Towards a Connectionist Model for Minimalist Syntax

Matthijs Westera

February 8, 2009

1 Introduction

Many neuroscientists, linguists and computer scientists attempt to unravel the secret of language. Most (though certainly not all) do so in relative independence, with their own means. The goal of this paper is to show how combining the three perspectives may help in making conceptual decisions in linguistics, designing neuroscientific experiments and developing an artificial language system. I do not mean to deny the great value of the findings in each field separately, and the interdisciplinary approach employed here is only able to touch on each subject from the outside. But I feel that it is a valuable, clarifying touch.

The strategy employed here will be as follows. I will formulate linguistic key elements in terms of connectionist mechanisms, derived from findings in computer science. During this translation process, I will relate decisions to neuroscientific data. The linguistic framework employed here will be the *Minimalist Program.* Its defining characteristic is the goal to remove all unnecessary theoretical artifacts and reduce the conceptual machinery to (ideally) the pure necessities. In this sense, the Minimalist Program is more fit than any other linguistic framework for the goal of this paper. A choice between two mutually exclusive views in the Minimalist Program is first made in section 2. Further specifics of the Minimalist Program will be reviewed in section 4.

The influence of computer science in this paper is apparent mainly in my decision to assume a *synchrony-based encoding* underlying the language faculty. Neuroscience is not yet fully able to actually destill from brain-imaging data the proof for a synchrony-based encoding, but certain findings, for instance on brain waves, do hint in this direction. Synchrony-based encoding is explained in section 3. My usage of this connectionist mechanism is based mainly on the dissertation of Makino ([16]) and earlier work by Henderson ([10]), although it differs in some important respects.

The Minimalist Program is just that: a program. It is not, and does not claim to be, a complete theory of language. There are many contradictory views within the Minimalist Program, and many open questions exist. This makes it difficult to isolate the key elements, and throughout this paper decisions will have to be made regarding which theory of language to pursue. These decisions will be made based on findings from psychology and computer science. They are not meant to be ultimate, but rather to illustrate the cooperation between the different fields. This does not mean the decisions were taken nonchalantly, or that the decisions should not be taken seriously by linguists - they should.

As we shall see, a constituent-less approach advocated by, among others, Bowers ([3]) and Collins ([4]) will turn out to be the best candidate for modelling by means of a sychrony-based encoding. This is explained in section 4. A synthesis of my findings, in the form of a set of design principles for a connectionist model, is finally given in section 5.

1.1 A note on terminology

This paper concerns the physical notion of a *phase*, i.e. the displacement of an oscillation as a fraction of the complete cycle. The symbol normally used to denote (physical) phase is the Greek letter theta (θ), not to be confused with the linguistic convention of denoting thematic roles by "theta roles". In neuroscience, oscillatory patterns found in EEGs have been assigned different names based on their frequency bands. One type of brain waves that is interesting for our purposes here is called 'theta waves', oscillating between 3 and 6 Herz, and not to be confused with the 8 Herz oscillations in rodent hippocampi that have, mistakenly, also been called 'theta waves'.

Linguists may find the terminology confusing for a different reason: within the Minimalist Program, the word "phase" refers to a part of a logical form that can be derived in relative isolation, in a separate stage. To rule out any source of confusion, let me just stress that the minimalist notion of a phase does not occur in this paper at all. I am confident that simply being aware of these possible sources of confusion solves the terminological problem.

2 The language faculty

For a long time the so-called T-model (1) has been the dominant view on the language faculty, at least with respect to language production. Sentence production begins with some Deep Structure, which is translated into a Shallow Structure. The derivation then splits into a trajectory towards a logical form (for the conceptual-intentional interface) and an phonetic form (for the motor interface). Changes to the DS that occur before SS result in overtly moved constituents. Changes after that, on the trajectory towards LF, are covert (i.e. not phonologically realised).

(1)
$$\begin{tikzcd}
\text{DS} \arrow[r] & \text{SS} \arrow[r] \arrow[d] & \text{LF} \\
& \text{PF} &
\end{tikzcd}$$

One of the major enterprises of the Minimalist Program has been to eliminate DS and SS from the conceptual machinery. The conceptual difference between overt and covert syntax is removed accordingly, and instead the Minimalist

Program makes a distinction between strong and weak features, strong features requiring structural material (e.g. categories) to move along in a pied-piping fashion when features are displaced to be checked. This view results in more elegant (minimalist) view of the language faculty, as a simple linear pathway from a collection of lexical items (the numeration N) to a logical form LF, which is then translated into some phonological form PF (2). For a more detailed overview and the empirical consequences of this change I refer the reader to chapter 9 in [12].

(2) N → LF → PF

A different view exists that can be classified as semantocentric minimalism. The most important argument against a model such as 2 is that common sense tells us that sentence formation does not begin with some words in the lexicon, but that it begins with a thought or an idea. Escribano, a proponent of semantocentric minimalism, writes the following:

> The linguistic process cannot start with a random choice of lexical items [...], and cannot possibly consist of a blind autonomous bottom-up computation that only by accident succeeds in producing a coherent output, being aborted or computed but thrown away as illegible or useless most of the time [...]. [7]

He proposes a model in which sentence formation begins with a logical form, shown schematically in 3 (his proposal is a lot more complex).

(3) LF → ... → PF

In my opinion, however, things need not be as black-and-white as Escribano suggests. I share his concern that the language faculty does not blindly select lexical items and produce constructions hoping they converge. Undeniably, linguistic derivations need to commence with some thought or idea. However, this thought need not be equated with the logical form. In fact, it is quite unlikely that all our thoughts are standardly represented in a neat, logical fashion.

Additionally, the numeration is not a blind collection of some words, as Escribano seems to suggest. Instead, it already contains a lot of valuable information on the meaning of the individual lexical items, but also on the way in which their meanings should be combined. Among others, Hornstein argues that theta roles (e.g. Agent, Theme, Place, Manner, for an overview see [19]) are simply a special kind of features ([11]), which would entail that they are already specified in the numeration. Theta roles contain a lot of semantic information, and are plausible building blocks for cognitive representations of the world ([9]).

When regarded as the T-model (either the standard or the linear minimalist version), the language faculty can be used to translate thoughts both into a serialised form, ready for communication, and into a predicate-argument structure that 1) functions as an extra check on the linguistic derivation and 2) may be used for reasoning, means-ends planning, etcetera - all those processes that seem to require logical representations of the world and seem to be uniquely human. From an evolutionary perspective, it is attractive to connect our ability for logical reasoning with our ability to speak and comprehend language. Given the concerns outlined in this section, the (minimalist version of) the T-model will be the backbone of our considerations throughout this paper.

3 Connectionist considerations

In this section I will explain the basic connectionist mechanism that will be employed in this paper: synchrony-based encoding. After introducing this encoding, I will propose a simple way of representing asymmetry, a prerequisite for dealing with language.

3.1 Synchrony-based encoding

Combining multiple neural representations of features, words, percepts, etcetera, to form a single composite representation is often called *binding* (not to be confused with the linguistic notion of binding, from the Government and Binding framework). Binding in connectionist systems is explained in detail by Makino ([16]), and I will only give a very short summary here. Because of the infinity of bindings possible, binding is required to be an additive operation, meaning that the larger signal can be built up from its two parts (i.e. pure *compositionality*). However, if a connectionist system is purely additive, then the representation of "John loves Mary" is the same as the representation of "Mary loves John", as figure 1 shows.

The only way to deal with the binding problem is to add complexity to the system. In [16], three possible sources of complexity are discussed. Spatial complexity, i.e. assigning different bindings to different locations in the brain, is unfit due to the infinitude of possible bindings. Intensive complexity, i.e. letting the intensity of the signal contribute to its meaning, is argued to lack representational precision. This leaves temporal complexity, i.e. let the timing of a signal contribute to its meaning, as the only candidate.

Through temporal complexity we arrive at a *synchrony-based encoding*, in which the synchrony of neural activations determines the bindings. This encoding, as Makino notes, is attractive from a connectionist perspective because synchrony-detectors are very elementary neural building blocks. Figure 2 illustrates how a synchrony-based encoding would work for a very simple example. The illustration seems to presume a localist view (i.e. one neuron, one meaning), but as Makino notes this is not required.

The synchrony-based encoding may seem a rather technical solution to a rather technical problem. However, neurological evidence suggests that oscillatory patterns play a key role in human cognition. So-called *brain waves* of various frequencies have shown to play an active part in human cognition. Brain

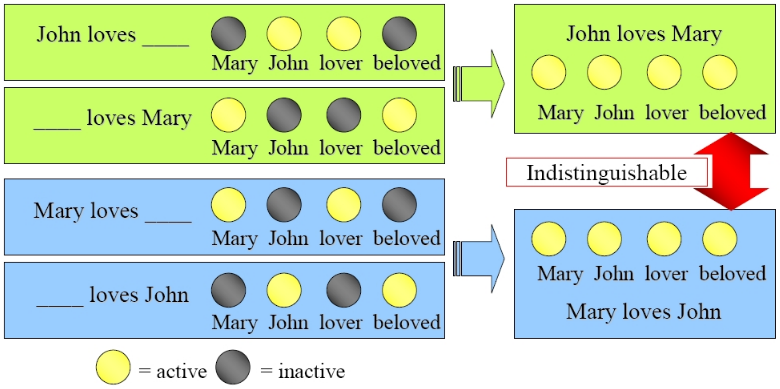


Figure 1: The binding problem in an additive representation of “John loves Mary”. Taken from [16].

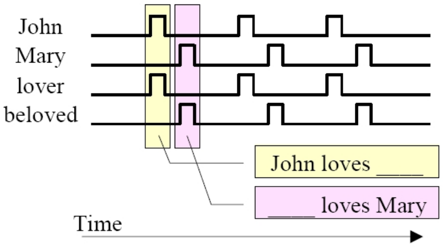


Figure 2: Synchrony-based encoding of “John loves Mary”. Taken from [16].

waves, i.e. oscillatory patterns visible in an EEG, occur when large amounts of neurons are firing in a synchronised way. When neurons desynchronise, the band power of a particular frequency decreases.

An overview of the role of theta (4-7 Hz) and alpha (8-12 Hz) oscillations in cognition is given in [14]. For instance, synchronisation in the theta band is associated with episodic memory, desynchronisation in the upper alpha band reflects semantic memory performance, and desynchronisation in the lower alpha band reflects attention. In [21], theta waves are associated with working memory, but not with language processing in particular. Alpha waves are associated mainly with the semantic part of language processing. Results in [15] suggest that the beta frequency is linked to orthographic encoding of words. Interesting links between alpha desynchronisation and the P600 peak, occurring after detecting a syntactic violation, and between theta synchronisation and the N400 peak, occurring after detecting a semantic violation, have been reported in [5].

Although it is too early to associate a particular frequency band with language processing, the existence of such encodings in the brain combined with the work in [16] provide sufficient indications that a similar mechanism may be used for language processing. In this paper I will assume that a synchrony-based encoding is the implementational basis of the main syntactic operations of the language faculty.

3.2 Asymmetry in synchrony-based representations

An important property of natural languages is that many constructions are to a large degree asymmetrical. One prominent example of this is headedness, i.e. the fact that some words seem to determine the behaviour of the constituent they are part of, whereas others do not. For instance, the verb “sleep” can occur wherever a verb is required, but so can the constructions “sleep well”, “sleep in the bedroom” and “rarely sleep”. The part “sleep” dominates the larger constituents it is part of.

Whatever the precise implementation of asymmetry in the language faculty, fact is that any connectionist model would have to be able to deal with it. The simplest form of asymmetry is rather abstractly an ordered pair, $\langle x, y \rangle$, corresponding to the set $\{x, \{x, y\}\}$. The kind of relations the system can deal with, furthermore, should be at least generic enough to represent (i) linear order in the phonological part and (ii) tree-like structures in the interpretation system, encoding a nested predicate-argument structure. As we will see, tree-like structures are also sufficiently powerful to realize the actual derivational part.

A simple linear order can be represented in a synchrony-based encoding by ordering the phases of the items. In representing the linearly ordered sentence “John kissed the girl”, for instance, ‘John’, along with its eventual features, would be in the first phase, ‘kissed’ in the second, ‘the’ in the third and ‘girl’ in the last. The neural activation patterns would be cycling in that specific order.

Things become more complex when instead of a linear order, a tree-like structure must be represented. Consider the (rather ad hoc) predicate-argument structure sleeps(the(man),in(the(garden))). Representing such a structure would require the item 'sleeps' to be succeeded in phase both by 'in' and by 'the', since these are the arguments of 'sleeps'. The item 'in' must be in the same phase as 'the' (otherwhise the one would preceed the other, representing an asymmetrical relation). But in the synchronisation based encoding, this would unjustifiably indicate that a binding exists between 'in' and 'the'. So, temporal order is insufficient to deal with tree-like asymmetries.

Instead, let us regard these relations as a kind of binding (which seems natural enough), but one which leaves a residual item in its own phase to denote the asymmetry. For example, 'the boy' with the determiner dominating the noun, would be represented as in figure 4. This kind of encoding is very similar to the set-notation of an ordered pair (recall that $\langle x, y \rangle = \{x, \{x, y\}\}$), and it is the minimal representation of tree-like relations possible when our only assumption is that our encodings are synchrony-based.

A more complex expression, like the predicate-argument structure given before, would be represented as in figure 5. The precise order of the phases does not matter for the representation. Regardless of the order of the phases, the asymmetry can be retained from the connectionist representation as long as one lexical item is distinguishably the root of the tree. An explicit starting symbol is no longer necessary. The loneliness of 'sleeps' in its phase represents the fact that this is where the hierarchy of relations begins.

4 Minimalist considerations

In this section I will discuss some key concepts of the Minimalist Program. I discuss feature checking, argue that the notion of constituents should be dropped, and briefly summarize the constituent-free approach of, among others, Collins ([4]) and Bowers ([3]).

4.1 Features

Theories in the Minimalist Program assume that rather than a set of grammatical rules, as traditionally used in context free grammar descriptions, our language faculty makes use of one very simple operation, which may be Merge, Unify, FormRel or something else, that is guided by a lexicon full of features. How plausible is this lexicalised, feature-based approach?

From an evolutionary perspective, it is easier to think of a selection process through which a lexicon is slowly enriched with more and more features, than to think of a process giving rise to a centralised, top-down system specialised in grammatical rules. Also from a computer science perspective, the more we can rely on self-organisational properties of the neural network, the easier it is to implement. Through the use of features rather than grammatical rules we can

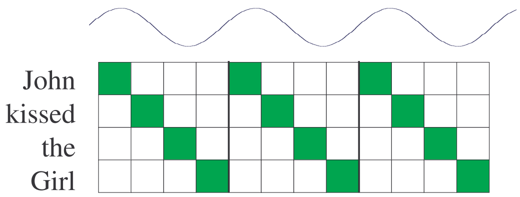


Figure 3: The simplest possible representation of linear order in a synchrony-based encoding. However, for tree-like structures (i.e. non-linear order) this manner of representing the asymmetries is insufficient. Here, three oscillations are shown. Columns represent phases. Coloured (green) cells represent activity.

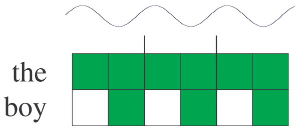


Figure 4: A synchrony-based representation of the asymmetry in ‘the boy’. Note that the order of the phases does not really matter.

throw away an important top-down part currently present in most connectionist models for language processing (for a detailed overview see [20]).

An application of information theory to article ommission shows that the information-theoretical notion of entropy is a very reliable measure of the computational complexity of different articles ([13]). The informational value of an article is calculated based on its features, like gender and definiteness. It seems very difficult indeed to come up with an equally reliable measure that does not presume the existence of such features.

Furthermore, an fMRI study ([8]) concerning artificial language learning supports the dominant idea in the Minimalist Program that constructions are built largely based on lexical features. To sum up, the minimalist reliance on features makes sense from a computer science perspective and complies with neuroscientific evidence.

After firing (depolarisation), neurons need to repolarise for a short interval before they can fire again. The maximal firing rate of neurons varies mostly between 10 and 100 Hz (although the 'fastest' neurons can fire up to several hundred times per second ([25])). Thus, for instance, in the alpha frequency range (around 10 Hz), most neurons could occupy somewhere between one and ten phases (i.e. fire one to ten times per oscillation).

Since all operations in a synchrony-based encoding involve cyclic patterns, a firing neuron can influence its own behaviour in the succeeding phases and even in the next oscillation. Activation of a neuron in one phase will inhibit its activation in the next phases. In a connectionist network relying on self-organisation, this inhibitory link across multiple phases is a key mechanism. If the satisfaction of uninterpretable features corresponds simply to the inhibition of their connectionist correlate, a limit on firing frequency may be responsible for this. More concretely, if two lexical items are active with the same feature, the feature activation for the first item will inhibit the activation of the feature for the second item. I will come back to this in more detail.

4.2 The burden of constituents

Within the Minimalist Program, the majority of linguistic phenomena are explained in terms of *constituents*. Following the more classical theories, a constituent is considered to inherit features from its head. Within the Minimalist Program, it is argued that this kind of redundancy is a computationally optimal implementation: it prevents the language faculty from having to traverse all the elements of a constituent each time in order to determine its behaviour ([12], [6]).

A solid implementation of ordered pairs of lexical items is not enough to represent a constituent structure. Consider again figure 5. Constituents are implicitly there, encoded as chains of relations. For example, the constituent 'sleeps in the garden' is encoded in figure 5 as the chain $(\langle sleeps, in\rangle, \langle in, the\rangle,$ $\langle the, garden\rangle)$. However, if this is all there is to representing constituents, then a constituent cannot form a relation with other items or constituents. After all, how would a connectionist system distinguish between relating to a lexical item and relating to a chain of lexical items?

A simple solution might be to *label* chains of relations. This corresponds to the minimalist view on constituents as labels for computational efficiency. If every chain of relations is preceded by a unique helper item to denote that a new constituent begins there, then relations between constituents could simply be relations between these helper items. However, there are at least three objections to this encoding. First, the system would require an infinite number of labels (or at least a whole lot of them), and equally many additional phases. Second, what would trigger the insertion of a label? Would they be added each time a new relation between lexical items is formed? Third, what is the nature of these helper items? Do they come from the lexicon, like words? Or do they reside in a different population of neurons, taking part in the same oscillatory activation pattern but in a different area of the brain?

A connectionist implementation of constituents would require a lot of extra assumptions to be made. If we wish to stick to our simple synchrony-based encoding model, then we might want to call into question the mainstream Minimalist view on constituents.

4.3 Syntactic relations

An interesting constituent-free alternative is the *syntactic relations* approach of, among others, Collins ([4]) and Bowers ([3]). They claim that constituents do not really exist (only implicitly, like the chains of relations mentioned before), and ground everything in a small number of simple relations that hold between lexical items only.

In this approach, the operation Merge is replaced with a much simpler operation, FormRel, which takes two lexical items and creates an ordered pair. Recall that, for a constituent representation, larger constructions should also be able to take part in relations. Here only lexical items can take part in relations. Lexical items form relations based on subcategorisation and selection conditions, which can be regarded as uninterpretable features (in this respect it actually sticks rather closely to the mainstream Minimalist Program). A determiner has a subcategorisation feature [N?] (i.e. it wants to relate to a noun), a verb has a selection feature [D?] (i.e. it wants to relate to the subject's determiner).

The mechanism Collins proposes works according to a *Locus Principle* (4). Together with the assumption that lexical items need to be satisfied themselves before they are allowed to satisfy another item's subcategorisation or selection condition, this imposes an order on the treatment of lexical items.

(4) **The Locus Principle** Suppose a lexical item λ, called the Locus, containing unsatisfied selection and subcategorisation features, is selected from a lexical array. Then all the subcategorisation conditions and selectional requirements of λ must be satisfied before a new lexical item can be selected as the Locus.

Every time a new relation is formed, it is immediately passed to spell-out (phonological form) and interpretation (logical form). This *immediate spell-out* and *immediate interpretation* are crucial parts of the mechanism. The different relations (subcategorisation and selection) behave differently under spell-out or interpretation. Leaving the details aside, subcategorisation relations are spelled out in order, whereas selection relations are spelled out inversely.

Furthermore, certain properties of spell-out are introduced in order to account for word order. For example, spelling out a pair of lexical items, where the first item is a phonetically null element (such as a light verb), leads to a phonological realisation of the second member of the pair instead. Interestingly, immediate spell-out and immediate interpretation are plausible also from a cognitive perspective, where incremental, continuous mechanisms are preferred over batch processes involving large stockpiled chunks to be transfered at once.

As I do not wish to go into the syntactic relations approach in much more detail, I will just show the derivation of a simple sentence, ignoring logical form, and refer to [3] and [4] for further details. A sentence "the boys read the books" is derived as in 5.

(5) {the, boys, read, the, books, v}

A FormRel(the, books). PF: the-books.

B FormRel(read,the). PF: the-books-read.

C FormRel(the, boys). PF: the-boys.

D FormRel(v, read). PF: read-the-books-⟨read⟩.

E FormRel(v, the). PF: the-boys-read-the-books-⟨read⟩.

There seems to be some indeterminacy here regarding the order of E and D. If E would come before D, the resulting phonetic form would be "read-the-boys-⟨read⟩-the-books", which is incorrect. The reason D comes before E is a constraint on interpretation, which I will omit here (but see chapter 3 in [3]). In the next section I will use this example again.

5 Synthesis

In this section, based on the previous sections and on neuroscientific data, I will propose an abstract model for language production. There are some ways in which my model will differ from the model in [16], although the base mechanism, synchrony-based encoding, is the same. First, the model employed by Makino is a model for language comprehension, whereas the model I will propose is one mainly for production. Theories in the Minimalist Program usually deal with production only, claiming that comprehension is dealt with by some kind of a parser. I think that the connectionist mechanisms I will describe can be used for parsing as well, but I will leave that for future work.

Second, Makino's model relies on global, top-down phase arbitration, i.e. the assignment of lexical items to unused phases. I will not make any particular claim concerning the phase arbitration mechanism. A local, self-organising system would be attractive from an evolutionary perspective - if phase arbitration can emerge from self-organisation - a key mechanism in our brain - why would evolution have installed a top-down manager? Also, the fundamental argument proposed by Makino for installing a top-down phase arbiter (see [16] page 58) is an argument only against strictly local systems, i.e. relying only on the neural correlates of the lexical items themselves, without any other neurons to assist. The argument loses most of its strength when some (local) assisting neurons are allowed.

Third, Makino's model makes use of explicitly implemented grammatical rules (hetero-associative rules of the form "if D and N are active together, NP will be active in the next oscillation"). The model presented here does not, and instead resolves everything by features and self-organisation, compliant with the Minimalist Program.

5.1 Assembling the numeration

An event-related fMRI study ([2]) strongly suggests that the lexicon behaves as a content-addressable memory system, i.e. lexical lookup is based on meaning. It furthermore reveals that a distributed set of left-hemispheric regions shows stronger activation for words than for non-words. A very promising enquiry is given in [17]. Through a series of fMRI studies, a method is developed to successfully predict the brain activity associated with a noun, by taking the weighted sum of brain activities associated with related verbs. Neural activity associated with verbs (and consequently nouns) occurs particularly in the sensori-motor areas. A different study ([1]) shows that lexical semantic information is stored and retrieved by means of oscillations in the theta frequency range (4-7Hz), and backs up [17] in the claim that sensori-motor brain areas are involved in a highly function-dependent way.

These results may suggest that the lexicon itself operates in this distributed, function-dependent way. However, they seem to concern mainly the semantic part of lexical lookup. Research done on brain activation effects of regular versus irregular verbs localises the lexicon in the left temporal regions and displays a dissociation between lexical lookup, involving the left temporal regions, and the grammatico-morphological system, involving left frontal regions ([23]). Thus, we see two views on the lexicon: the lexicon as a distributed and mainly semantic system, and the lexicon as a more localised, less semantic system. Which of these 'lexicons' correspond to the Minimalist notion of a lexicon?

In the Minimalist Program, the lexicon is (presumably) a compact collection of sets of simple features, rather than a rich and widely distributed collection covering all the aspects of meaning a lexical item can possibly have. The latter is more accurately called 'semantic memory' and should (at least for linguists) not be equated with the lexicon. I see the activation patterns found in [2], [17] and [1] as having to do with thoughts and ideas that would exist even if we did not have language. The success of corpus analysis in [17] is a result of the corpus reflecting the structure of the world very well, rather than an indication of some relation between these activations and something linguistic like a lexicon. The findings in [23] seem to target the lexicon more accurately.

Somehow, the vague, distributed representation of a thought can be reliably linked to very specific items in a numeration, where compact and somewhat isolated representations of words as sets of features, phonological and semantical, seem a prerequisite for the language faculty to be able to do its job effectively and efficiently. Importantly, as mentioned in section 2, the fuzzy meaning representation that precedes numeration assembly and sentence production (and which succeeds sentence perception) should be distinguished from the logical form that allows to be reasoned with and functions as a check on the convergence of the language faculty.

Somehow, based on a thought or idea, which is likely to exist in a synchrony-based encoding in the theta band ([1]), lexical items are activated in a system in the parieto-temporal area, each in a different phase, possibly in the alpha frequency range (8-12Hz) - although it is too early to claim this with much certainty. Such an acceleration from theta to alpha would make sense from a computational perspective. The higher the frequency, the faster information can be processed. A higher frequency encoding is only reliable when the neurons are sufficiently interconnected and relatively close together, quite unlike activations in the theta band.

It is likely that the activation of lexical items is paired with an activation of the appropriate features, either directly or with a short delay via auto-associative rules (e.g. ‘the’ excites [D]). Lexical items that do not carry real meaning, like the expletive ‘there’ or the light verb, cannot originate from the original meaning representation. It is assumed here that such items are more easily activated during the derivation, or are simply always there. This account of light verbs as being more easily activated fits with an interesting case reported in [24], of a woman with damage to her left temporal and parietal lobes, who tended to use a lot of nonspecific verbs (i.e. light verbs) to circumvent her verb retrieval problems (e.g. “using a saw” instead “sawing”).

5.2 Forming relations

When in the lexicon the appropriate items have been activated and assigned a phase, the brain is ready to add some structure to the system. Recall that this happens based on features. Recall also the (variable) maximum firing rate of neurons and the inhibition of a pattern in phases succeeding the pattern’s activation, as discussed in section 4.

Consider the phase configuration in figure 6, and the way it changes when a relation is established. Words like ‘the’, seeking an N, and ‘boy’, being of category N, share the same feature [N], the difference being the interpretability. If our connectionist model is in any way minimalist, the neural correlates of an interpretable feature [N] and an uninterpretable feature [N?] should be the same. What, then, causes the difference between interpretable ([N] of ‘boy’) and uninterpretable ([N?] of ‘the’) features? Why does the uninterpretable [N?]

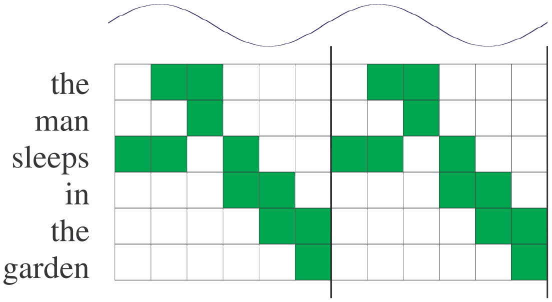


Figure 5: A synchrony-based representation of the simple predicate-argument structure sleeps(the(man),in(the(garden))).

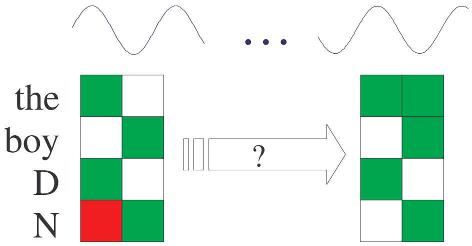


Figure 6: The formation of a relation between 'the' and 'boy'. The uninterpretable feature [N?], denoted by the red cell in the first phase, disappears, while a duplicate activation 'the' in the second phase appears. But how?

disappear in this transition rather than [N]? And why, as the figure shows, is 'the' activated together with 'boy' after their features collide?

I propose that the principles in 6 provide a very simple mechanism to establish a relation. The four principles have been formulated in a rather localist fashion, seemingly assuming that a particular uninterpretable feature corresponds to a particular neuron, with one outgoing excitatory connection. As before, it must be emphasised that the neural correlate of a feature is a certain fuzzy set of similar activations, rather than a single firing neuron.

(6) (i) The neural correlates of interpretable and uninterpretable features are the same.

(ii) Categorical features, both interpretable and uninterpretable, have an inhibitory effect on all other categorical features.

(iii) Uninterpretable features have an excitatory effect on the (rest of the) lexical item, whereas interpretable features are themselves excited by the (rest of the) lexical item.

(iv) The neural correlates of categorical features have a lower maximal firing frequency than the other parts of a lexical item (e.g. phonetic features).

More concretely, words excite their categories and are excited by their selection and subcategorisation conditions. Thus, 'the' excites [D] but is excited by [N?], and 'boy' excites [N]. Furthermore, both 'the' and 'boy' have a higher maximal firing rate than their categorical features. The latter is no particularly exotic assumption, since the firing frequencies of different neurons may differ as much as a factor of ten ([25]). Additionally, each phase ideally only contains one active categorical feature (either interpretable or uninterpretable, e.g. [D], [N?]), so multiple active categorical features in the same phase will inhibit eachother a bit.

According to these principles the relation between 'the' and 'boy' can be established as in figure 5.2. First [N?] fades, because the firing rate of [N] is just a bit too low to keep up with two activations in every oscillation. Therefore, one of the [N] activations has to die out. Since 'the' does not excite [N] whereas 'boy' does, it is the latter that survives. Then, since [N] excites 'the', 'the' is gradually coactivated with [N]. The multiple occurrences of 'the' cause no trouble, because according to the fourth principle in 6 it has a much higher maximal firing rate than features. One might wonder why, given that neurons fire in an all-or-nothing fashion, such gradual transitions are possible. However, recall that one coloured cell in the figure does not correspond to one firing neuron, but rather to a certain activation pattern.

Admittedly, of the four principles proposed, none can be justified from a neuroscientific perspective, simply because science *isn't quite there yet.* However, the resulting mechanism is very straightforward, following from a minimal set of assumptions, and complies with the minimalist philosophy. It relies completely on self-organisation and requires no top-down arbiter. In the last section I will come back to this, and discuss the scientific value of the four principles.

5.3 Derivational order

We have not yet seen any use of the second principle in 6, that categorical features, both interpretable and uninterpretable, have an inhibitory effect on all other categorical features. This principle complies perfectly with the fact that in a convergent phase configuration (such as the last oscillation in figure 5.2), each phase holds the activity of only one categorical feature. But what is the purpose?

Recall that the Locus Principle (4), together with the criterion that lexical items need to be satisfied before they can satisfy, imposes an order on the treatment of lexical items. Both criteria still have to be translated into connectionist mechanisms. If a connectionist network is in any way close to optimal, then the selection of an appropriate Locus would not be done at random. Rather, an inventarisation must be made of the available lexical items that are already satisfied. A lexical item must then only be selected as a Locus when all its subcategorisation and selection features have proper (satisfied) counterparts to match on. When this is the case, all the matching can happen instantaneously. Because the locus is, in this sense, only a locus for an instant, the Locus Principle is rendered obsolete (or at least it can be left implicit).

The criterion that lexical items need to be satisfied before they can satisfy other lexical items, is incorporated by the second principle in 6. If, as proposed, categorical features slightly inhibit eachother when in the same phase, the presence of an uninterpretable feature would inhibit the categorical label (e.g. for 'the' the presence of [N?] would inhibit [D]) or vice versa. If the uninterpretable feature wins, so to speak, and inhibits the categorical label, a relation will not be formed - and that is what we wish to achieve.

Luckily, the situation is already such that the uninterpretable feature will win. To see this, consider figure 5.3, where the noun is for the sake of clarity accompanied by an unsatisfied feature [X?]. Since the label [N] in the second phase was already less powerful due to the activation of [N?] the phase before, [X?] will win in phase 2, thus inhibiting [N] and blocking the relation.

Following the connectionist apparatus I have explained, the Appendix (end of document) contains a complete derivation of "the boys read the books" (see also the derivation in 5). Six different phases are occupied. As before, coloured (green and red) cells represent activity, where red cells represent uninterpretable features (i.e. outgoing excitatory connections). The formation of a relation is shown in one step rather than as a gradual transition as in figure 5.2.

6 Final remarks

Based on insights from computer science and neuroscience I have formulated in connectionist terms the main operations of a promising branch of the Minimalist Program. First I argued why the minimalist version of the T-model is a good starting point. Then I elucidated my choice for a synchrony-based

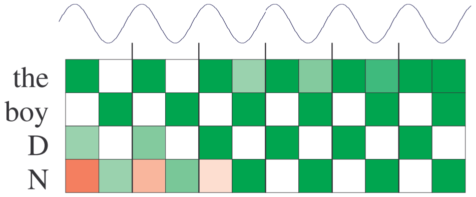


Figure 7: The gradual formation of a relation between ‘the’ and ‘boy’. The uninterpretable feature [N?], denoted by the red cell in the first phase, slowly fades, because its maximal firing rate is too low and, unlike in the second phase, it is not explicitly excited in the first phase.

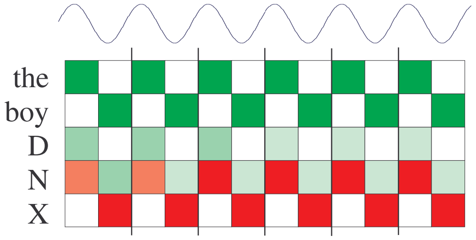


Figure 8: The gradual formation of a relation between ‘the’ and ‘boy’ is disturbed when the latter still contains some unsatisfied feature [X?] (the red cell in the second phase). This desired effect can be achieved by assuming that categorical features inhibit eachother.

encoding and explained how, in a connectionist system, asymmetries could be represented. After that, I captured some key notions of the Minimalist Program in connectionist terms and decided to employ a constituent-free approach. Finally I brought everything together, along with psychological data, into a connectionist model for language production.

The core of the model is a small set of principles (7).

(7)
(i) The neural correlates of interpretable and uninterpretable features are the same.
(ii) Categorical features, both interpretable and uninterpretable, have an inhibitory effect on all other categorical features.
(iii) Uninterpretable features have an excitatory effect on the (rest of the) lexical item, whereas interpretable features are themselves excited by the (rest of the) lexical item.
(iv) The neural correlates of categorical features have a lower maximal firing frequency than the other parts of a lexical item (e.g. phonetic features).

What is their empirical and theoretical value? The first principle can be motivated from an evolutionary or minimalist perspective: if interpretable and uninterpretable features can be dealt with in this way, why would natural selection, or our language acquisition mechanism, increase the number of representations required? The simple principle that neural correlates of interpretable and uninterpretable features are the same (although there may be a difference in connectivity) implies that in natural language *there are no lexical items that subcategorise or select their own category.* Fortunately, this seems to be the case.

The second principle, invoking a kind of competition between (both interpretable and uninterpretable) categorical features, makes sense as well. A lexical item is typically of only one category, not two. Besides, a convergent derivation typically ends in a phase configuration with one categorical label per lexical item. It is in that sense the optimum.

The third provides a temporary solution that needs to be taken for granted until neuroscience either confirms or disconfirms it. However, since interpretable and uninterpretable features have to differ at least in some sense, it is likely that the difference is in the connectivity patterns. I have simply made a particular choice here, in order to show that a connectionist model of minimalist grammar is really possible. It is likely that connectivity differences, other than the one proposed here, exist which can account for the difference between interpretable and uninterpretable features.

Likewise, the fourth principle awaits confirmation or disconfirmation from neuroscience. Categorical features have a cognitive function that is much different from phonological or semantic features, so it can be expected that their neural correlates differ. Given that large differences in maximal frequencies have already been recorded, why wouldn't our brain use this property? Again, I do not deny that there are valid alternatives. I have simply made a choice here in order to prove the existence of at least one valid connectionist mechanism for minimalist grammar.

In this paper I did not explain how the resulting relations are transmitted to the phonological and interpretational interfaces, and what happens there. The abstract functions translating the set of relations into either phonological or logical form, as formulated in [3], are very simple and do not seem particularly difficult to implement in a connectionist network. This is something I will leave for future work.

Another opportunity for future work is to actually implement the system. The mechanisms proposed in this paper have the advantage over existing approaches of being very reliant on the self-organisational properties of the network, and making use of very simple connectivity conditions rather than grammatical rules imposed in a top-down fashion. It would also be the first connectionist system to explicitly implement theories from the Minimalist Program.

The relative ease in which concepts from the Minimalist Program (features and relations in particular) can be formulated in connectionist terms should be a welcome finding for linguists working on the program. They should also see the merits of the interdisciplinary approach employed in this paper. My decision to take on a constituent-free approach did not come out of thin air, but rests on implementational considerations. Similarly, many ongoing debates in the Minimalist Program could be tackled from an interdisciplinary standpoint.

Finally, some weight rests on the neuroscientist's shoulders to verify or disconfirm the third and fourth principles formulated here, and to find out which frequency band in particular may be associated with the relation-forming part of the language faculty. Currently, promising research is being conducted by Klimesch and his colleagues. Also the work of Davidson and Indefrey on the relation between P600/N400 peaks and brain waves may shed light on some open issues.

References

[1] Marcel C.M. Bastiaansen, *et al.*. 2008. *I see what you mean: Theta power increases are involved in the retrieval of lexical semantic information.* Brain and Language, 106, pp. 15-28.

[2] J.R. Binder, *et al.*. 2003. *Neural Correlates of Lexical Access during Visual Word Recognition.* Journal of Cognitive Neuroscience, Vol. 15, no. 3, pp. 372-393.

[3] John Bowers. 2001. *Syntactic Relations.* Manuscript.

[4] Chris Collins. 2002. *Eliminating labels.* In Epstein & Seely, eds., Derivation and Explanation in the Minimalist Program. Blackwell publishing.

[5] D.J. Davidson & P. Indefrey. 2007. *An inverse relation between event-related and time-frequency violation responses in sentence processing.* Brain Research 1158, pp. 81-92.

[6] Catarina Donati. 2006. *Labels and merge*. Handout for University of Cyprus.

[7] Jos L.G. Escribano. 2005. *Semantocentric Minimalist Grammar*. Atlantis 27.2, pp. 57-74.

[8] Christian Forkstam, *et al.*. 2006. *Neural correlates of artificial syntactic structure classification*. NeuroImage 32, pp. 956-967.

[9] Y. Grodzinsky. 1995. *Trace deletion, theta-roles, and cognitive strategies*. Brain & Language 51, pp. 467-497.

[10] James Henderson. 1994. *Connectionist Syntactic Parsing Using Temporal Variable Binding*. Journal of Psycholinguistic Research, Vol. 23, no. 5.

[11] Norbert Hornstein. 2001. *Move! A Minimalist Theory of Construal*. Blackwell Publishers.

[12] Norbert Hornstein, Jairo Nunes, and Kleanthes K. Grohmann. 2005. *Understanding Minimalism*. Cambridge: Cambridge University Press.

[13] Joke de Lange. 2008. *Article Omission in Headlines and Child Language: A Processing Approach*. Thesis. LOT, Utrecht.

[14] Wolfgang Klimesch. 1999. *EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis*. Brain Research Reviews 29, pp. 169-195.

[15] Wolfgang Klimesch, *et al.*. 2001. *Alpha and beta band power changes in normal and dyslexic children*. Clinical Neurophysiology, pp. 1186-1195.

[16] Takaki Makino. 2002. *A Pulsed Neural Network for Language Understanding: Discrete-Event Simulation of a Short-Term Memory Mechanism and Sentence Understanding*. Ph.D. dissertation, Tokyo University.

[17] Tom M. Mitchell, *et al.*. 2008. *Predicting Human Brain Activity Associated with the Meanings of Nouns*. Science Vol. 320. no. 5880, pp. 1191-1195.

[18] John O'Keefe, Michael L. Recce. 2008. *Phase relationship between hippocampal place units and the EEG theta rhythm*. Hippocampus Vol.3 No.3, pp. 317-330.

[19] Andrew Radford. 2006. *Minimalist Syntax Revisited*. http://courses.essex.ac.uk/lg/lg514 .

[20] Douglas L. T. Rohde & David C. Plaut. 2003. *Connectionist Models of Language Processing*.

[21] D. Rhm, *et al.*. 2001. *The role of theta and alpha oscillations for language comprehension in the human electroencephalogram*. Neuroscience Letters 310, pp. 137-140.

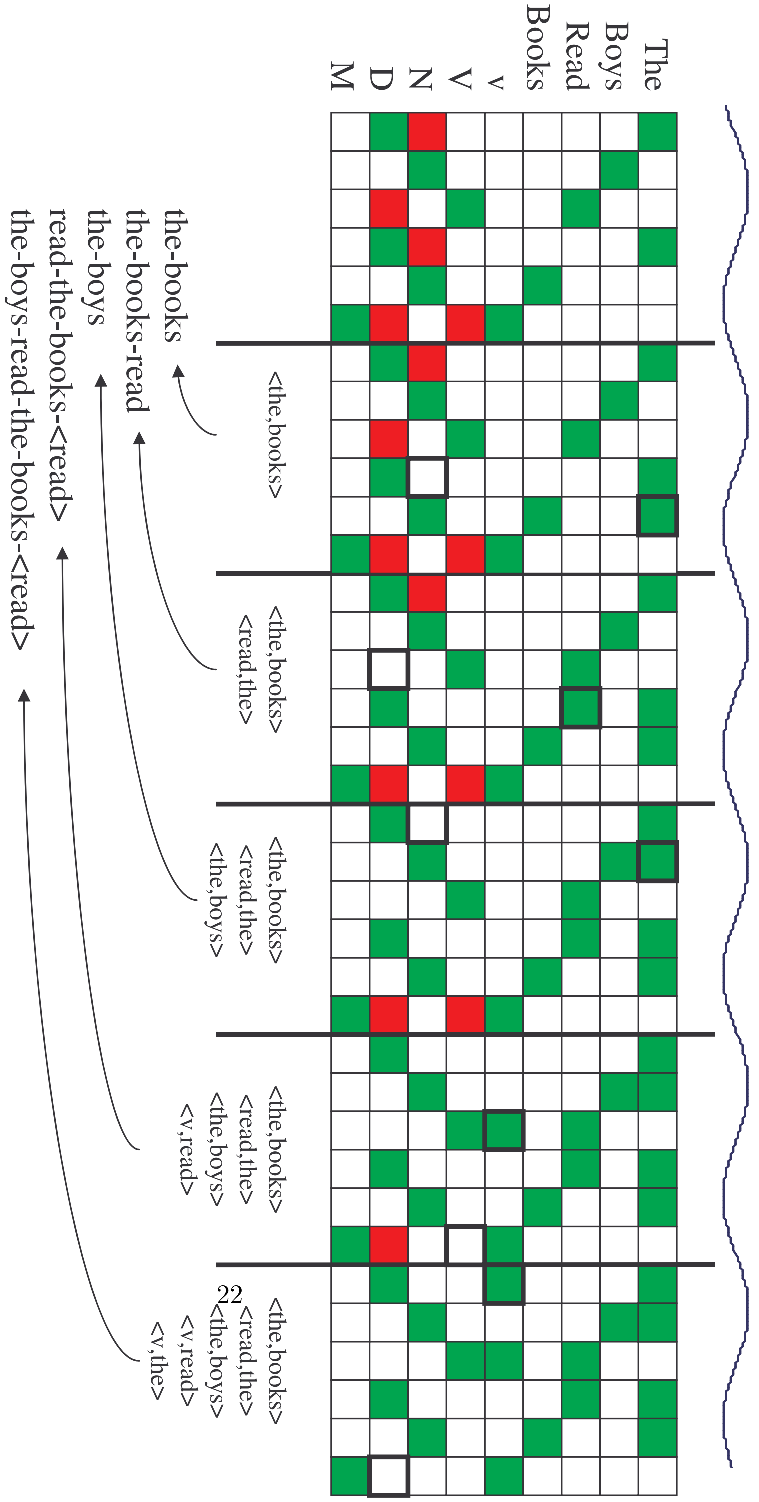
[22] Thomas Suddendorf and Janie Busby. 2003. *Mental time travel in animals?*. TRENDS in Cognitive Sciences Vol.7 No.9.

[23] M.T. Ullman, *et al.*. 2005. *Neural Correlates of Lexicon and Grammar: Evidence from the Production, Reading, and Judgment of Inflection in Aphasia*. Brain and Language, Vol. 93 no.2, pp. 185-238.

[24] Christina E. Wierenga, *et al.*. 2006. *Neural substrates of syntactic mapping treatment: An fMRI study of two cases.* Journal of the International Neuropsychological Society, Vol. 12, pp. 132-146.

[25] Frank Wijnen & Frans Verstraten. 2004. *Het brein te kijk.* Harcourt, Amsterdam.

7 Appendix


The
Boys
Read
Books
v
V
N
D
M
<the,books>
<the,books>
<read,the>
<the,books>
<read,the>
<the,boys>
<the,books>
<read,the>
<the,boys>
<v,read>
<the,books>
<read,the>
<the,boys>
<v,read>
<v,the>
the-books
the-books-read
the-boys
read-the-books-<read>
the-boys-read-the-books-<read>