This data is then analysed using eLORETA (latest software release). To confirm the validity of the approach, results are then compared against previously published anatomical lesion studies (e.g. Shapiro, Moo & Caramazza, 2006). See Chapters 4 – 6.

1.8.1.1. PRATT

PRATT (Boersma & Weenink, 2001) is freely available acoustic analysis software used widely in the areas of phonology and phonetics (e.g. Pater, Jesney & Tessier, 2007; So, Jia & Cai, 2012; Amir & Grinfeld, 2011). Like professional sound editing software (e.g. Adobe Audition), PRATT provides the user with a variety of tools for trimming sound files, as well as modifying sound files (e.g. changing phase) and allowing users to zoom to sub-millisecond changes in wave form patterns. PRATT further provides tools for extracting acoustic measures from either the complete sound file, or selected sections of the file. These measures include amplitude (minimum, maximum, mean, SD – although note that PRATT provides measures in Pascals rather than decibels), root mean square (RMS), total energy, number of pulses/periods, voicing (fraction of locally unvoiced frames), pitch (mean, SD), jitter (RAP), shimmer (local), mean harmonics to noise ratio and intensity/loudness (mean).

In the current thesis, PRATT was used to trim sound files to 500 ms (+/- 0.3 ms). This length was chosen based on the length of the shortest word to ensure that no sample was contaminated by unrelated silence. Files were then further divided into four additional time windows – 0 - 100ms, 100 - 200ms, 200 - 300ms and 0 - 300ms. PRATT was then used to extract a visible spectrogram, pitch contour, intensity contour, formant contour and visible pulses for each of the sound segments. To gain more precision in the measurements, PRATT's pitch settings were optimised for voice analysis (cross-correlational) rather than intonation analysis (auto-correlational), with the default minimum pitch of 75 Hz was reduced to 50 Hz to allow for the pitch and resonance of the speakers used. PRATT's Voice Report and Information functions were then used to extract each file's acoustic values, with these values then exported to and analysed by SPSS V21 (IBM Corporation) to assess the statistical significance of any differences.

1.8.1.2. LORETA

Low resolution brain electromagnetic tomography (LORETA; Pascual-Marqui, Michel & Lehmann, 1994) specifically addresses the inverse problem inherent in localising neural sources from scalp recorded EEG data; that is, the electrical fields topography observed at the scalp is not a direct representation of neuronal activity on the underlying cortical surface let alone in deeper cortical sources. LORETA, and its subsequent developments sLORETA (Pascual-Marqui, 2002) and eLORETA (Pascual-Marqui, 2007), address the problem by working on the basis that EEG is the manifestation of current density distributions resulting from post-synaptic neural processing of cortical pyramidal cells arranged in parallel columns perpendicular to the folded cortical surface.

Based on the scalp-recorded electric potential distribution, sLORETA (standardized low resolution brain electromagnetic tomography) software (http://www.uzh.ch/keyinst/loreta.htm) was used in the present study to compute the cortical three-dimensional (3D) distribution of the ERD/ERS (see Section 1.6.3) transform of current density. ERD/ERS is a linear transform of the EEG signal and is therefore consistent with the mathematical assumptions of sLORETA which is itself a form of linear decomposition of the EEG signals at scalp electrodes that yields a set of latent variables corresponding to the signal at voxel sources; that is, it is a properly standardized discrete, 3D distributed, linear, minimum norm inverse solution. This endows sLORETA tomography with the property of exact localization to test point sources, yielding images with exact localization although low spatial resolution (i.e. neighboring neuronal sources will be highly correlated). The description of the method can be found in Pascual-Marqui (2002) and proof of its exact, zero-error localization property in Pascual-Marqui (2007, 2009).

The sLORETA software package was also used to perform the statistical analyses. The methodology used is non-parametric and based upon estimating, via randomisation, the empirical probability distribution for the max-statistic (e.g. the maximum of a t or an F statistic), under the null hypothesis. This methodology also corrects for multiple testing and due to the non-parametric nature, does not rely on any assumption of Gaussianity. For a complete overview of this method, see Nichols and Holmes (2002). Used together, these two features of sLORETA provide visual localisations as 3D cortical maps and slice maps as well as a maximum t or F statistic, indicating significant differences between conditions. See Figure 1.9.
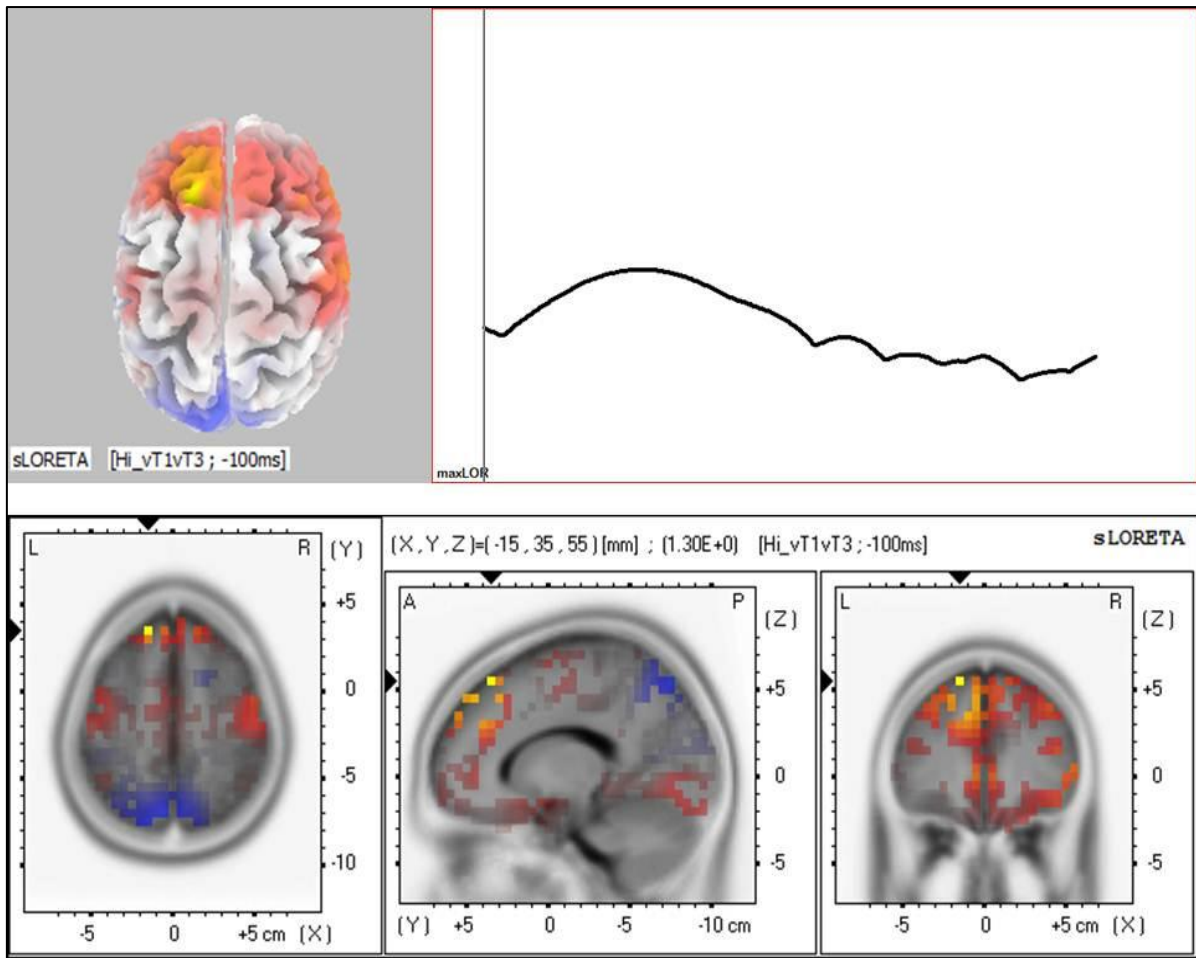
Figure 1.9: sLORETA output example - 3D cortical map (top left), maximum voxel wave form (top right), and slice map (bottom).

LORETA/sLORETA have subsequently been validated numerous times, and applied in a variety of research and clinical contexts. For example, LORETA forms an integral part of NeuroGuide™, a commercial Z score EEG biofeedback system used extensively in clinical neurofeedback settings (Thatcher, Biver & North, 2004). sLORETA has also been used to investigate neurocognitive deficits in male alcoholics (Pandey et al., 2012); identify the sources of symptom provocation in phobias (Scharmüller, Leutgeb, Schäfer, Köchel & Schienle, 2011); assess cerebral ischemia (Sakamoto et al., 2010); detect medial pre-frontal default network anomalies in attention-deficit hyperactivity disorder (Cannon, Kerson & Hampshire, 2011); explore memory enhancement through hypothalamic/fornix deep brain stimulation (Hamani et al., 2008); monitor the effects of acute hypoxia on cortical activity (Schneider & Strüder, 2009); and search for the neural correlates of moral decision making in different cultures (Wang, Deng, Sui & Tang, 2014).

1.8.2. A Novel LORETA Analysis Approach

As shown in Figure 1.9, sLORETA provides a maximum voxel statistic waveform output. By identifying the turning points (local maxima/minima) of the time series of this voxel statistic, the cortical sources where event related activity maximally differentiates experimental and control conditions can be retrieved from sLORETA for each turning point. See Figure 1.10.
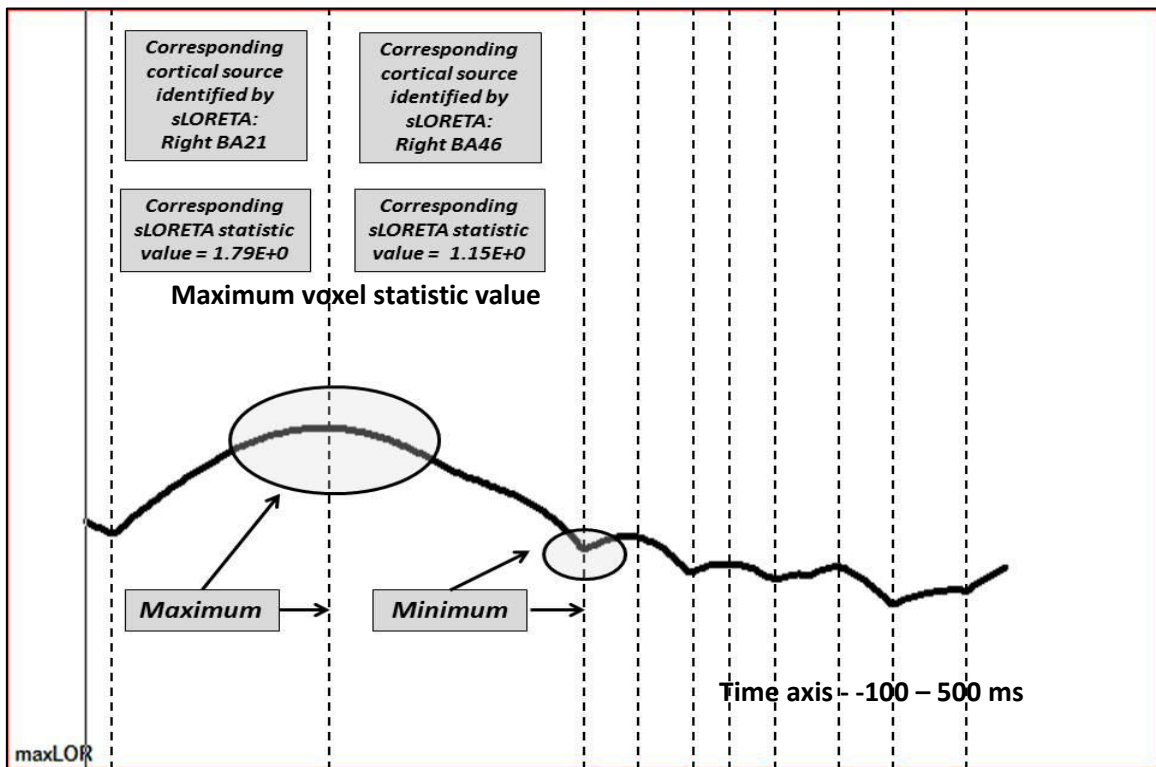

Figure 1.10: Corresponding statistic values and cortical sources can then be retrieved from sLORETA outputs.

If interpreted within the framework of travelling alpha waves spreading throughout the cortex (Klimesch, 2011) and that such waves gate the flow of functional processing activity in specific regions at specific time points as a specific processing response unfolds, it becomes possible to visualise the sequence of regional processing differences (corresponding to the turning points in the flow of cortical information processing) as a functional processing map. See Figure 1.11.
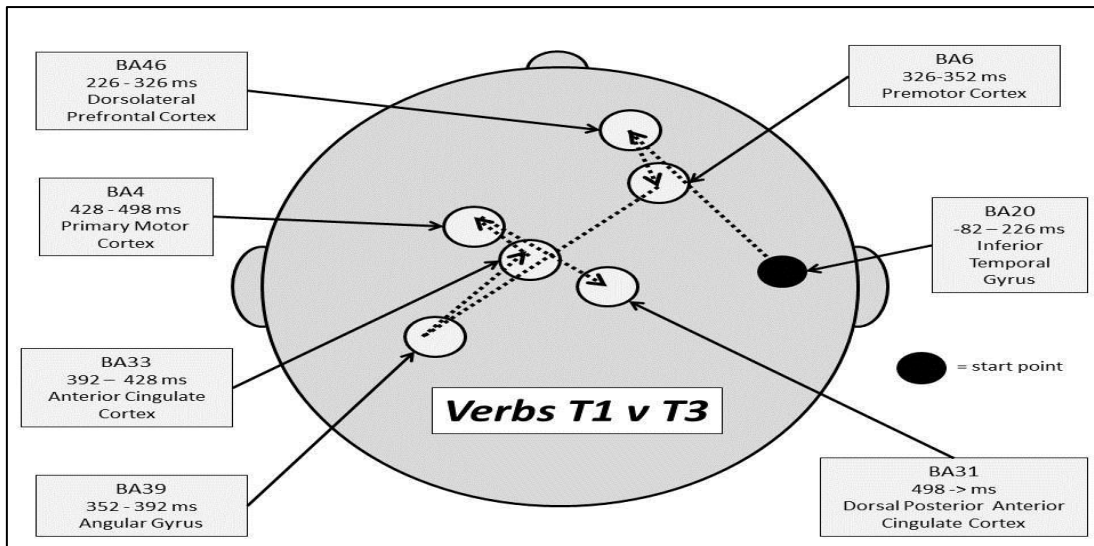
Figure 1.11: Cortical topography can then be plotted including direction of processing.

Although this approach has not (to the best of the authors' knowledge) previously been used, it will be validated in later chapters by comparing the results obtained in this manner with previously published anatomical studies, most notably lesion studies such as Damasio and Tranel (1993).

1.9. Things to Note

The current thesis seeks to demonstrate neurophysiological evidence for early (<300ms) differences in grammatical class processing. As such, the focus is on neurophysiological processes preceding conscious awareness and behavioural responding. It is certainly possible that such processes may be open to modulation by executive control through mechanisms such as attention however that is not a matter of concern in this thesis. However participants could not *not-hear* (fail to register) the stimulus therefore could not *not-process* the stimulus (early processing is automatic and obligatory). Therefore, this thesis restricts itself to examining neural measures (EEG) and does not address the later behavioural measures (such as error rates or reaction times) which may be related to each task. Such measures are generated by later neural processing that follows the activity of central interest here.

It should also be noted that this thesis is being presented in journal-article style format with Chapters 2 – 6 designed to stand-alone. As a result, several sections of the following chapters (most notably the Methodology and broad data analysis approach) will appear highly repetitive.