2 Light and Deep Parsing: A Cognitive Model of Sentence Processing

Philippe Blache

Abstract

Humans process language quickly and efficiently, despite the complexity of the task. However, classical language-processing models do not account well for this feature. In particular, most of them are based on an incremental organization, in which the process is homogeneous and consists in building step-by-step a precise syntactic structure, from which an interpretation is calculated. In this chapter, we present evidence that contradicts this view, and show that language processing can be achieved at varying levels of precision. Often, processing remains shallow, leaving interpretation greatly underspecifie.

We propose a new language-processing architecture, involving two types of mechanisms. We show that, in most cases, shallow processing is sufficient and deep parsing is required only when faced with difficulty. The architecture we propose is based on an interdisciplinary perspective in which elements from linguistics, natural language processing, and psycholinguistics come into play.

2.1 Introduction

How humans process language quickly and efficiently remains largely unexplained. The main difficulty is that, although many disciplines (linguistics, psychology, computer science, and neuroscience) have addressed this question, it is difficult to describe language as a global system. Typically, no linguistic theory entirely explains how the different sources of linguistic information interact. Most theories, and then most descriptions, only capture partial phenomena, without providing a general framework bringing together prosody, pragmatics, syntax, semantics, etc. For this reason, many linguistic theories still consider language organization as modular: linguistic domains are studied and processed separately, their interaction is implemented at a later stage. As a consequence, the lack of a general theory of language, accounting for its different aspects, renders difficult the elaboration of a global processing architecture. This problem

has direct consequences for natural language processing: the classical architecture relies on different subtasks: segmenting, labeling, identifying the structures, interpreting, etc. This organization more or less strictly imposes a sequential view of language processing, considering in particular words as being the core of the system. Such a view does not account for the fact that language is based on complex objects, made of different and hetergeneous sources of information, interconnected at different levels, and which interpretation cannot always be done compositionally (each information domain transferring a subset of information to another).

Cognitive approaches to language processing (LP) face the same difficulties. Even more crucially than for linguistics, psycholinguistics models mainly rely on a sequential and modular organization. Language is usually considered to be strictly incremental, relying on a word-by-word processing, each word being integrated into a partial syntactic structure, starting from which an interpretation can be calculated. In this organization, the different steps used in classical natural language processing (NLP) architectures are implemented: segmentation, lexical access, categorization, parsing, interpretation, etc.

This perspective is also adopted in neurolinguistics, trying to identify in a spatial or in a temporal dimension the brain basis of LP. The question there consists in studying what parts of the brain are involved in LP and in what manner. What is interesting is that even though the different works focus on only one linguistic dimension (e.g., lexicon, lexical semantics, prosody, morphosyntax), they also show that they are strongly dependent on each other.

We propose in this paper an approach bringing closer the different types of knowledge about LP coming from these disciplines that makes it possible to draw a broader and more integrated architecture.

2.2 An Interdisciplinary View of Language Processing

This section gives an overview of language processing through different disciplines (linguistics, psycholinguistics, computational linguistics, and neurolinguistics). We show in particular that classical architectures (modular and serial) are now challenged, in particular when taking into account language as a whole, in its natural environment, opening the door to a more flexible approach.

2.2.1 A Classical View: Modular and Incremental LP

This section presents the main features of the classical modular architecture of LP, from the generative framework, that has influenced other disciplines.

Modularity: A view from linguistics: The classical generative architecture relies on a succession of different modules, each one specialized for a

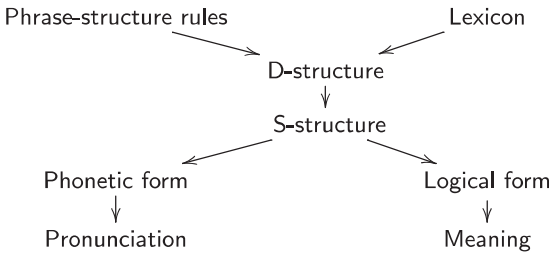


Figure 2.1 A classical generative architecture of language processing.

specific linguistic dimension. The Figure 2.1 illustrates such organization for the "Government and Binding" approach [Chomsky, 1981]. Starting from the lexicon and the rules (the set of local trees), the process consists of generating an underlying structure, subject to modifications and transformations that lead to another structure, closer to the surface form. From this structure, the phonological and logical forms are produced, making it possible to access the meaning:

This organization considers the different modules as not only separate but also sequential. Many linguistic theories propose a similar organization in which each domain produces a structure that is transferred to another one. One of the reasons for this is that linguistic theories are usually *syntactocentric*: all domains are considered in terms of their relation to the syntactic structure.

Module interaction: A view from computational linguistics: LP is taken from a specific perspective in NLP because of implementation constraints: LP is usually considered as a set of tasks, implementing the different modules in a serial manner. In this architecture, modules are synchronized, the input of one module being the output of the previous one. Up to now, no real answer to the question of the integration of the different sources of linguistic knowledge is given and their interaction is described in terms of specific synchronization rules [Jackendoff, 2007]. More precisely, even though many works have been done concerning the study of the interaction between the domains (e.g., prosody/syntax, syntax/semantics, etc.), solutions are proposed by giving the priority to one domain, usually syntax. For example, the compositional view of semantics [Werning et al., 2012] is implemented by the construction of a syntactic structure starting from which the interpretation can be calculated [Copestake et al., 2001]. The same kind of approach can be found for the prosody/syntax interface, in which prosodic information is integrated to the syntactic structure [Steedman, 2000]. As presented in the previous section, Modularity: A View from Linguistics, this is a syntactocentric organization, which induces an incremental and modular view of LP. As a consequence,

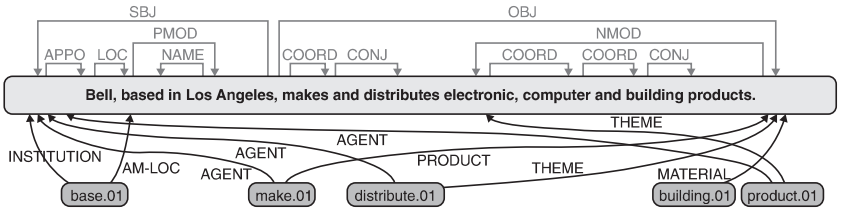


Figure 2.2 Output of the Stanford parser.

classical NLP architectures are organized around a series of subprocesses: tokenization, tagging, parsing, discourse organization, semantic interpretation. A parser builds complex structures (trees, feature-value structures, etc.) involving information at different levels, as shown in Figure 2.2, based on information produced by parsers such as the Stanford [de Marneffe and Manning, 2008], enriched with semantic information.

Other processes can be added to this general schema when studying audio (phonetics, prosody) or multimodality (gestures). Even though the question of parallelization in NLP is regularly addressed [Adriaens and Hahn, 1994; Jindal et al., 2013], the answer is usually given in terms of different parallel processes with meeting points, instead of an integrated view.

Incrementality: A view from computational psycholinguistics: In psycholinguistics, the LP classical architecture relies on the idea that processing is incremental, consisting in integrating each new word into a partial structure under construction [Fodor and Ferreira, 1998; Grodner and Gibson, 2005; Sturt and Lombardo, 2005; Keller, 2010; Altmann and Mirković, 2009; Rayner and Clifton, 2009]. In this approach, the basic units are considered to be the words: all information related to the lexical item is accessed when encountering a new word and is used to integrate the item into a partial syntactic structure (often called the *current partial phrase marker*). This operation consists in finding a site in the structure to which to attach the word. If this becomes difficult, the word is integrated where it least severely violates the grammar, following the "attach anyway" principle proposed by Fodor and Inoue [Fodor and Inoue, 1998].

This concept is also, syntactocentric, organizing information around the syntactic structure. Moreover, it is essentially sequential, in the sense that lexical information is processed before syntax, from which interpretation becomes possible. In other words, it is a *modular syntax-first* concept, supported by several classical works [Fodor, 1983; Frazier and Fodor, 1978] and still at work in many psycholinguistics models.

In terms of interpretation or meaning access, these approaches are also basically compositional, whether they are serial or parallel [Gibson, 2000]. In serial

models, the language processor initially computes only one of the possible interpretations [Fodor and Ferreira, 1998; Gorrell, 1995]. When this interpretation becomes difficult or even impossible, another interpretation is built. In parallel models, all possible interpretations are computed at once, the analysis with the greatest support being chosen over its competitors [MacDonald et al., 1994; Marslen-Wilson and Tyler, 1980; Spivey and Tanenhaus, 1998]. These two options both rely on an incremental view: interpretation is built at each new word, on the basis of a word-by-word syntactic and semantic analysis. Many issues are raised with these models. First, they both consider incrementality in a strict manner: an interpretation covering all the words at a given position is built, even it if builds an ill-formed structure [Fodor and Ferreira, 1998]. Moreover, the question of memory remains an issue in both cases: What elements are to be stored, under what form, requiring what capacity?

Brain basis of a modular architecture: A view from neurolinguistics: The study of the physiological and brain basis of language processing also leads to different LP architecture. Among the possible investigation techniques for the exploration of LP neural correlates, electrophysiological studies are frequently used. These experiments focus on the study of event-related potentials (ERPs), which are potential changes measurable from the scalp and which reflect the activity of a specific neural process [Luck, 2005]. LP modulates a number of ERP components, located between 100 and 600 ms after the stimulus (for example, reading a word). Figure 2.3 shows the main positive (P1, P2, P3) and negative (N1, N2) deflections that can be elicited by language processing.

Many effects have been explored, related to different linguistic domains (prosody, morphology, syntax, semantics in particular) [Kutas et al., 2006; Kaan, 2007]. Even though no electric component is strictly related with one domain, we can find in the literature. Although no electrical component relates

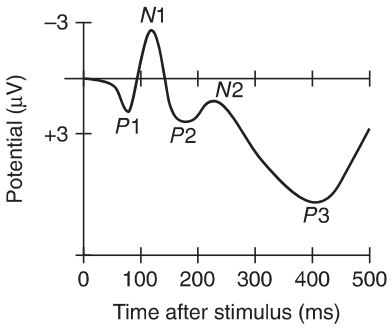


Figure 2.3 The main ERP components in language processing.

strictly to one domain, some early effects are reported for speech perception at 100 ms after the stimulus, word production at approximately 200 ms, semantics at approximately 400 ms, and syntax at approximately 600 ms. This is only a very rough picture, and all the observed effect depends on the linguistic material, in particular the amount of information coming from each domain.

Several works in neurolinguistics support a modular and serial view of language processing. Typically, the three-phases model [Friederici, 2002; Friederici, 2011] proposes an organization into three different steps, after an initial phase of acoustic-phonological analysis:

Phase 1: Local phrase structure is built on the basis of word category information.
Phase 2: Syntactic and semantic relations (verb/argument, thematic role assignment).
Phase 3: Integration of the different information types and interpretation.

This organization can be completed, in an auditory comprehension model, by adding interaction of prosody at each of these stages. Different language ERP components are in relation with these phases (see Figure 2.4).

- Early left anterior negativity (ELAN, 120–200 ms): Initial syntactic structure-building processes.
- Centroparietal negativity (N400, 300–500 ms): Semantic processes.

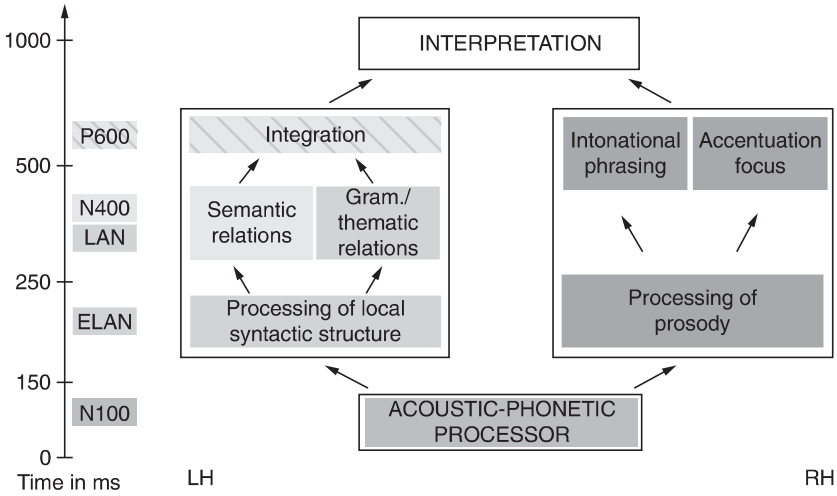


Figure 2.4 General organization of the three-phases model [Friederici, 2011]. ELAN, early left anterior negativity; LAN, left anterior negativity; N100, _____; N400, centroparietal negativity; P600, late centroparietal positivity.

- Left anterior negativity (LAN, 300–500 ms): Grammatical relation between arguments and verb, assignment of thematic relations.
- Late centroparietal positivity (P600): Late syntactic processes.

This model is syntax-first, and consists in building a syntactic structure from which interpretation can be done.

2.2.2 An Integrated View of LP: Constructions

The classical modular view is now challenged by recent theories considering that no specific domain is at the center of the architecture, and the processing is described in terms of interaction between them. Instead of being serial, processes are considered parallel, as described in [Jackendoff, 2007].

Construction grammar [Fillmore, 1988; Goldberg, 1995] is one of those theories proposing an alternative organization. Here, no structure is predefined, and no domain need to be described before another; all linguistic phenomena are described thanks to a set of interacting properties. As presented in Goldberg, 2003, constructions are *form and meaning* pairings: a set of properties makes it possible to characterize a construction, the meaning of which is accessed directly. Constructions can be of different types, as presented in the following examples:

- Ditransitive construction: Subj V Obj1 Obj2: *She gave him a kiss*.
- Covariational conditional construction: The *X*er the *Y*er: *The more I read the less I understand*.
- Idiomatic constructions: *Kick the bucket*; *to put all eggs in one basket*.

What is important with constructions is the fact that they are defined on the basis of different properties, possibly coming from different linguistic domains, without requiring preliminary complete analysis of each of these domains. For example, a syntactic tree is of no use in understanding an idiomatic construction. In such cases, instead of being built compositionally, the meaning of the construction is accessed directly.

This means that two types of mechanisms coexist in LP: one based on a compositional architecture, and another relying on direct access. In the first case, the architecture consists in analyzing all sources of information and their interactions. Each source or combination of sources contains a partial meaning, and their composition leads to a complete interpretation of the message. In the direct access case, the different properties, instead of bearing part of the meaning, play the role of cues in identifying the construction. The recognition of such pattern leads to a direct interpretation, without any composition. In some cases, only a few properties makes it possible to recognize and to interpret an entire construction.

2.2.3 Different Levels of Processing: LP Is Often Shallow

A flexible model of LP, the *good-enough theory* [Ferreira and Patson, 2007], has been proposed. It is based on the observation that interpretation of complex material is often shallow and incomplete. For example, Swets and colleagues [Swets et al., 2008] showed in a self-paced reading study that when participants expect superficial comprehension questions, ambiguous sentences are read faster, showing that no precise attachment resolution is done, leading to underspecified semantic representations. In this case, it is suggested that the ambiguity is not resolved, explaining the facilitation effect.

Several experiments confirm this observation that sentence comprehension can be quite shallow. For example, thematic role assignment can be subject to a simple heuristic: the first NP is the agent, the second the entity affected by the action. The use of such a heuristic has been exhibited by simple experiments showing that the interpretation of sentences contradicting this heuristic leads more often to misinterpretations than those satisfying it. These observation tend to show that, in several cases, no compositional processing is at work. Instead, as Ferreira and Patson explained, "the comprehension system tries to construct interpretations over small numbers of adjacent words whenever possible and can be lazy about computing a more global structure and meaning." The building of a complete and precise interpretation is often delayed or even never done, replaced by the identification of "*islands*," from which a general interpretation can be approximated.

Note that this theory contradicts several classical language-processing models. In this case, there is no systematic instantiation of thematic roles, at least in a first stage, contrary to what is required in generative theories. This is contradicts some psycholinguistic diffculty models [Gibson, 2000], which stipulate that NP without thematic roles (as well as unassigned thematic roles) impose a burden on working memory. Conversely, the good-enough theory proposes that shallow semantic processing, even involving such underspecification, can be an element of facilitation.

Several works in neurolinguistics focusing on semantic processing also suggest that some type of basic and shallow processing can be at play. In line with the good-enough theory, meaning integration can be switched off when the context renders it unnecessary. This is the case when processing idioms: semantic composition might be not fully engaged during comprehension [Rommers et al., 2013]. In particular, the activation of literal word meanings is only carried out when necessary. Analyzing the cortical responses of semantic violations (see Figure 2.5) shows no significant difference at N400 (the component related to semantic surprisal) between hard and soft violations for idiomatic context, whereas an important reduction in N400 amplitude appears for soft violation compared with hard violation for literal contexts.

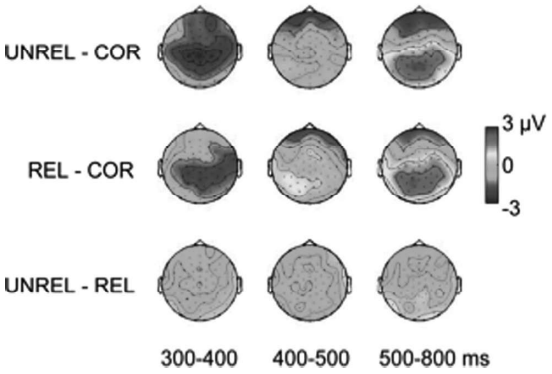


Figure 2.5 The differences between conditions in idiomatic context (COR, correct sentence; REL, soft violation; UNREL, hard violation). The schema shows the difference in the potentials when compared the correct and both types of violation, but no difference between hard and soft violations (UNREL-REL), from [Rommers et al., 2013].

2.2.4 Basic Processing Units: Words or Chunks?

One important question is that of the types of units that are used during LP. The classical option, considering LP as strictly incremental and compositional, consists in recognizing atomic elements and aggregating them progressively (phonemes, morphemes, words, phrases, etc.). This mechanism leads to an interpretation built by composition of the semantic information available at each level (in particular words and phrases). An alternative option considers that the data input stream (heard or read) is stored in the working memory on the basis of larger units, made of sets of words (also called *chunks*), grouped thanks to a shallow processing, which becomes the basis of the interpretation. Several experimental observations support this idea of a more global not strictly incremental processing.

Many NLP applications are based on such low-level information, relying on the identification of basic relationships between words (co-occurrency, order). This technique, *shallow parsing* [Uszkoreit, 2002; Balfourier et al., 2002; Baldwin et al., 2002], leads to the construction of chunks [Abney, 1991] that consist of groups of adjacent words, usually identified on the basis of their boundary markers rather than the syntactic relationships between their constituents:

```
[When I read] [a sentence], [I read it] [a chunk] [at a time]
```

Several works have shown that chunks can be considered a relevant basic unit for LP. For example, studying eye movements when reading a text shows

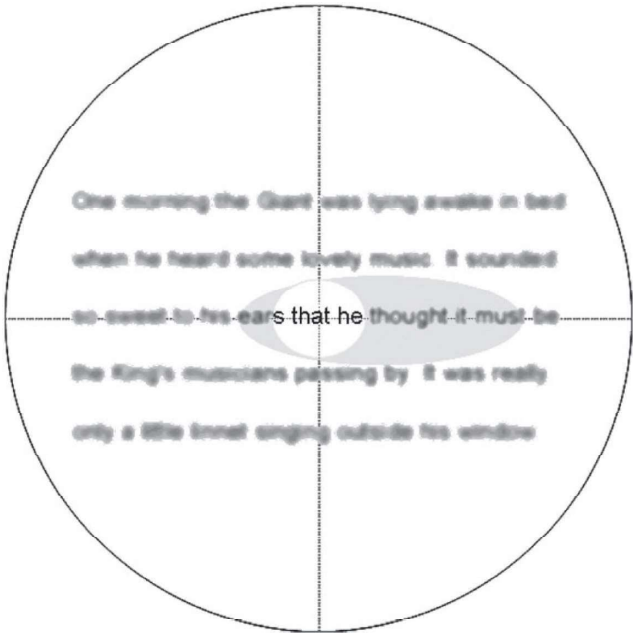


Figure 2.6 Parafoveal vision. Extracting features from the surrounding words.

that fixations are only done from time to time. This is very well known and is the result of *parafoveal vision*, which consists of a preview of adjacent words. As shown in Figure 2.6 (from Schuett et al., 2008), readers extract during a fixation visual information from the foveal visual field (central white oval) and the parafoveal visual field (grey ellipse).

This process makes it possible to extract information about upcoming words, opening the capacity to deal with entire sequences, not only separate words. Moreover, Rauzy and Blache [2012] showed that fixation can be done in chunks (defined here by a sequence of function word and content word).

This observation is an argument in favor of a more global treatment, including at the physiological level. It can be supported by other observations focusing on the neural correlates of LP: several works have specifically studied the question of syntax and, more precisely, its role in the processing of basic properties. In particular, some morphosyntactic properties can be assessed automatically, at a low level, when studying differences between chunks with or without *Det-N* or *Pro-V* agreement violations [Pulvermüller et al., 2008; Pulvermüller, 2010]. When comparing the two conditions, one observes a difference in the cortical reaction at a very early stage (around 100 ms after the stimulus, see Figure 2.7).

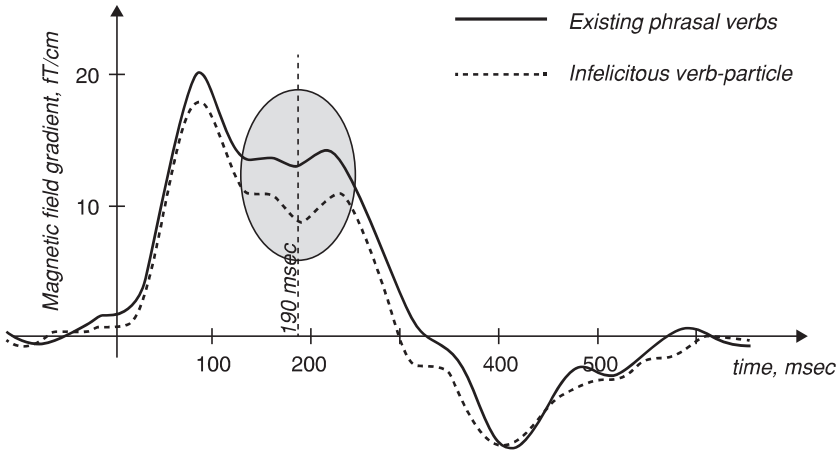


Figure 2.7 Early EEG effects of syntactic violation (mismatch negativity), from [Cappelle et al., 2010].

This effect, called *mismatch negativity*, occurs in a time range and experimental design in which there is no strictly conscious activity. This research suggests that syntax can function as a discrete combinatorial system implemented by discrete combinatorial neuronal assemblies in the brain [Pulvermüller, 2010] that can connect categories into larger constructions.

2.2.5 *Intermediate Summary*

Many of the observations presented to this point tend to indicate that the organization of language processing is not homogeneous. The classical, but also simplest way to explain LP consists in distinguishing different modules that are organized in a serial schema. Supporting this idea, many experiments (judgments, eye movement, brain activity) have shown the respective effects and contribution to these modules in global processing. They usually rely on a theoretical concept of language in which each module bears a subpart of the information, the global interpretation resulting from their composition.

However, this approach appears to be too simple when faced with the description of natural data. First, we know that language is intrinsically heterogeneous, made of different sources of information, different modalities, that interact at any time, producing a complex signal. In a natural environment (typically conversation), the linguistic signal is made of different sources that cannot be strictly separated and analyzed independently from the others. It is made of multiple streams that are not strictly temporally aligned.

This view of language fits better with many observations presented here so far. In particular, language processing often stays at a shallow level, leading to incomplete processing: chunks, identified in terms of basic properties instead of complete analysis, can be considered the basic processing unit, and give access to a certain level of meaning and interpretation.

These objects can be at different levels and result from the convergence of different sources of information. We distinguish between chunk and construction as follows. A *chunk* is a group of words that are gathered on the basis of low-level morphosyntactic properties. A *construction* is a chunk or a set of chunks that can be associated with a global meaning (which can be figurative).

Chunks and constructions are described as sets of interacting properties instead of structures that are built step-by-step from atomic to complex objects. These properties can be at a low level and are automatically assessed.

At the interpretation level, in line with the notion of construction, we have seen that meaning can in some cases be accessed directly instead of compositionally (e.g., idioms, multiword expressions). Moreover, interpretation is often incomplete or underspecified. In particular, it has been shown that ambiguity can be left unresolved and interpretation delayed (or even never completely built).

We propose to take into consideration these different features, gathering them into a language-processing architecture capable of accounting for different types of processing, at different levels, depending on the type of input available. The objective is to describe any type of situation, from the more controlled (e.g., laboratory speech, isolated words) to the more natural (that is, conversation). The proposal relies on the idea that, according to the context and the sources of information, LP can be either serial, modular, and compositional or, conversely, parallel, integrated, and directly interpretable. This approach induces a hybrid processing: one, at a low level, is shallow and partial and supposed to operate by default. The second, which relies on deep, modular, and compositional parsing, is activated when processing complex material (in other words, when interpretation becomes difficult). This organization comes with several assumptions:

- Instead of a word-by-word parsing, LP is based on chunks.
- Chunks are specified in terms of low-level properties, automatically assessed (i.e., without needing deep analysis).
- Semantic interpretation can be delayed.
- Chunks offer the possibility of direct access to the meaning.

In the remainder of this chapter, we will investigate these aspects by addressing specific questions:

- What is the nature of basic properties?: constraints.
- How can basic properties specify entire chunks?: constraint interaction.
- How to access directly from low-level properties to meaning: constructions.

2.3 The Theoretical Framework: Property Grammars

We present in this section the main features of *property grammar* (PG) [Blache, 2000]. PG is a linguistic theory that proposes a constraint-based processing architecture. More precisely, all linguistic information in PG is represented by means of different properties (implemented as constraints). At the difference with the classical generative paradigm, there is no specific module: all properties are mutually independent, offering the possibility to represent separately the different types of information, whatever their domain (morphology, syntax, semantics, etc.) or their level (relationships between features, categories, chunks, etc.). These properties connect the different words of a sentence when processing an input. As a consequence, instead of building a structure, the processing mechanism consists here in describing the input by identifying its different properties. Focusing on syntax and semantics, the following list summarizes the possible relationships between words:

- *Linearity*: Linear order that exists between two words.
- *Co-occurrence*: Mandatory co-occurrence between two words.
- *Exclusion*: Impossible co-occurrence between two words.
- *Uniqueness*: Impossible repetition of a same category.
- *Dependency*: Syntactic-semantic dependency between two words. Different types of dependencies are encoded: complement, subject, modification, specification, etc.

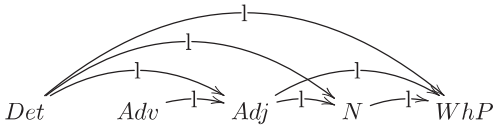
A grammar in PG is a set of all the possible relationships between categories, describing the different constructions. When parsing a given sentence S, assessing a property of the grammar consists in verifying whether the relations between two categories corresponding to words of S are satisfied or not. We present an overview of each type of property.

Linearity: This property implements the same kind of linear precedence relationship as proposed in generalized phase structure grammar [Gazdar et al., 1985]. For example, the nominal construction in English must follow the linearity properties:

$$Det \prec Adj; \quad Det \prec N; \quad Adj \prec N; \quad N \prec WhP; \quad N \prec Prep \qquad (2.1)$$

Note that relationships are expressed directly between the lexical categories. As such, the $N \prec Prep$ property indicates precedence between these two categories regardless of their other dependencies. The following example illustrates

the linearity relationships in the nominal construction "*The very old reporter who the senator attacked*":



(2.2)

In general, properties are also used to control attribute values. For example, one can distinguish linearity properties between the noun and the verb, depending on whether *N* is subject or object by specifying this value in the property itself:

$$N[subj] \prec V; \qquad V \prec N[obj] \tag{2.3}$$

Co-occurrence: This property typically represents subcategorization, implementing the situation in which two categories must be realized together. An example of co-occurrence within a verbal construction concerns nominal and prepositional complements of ditransitive verbs, which are represented by means of the following properties:

$$V \Rightarrow N; \qquad V \Rightarrow Prep \tag{2.4}$$

It should be noted that co-occurrence not only represents complement-type relationships it can also include co-occurrence properties directly between two categories independent from the head. For example, the indefinite determiner is not generally used with a superlative:

a. *The most interesting book of the library*
b. **A most interesting book of the library*

In this case, there is a co-occurrence relation between the determiner and the superlative, which is represented by the property:

$$Sup \Rightarrow Det[def] \tag{2.5}$$

Exclusion: In some cases, restrictions on the possible co-occurrence between categories must be expressed (e.g., lexical selection, concordance). The following properties describe some restrictions in nominal constructions:

$$Pro \otimes N; \qquad N[prop] \otimes N[com]; \qquad N[prop] \otimes Prep[inf] \tag{2.6}$$

These properties stipulate that a pronoun and a noun, a proper noun and a common noun, and a proper noun and an infinitive construction introduced by a preposition cannot be realized simultaneously.

Uniqueness: Certain categories cannot be repeated inside a governing domain. More specifically, categories of this kind cannot be instantiated more than once in a given domain. The following example describes the uniqueness properties for nominal constructions:

$$Uniq = \{Det, Rel, Prep_{[inf]}, Adv\} \tag{2.7}$$

These properties are classical for the determiner and the relative pronoun. They also specify here that it is impossible to duplicate a prepositional construction that introduces an infinitive ("*the will to stop*") or a determinative adverbial phrase ("*always more evaluation*").

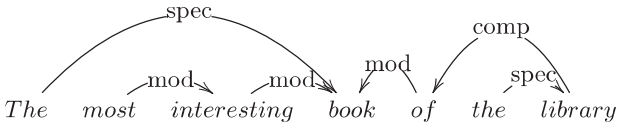
Dependency: This property describes syntax-semantics relationships between categories, indicating that the dependent category complements the governor and contributes to its semantic structure. Dependency relationships are type-based, following a type hierarchy, making it possible to vary the level of precision of the relationship, from the most general to the most specific. These types and subtypes correspond to a classical syntactic relationship:

dep: Generic relationship, indicating dependency between a constructed component and its governing component.
mod: Modification relationship (typically an adjunct).
spec: Specification relationship (typically *Det-N*).
comp: The most general relationship between a head and an object (including the subject).
subj: Dependency relationship describing the subject.
obj: Dependency relationship describing the direct object.
iobj: Dependency relationship describing the indirect object.
xcomp: Other types of complementation (e.g., between *N* and *Prep*).
aux: Relationship between the auxiliary and the verb.
conj: Conjunction relationship.

Dependency is noted $\rightsquigarrow$, possibly completed with the dependency subtype as an index. The following properties indicate the dependency in nominal constructions:

$$Det \rightsquigarrow_{spec} N[com]; \qquad Adj \rightsquigarrow_{mod} N; \qquad WhP \rightsquigarrow_{mod} N \tag{2.8}$$

The following example illustrates some dependencies in a nominal construction:



(2.9)

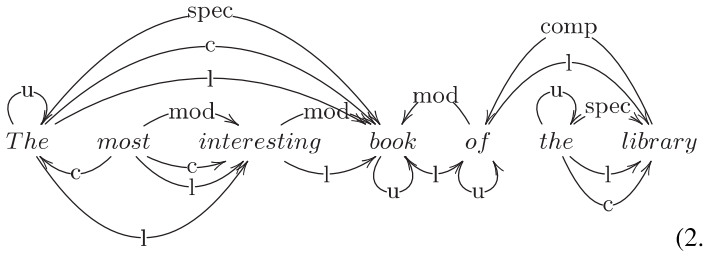
This schema illustrates the specification relationship between the determiner and the noun, the modification relationship between the adjectival and prepositional constructions, and the modification between the adverb and the adjective inside the adjectival construction.

Example: Each property as defined earlier corresponds to a certain type of syntactic information. In PG, the description of syntactic units or linguistic phenomena (chunks, constructions) in the grammar consists of gathering all the relevant properties into a set. Schema 2.10 summarizes the properties describing the nominal construction.

$Det \prec \{Det, Adj, WhP, Prep, N\}$ $N \prec \{Prep, WhP\}$	$Det \rightsquigarrow_{spec} N$ $Adj \rightsquigarrow_{mod} N$
$Det \Rightarrow N[com]$ $\{Adj, WhP, Prep\} \Rightarrow N$	$WhP \rightsquigarrow_{mod} N$ $Prep \rightsquigarrow_{mod} N$
$Uniq = \{Pro, Det, N, WhP, Prep\}$	$Pro \otimes \{Det, Adj, WhP, Prep, N\}$ $N[prop] \otimes Det$

(2.10)

A syntactic description, instead of being organized around a specific structure – a tree, for example – consists of a set of independent properties together with their status (satisfied or violated). The graph in Schema 2.11 illustrates the PG description of the nominal construction: *The most interesting book of the library*, where l represents linearity, u is uniqueness, and c is co-occurrence.



(2.11)

In PG, a syntactic description is therefore the graph containing all the properties of the grammar that can be evaluated for the sentence to be parsed. As illustrated in the example, this property graph represents explicitly all the syntactic characteristics associated with the input, and each is represented independent from the others.

2.4 Chunks, Constructions, and Properties

As we noted in Section 2.2 An Interdisciplinary View of Language Processing, many observations tend to show how important chunks and constructions can

be in LP architecture. Our hypothesis is that two different types of processing coexist: one serial, incremental, word-by-word and compositional (the classical LP organization in the literature) and another shallow, based on chunks or constructions, recognized as a whole, giving, when possible, direct and global access to the meaning. This hypothesis is supported by several observations, showing the existence of such units in particular when studying the brain correlates of language processing. Moreover, some basic properties (typically agreement) are identified at a very early stage, indicating an automatic and low-level process. We will first explain what these basic properties are and then how chunks or constructions can be recognized on the basis of the properties they contain.

2.4.1 Basic Properties

The different properties presented in earlier sections can be assessed directly when processing a sentence: for each set of categories, it is possible to verify whether some properties link them in the grammar and whether, in the specific context of their realization in the sentence, they are satisfied or not. A property plays exactly the role of a constraint, describing an input consists in assessing the properties, assigning them a truth value.

Two types of properties can be distinguished, according to the way they can be evaluated and their sensitivity to the context [Blache and Rauzy, 2004]. More precisely, a property can be assessed as soon as the categories they concern are recognized in a sentence. The difference between the two types of properties is that in one case, their satisfaction remains the same whatever the window of words taken into consideration, and in the other case, this value can vary depending on the window (being sensitive to the context).

- *Success-monotonic properties*: When a property between two categories becomes satisfied, it remains satisfied for the entire sentence. For example, the linearity between *most* and *interesting* in Schema 2.11 holds as soon as it can be assessed, and remains satisfied until the end, whatever the span of words.

  More formally, the linearity relationship $a \prec b$ is satisfied in the sequence of words $s = [\gamma, a, b, \eta]$, whatever the composition of γ and η. Two types of properties are success-monotonic: *linearity* and *co-occurrence*.
- *Success-nonmonotonic properties*: A property can be satisfied locally and become violated at a larger span: the evaluation of a property depends on the set of categories taken into account. For example, an *exclusion* relationship between the words a and d is satisfied within the set of words $s1 = \{a, b, c\}$, but false when adding a new category d to this sequence $s2 = \{a, b, c, d\}$. In

this case, it is necessary to specify the sequence (or the partition) for which the constraint is evaluated.

Success-monotonic properties are computationally simpler than nonmonotonic properties because they do not need to be re-evaluated at each step; when such a property becomes satisfied, this assessment cannot be reconsidered. We characterize these types of properties as basic. They are low-level properties, automatically assessed at an early stage in the brain. Moreover, they encode the two types of information used when evaluating transition probabilities between categories (linearity and co-occurrence), reinforcing the proposal to consider them at a first level.

2.4.2 Chunks from Properties

We have seen how to recognize and assess properties. The question is now how they can be used to identify a chunk or a construction. Several experiments have shown that some syntactic properties can be assessed very early, without any deep and precise analysis (see in particular Pulvermüller et al., 2008). In our hypothesis, they correspond to "basic properties," which can be evaluated based on the immediate context. The next challenge is to learn how to identify higher-level organizations such as chunks and constructions.

Our proposal relies on the idea that in some cases, there exists a link between properties: the existence of some properties can also activate other properties. For example, the verification of a linearity property between a *Det* and a *N* activates a dependency relationship between them. This same kind of relationship exists in lexical selection, collocations, etc.: the realization of a given word activates or predicts that of another one.

As a consequence, the description of a construction in the grammar consists in two types of information: the set of properties and the identification of those basic properties that can activate other ones. We propose to add this information to the representation of the properties by means of a new argument encoding the properties that can be linked to the current as follows:

```
<id, type, source, target, weight, linked_props>
```

The `linked_props` argument is a set of indexes that point to other properties describing the same construction. For example, the dependency relationship between a preposition and a noun depends on the linearity: if $Prep \prec N$, then *Prep* is the head and *N* depends on it. Reciprocally, when $N \prec Prep$, the *Prep* depends on *N*. These relationships between properties are represented as follows:

```
<1, lin, Prep, N, H, {}>   <2, comp, N, Prep, S, {1}>
<3, lin, N, Prep, H, {}>   <4, mod, Prep, N, S, {3}>
```

The example of the ditransitive construction can be implemented in the same manner, specifying different dependency types[1] according to the form (the first noun is the indirect object, the second the direct):

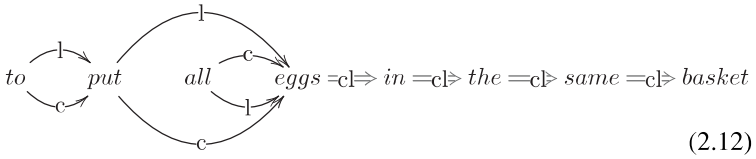
```
<1, lin, V[dit], N1, H, {}>   <4, iobj, N1, V, H, {1,2,3}>
<2, lin, V[dit], N2, H, {}>   <5, obj, N2, V, H, {1,2,3}>
<3, lin, N1, N2, H, {}>
```

2.4.3 Processing Idioms

The activation mechanism based on linked properties can be generalized to the processing of other types of constructions. We know, for example, that processing idioms (see Vespignani et al., 2010 and Rommers et al., 2013), is done in two different steps, before and after the word starting from which the idiom is recognized (called the *recognition point*, or RP). Before the RP, the processing consists of assessing basic properties. At the RP, the idiom is recognized, its meaning is globally accessed, without any need to analyze the rest of the idiom. All the remaining words become fully predictable. In terms of properties, this means that a set of mandatory co-occurrences as well as linearity between the list of words is activated.

This phenomenon can be implemented with the mechanism of *linked properties*: reaching the RP means having already assessed a certain amount of basic properties, relating the initial words of the idiom. Recognizing the RP consists of inferring a set of linked properties from the basic ones.

The following figure illustrates this mechanism for the idiomatic expression *to put all eggs in the same basket*:



(2.12)

In this idiom, the RP is at the word *eggs*. Before the RP, the basic properties are assessed, linking the first words of the idioms. After this point, all the other properties can be automatically inferred, as well as the association of a global meaning. The rest of the process consists only of verifying whether the prediction matches the remaining words.

The property-based description of this idiom can be implemented with the following properties:

[1] At this stage, to be as generic as possible, representation of the dependencies usually requires underspecification.

(1) $put \prec all$
(2) $all \prec eggs$
(3) $put \Rightarrow \{in, one, basket\}$
(4) $eggs \prec in \prec one \prec basket$
(5) $sem(put) = [[risk\_losing\_everything]]$

In this description, we only describe the basic linearity and co-occurrence properties. The RP is implemented by factorized properties (3) and (4). The general mechanism is described by the following formula, indicating that properties (3), (4), and (5) can be inferred directly from the basic properties (1) and (2):

$$(1 \wedge 2) \Rightarrow (3 \wedge 4 \wedge 5)$$

Note that the semantics of the idiom is represented by a denotation attached, arbitrarily, to the verb (the idiomatic construction being verbal in this case).

2.5 The Hybrid Architecture

The language-processing architecture we propose is an alternative to the classical incremental, modular, and serial organization. We think that, instead of processing word-by-word by trying to integrate each new word into a partial structure and interpreting the result compositionally (the meaning of the whole being a function of each component), it is preferable to propose a flexible architecture, more in line with what is observed in human LP.

2.5.1 General Organization

The first general idea is that processing is not strictly incremental and meaning access not compositional. We have described a basic processing level in which words are gathered into larger units. There are some situations, typically constructions, in which meaning is assessed directly. This means that two different types of processing are juxtaposed: the first type, which relies on low-level mechanisms and is considered the default level, directly identifies chunks that offer the potential for global processing. The second type is classical, word-by-word, serial, and compositional, and is applied when the first is not possible.

Generally speaking, we consider that the first level of processing is superficial and delay as much as possible the interpretation. Whenever possible, larger units grouping several words are built. Such groups make it possible to gather different sources of information, preparing a first level of interpretation. In some cases, they even constitute entire constructions, offering the possibility to directly access the meaning. This first level of processing is supposed to be done automatically, on the basis of low-level mechanisms.

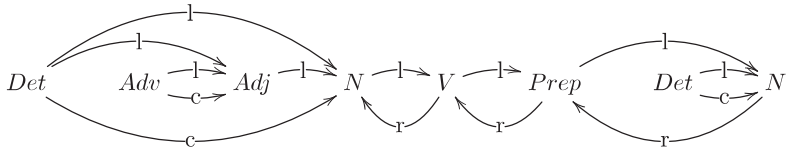
These units are stored in the working memory, together with their interpretation when it exists. Global interpretation then groups the different local meanings together. When no grouping is possible, then words (instead of groups) are stored in the memory. In both cases, the interpretation process is only done after gathering a certain amount of information (or when reaching the maximal capacity of the working memory).

As a result, different types of objects coexist: words, chunks, and constructions. The existence of units grouping words facilitates the processing: these objects are recognized by means of low-level mechanisms and offer the possibility to directly contribute to the meaning. We propose a method for identifying these units.

2.5.2 *Recognizing chunks*

Chunks are set of words, usually adjacent (but not necessarily), linked by tight morphosyntactic relationships (typically *Det-N*). As noted earlier, such relationships mainly correspond to what we call basic properties, that is, linearity and co-occurrence. When parsing an input, processing a new word consists of checking such properties with adjacent words. The resulting graph makes it possible to identify subgroups, formed by the set of words that are connected by such properties. When looking at a constraint graph obtained from basic properties, such subgroups can be immediately identified: they correspond to the complete subgraphs (the set of nodes in a graph that are directly connected).

The example in Schema 2.13 illustrates a constraint graph, containing basic properties only:

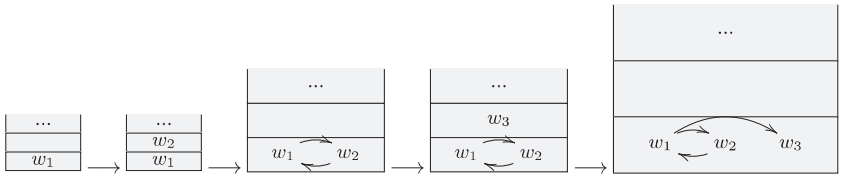


(2.13)

In this graph, we can identify several subset of nodes that are all connected. For example, the subset *Adv-Adj* forms a complete subgraph, but not *Det-Adv-Adj*: in this last case, *Det* is not connected with *Adv*. A complete subgraph is then made of words with morphosyntactic relationships that corresponds to our definition of chunks. The list of complete subgraphs in Schema 2.13 is: *Adv-Adj; Det-Adj-N; N-V; V-Prep; Prep-N; Det-N*. This corresponds to the list of the chunks, identified only by means of basic properties.

Chunk recognition constitutes the first level of processing. In the following, we leave aside the question of word segmentation and recognition during human LP (even though it remains an open question) and consider the initial

input as a set of words. The processing consists of scanning the input word by word. The initial word of the input is simply stored in the working memory, which is made of several buffers (represented here as a stack). The next step consists of scanning a new word, pushing it into the stack of buffers, and then determining whether it can be connected by basic properties with the word in the lower buffer. When the new word can be linked to the previous word, the two words form a chunk, which is pushed into the buffer. In the same way, a word can be linked to an existing chunk, forming a new chunk to replace the previous one, as illustrated in Figure 2.8.



2.5.3 Recognizing Constructions

The identification of constructions can be explained thanks to linked properties and cohesion evaluation. They constitute a distinct mechanism that makes it possible, starting from a chunk, to identify the type of the construction and complete its description with new properties.

A chunk is made of words connected with basic properties. In some cases, this set of properties can be associated to linked properties, resulting in the inference of new relations completing the constraint graph.

As shown in Figure 2.8, this mechanism explains the effect of the recognition point in idiom processing. Before this point, the preceding words are processed as explained in Section 2.4.3. Processing Idioms, building chunks when possible. When reaching the word corresponding to the RP, a set of linked properties is directly assessed. We obtain then a new constraint graph which also bears for idioms a complete interpretation; in this case, the entire set of properties describing this specific construction is formed by linked properties, making it possible to infer directly the description and its interpretation.

For other types of constructions, with a certain level of flexibility [Goldberg, 2006], only subparts of the properties are linked and can be automatically inferred from the basic ones. In this situation, an evaluation of the graph density completes the mechanism. The constraint graph, completed by the possible linked properties, is analyzed. If its density value reaches a certain threshold, then the construction is recognized and the entire set of linked properties is activated. This mechanism gives direct access to the meaning associated with the construction, to be completed when scanning the rest of the input.

2.5.4 Light Parsing with Chunks and Constructions

Recognizing chunks and a fortiori constructions facilitates language processing because it allows direct access to a certain interpretation thanks to basic properties. In our approach, this constitutes the first level of processing based on the assessment of basic properties that can trigger the inference of other types of properties thanks to the mechanism of linked properties. In this hypothesis, constructions are encoded in the memory as *recurrent networks*, encoding directly linked properties. This view fits with the *MUC model* (memory, unification, control) proposed by Hagoort [Hagoort, 2005; Hagoort, 2013] in which memory contains lexical building blocks that encode complex lexical entries, including syntactic and semantic relations. MUC is in line with several linguistic theories such as head-driven phrase structure grammars [Pollard and Sag, 1994] or tree-adjoining grammars [Joshi and Schabes, 1997] in which most syntactic and semantic information is encoded in the lexicon. The memory stores such units when the unification component is in charge of integrating them. In our model, the linked properties are stored in the memory together with lexical units. The unification component can then directly assemble units thanks to a simple mechanism: basic properties assessment plus linked properties inference.

Finally, light processing architecture distinguishes two levels of unification during sentence processing:

- *Light level*: Used as default, storing words in the working memory, assembling them into chunks, inferring linked properties and activating constructions when possible.
- *Deep level*: Used when the light level does not lead to interpretation. The processing is classical: strictly incremental and serial, interpretation being built compositionally, starting from a syntactic structure.

2.6 Conclusion

Sentence processing is fast and happens in real time despite the complexity of linguistic mechanisms. One explanation is that language makes use of frequent structures or patterns that can be learned. This justifies the use of probabilistic approaches that is at work today in most LP models. However, several experiments have shown that even some types of rare structures can be processed in real time [Pulvermüller, 2010]. We have described in this chapter a new way of representing linguistic information, based on properties, and we have shown how two types of such properties can be distinguished, according to the way they can be verified. Some properties, known as basic properties, are assessable simply and directly. Moreover, we have shown how properties, whatever their type, can be linked and directly inferred from each other. This

mechanism (basic assessment+inference) is the basis for recognizing chunks and constructions. It constitutes the first level of parsing in our model, based on *light parsing*. This processing mechanism is the default one, explaining why language processing can be often shallow but fast: interpretation can be directly accessed thanks to such basic mechanism. In some difficult and complex cases, light parsing does not lead to any interpretation. In such cases, a classical *deep parsing*, incremental and serial, is used.

The *light and deep parsing model*, therefore, constitutes a candidate for a new language processing architecture, explaining why human LP is efficient and opening the way to new types of experiments in neurolinguistics.

References

Abney, S. (1991). Parsing by chunks. In *Principle-Based Parsing: Computation and Psycholinguistics*, Dordrecht; Boston: Kluwer Academic Publishers, pages 257–278.

Adriaens, G. and Hahn, U., editors (1994). *Parallel Natural Language Processing*. Norwood, NJ: Ablex Publishing Corporation.

Altmann, G. T. M. and Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, 33(4):583–609.

Baldwin, T., Dras, M., Hockenmaier, J., King, T. H., and van Noord, G. (2002). The impact of deep linguistic processing on parsing technology. In *Proceedings of IWPT-2007*.

Balfourier, J.-M., Blache, P., and Rullen, T. V. (2002). From shallow to deep parsing using constraint satisfaction. In *Proc. of the 6th Int'l Conference on Computational Linguistics (COLING 2002)*.

Blache, P. (2000). Property grammars and the problem of constraint satisfaction. In *Linguistic Theory and Grammar Implementation*, ESSLLI 2000 workshop.

Blache, P. and Rauzy, S. (2004). Une plateforme de communication alternative. In *Actes des Entretiens Annuels de l'Institut Garches*, pages 82–93, Issy-Les-Moulineaux, France.

Cappelle, B., Shtyrov, Y., and Pulvermüller, F. (2010). Heating up or cooling up the brain? MEG evidence that phrasal verbs are lexical units. *Brain and Language*, 115(3), 189–201.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht; Cinnaminson, NJ: Foris Publications.

Copestake, A., Flickinger, D., Pollard, C., and Sag, I. (2001). Minimal recursion semantics: An introduction. In *Language and Computation (L&C)*, volume 1, pages pp. 1–47. Oxford: Hermes Science Publishing.

de Marneffe, M.-C. and Manning, C. D. (2008). Stanford typed dependencies manual. Technical report, Stanford Parser v. 3.5.2.

Ferreira, F. and Patson, N. D. (2007). The "good enough" approach to language comprehension. *Language and Linguistics Compass*, 1(1).

Fillmore, C. J. (1988). The mechanisms of "construction grammar." In *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*, pages 35–55.

Fodor, J. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.

linguistics, through both experts in controlled environments and crowdsourcing of naive annotators (Snow et al., 2008). Experimental psycholinguists have used a range of methods that do not rely on intuition, judgments, or subjective reflection, such as the speed of self-paced reading, or the order and timing of gaze events as recorded with eye-tracking technologies (Rayner, 1998).

Brain-recording technologies offer a different kind of evidence, as they are the closest we can get empirically to the object of interest: human cognition. Despite the technical challenges involved, especially the complexity of the recorded signals and the extraneous noise that they contain, brain imaging has a decades-long history in psycholinguistics. Particular patterns of electrophysiological activity are associated with processing difficulties in the meaning and structure of sentences, and relative changes in blood flow can reveal parts of the brain whose activity is modulated by the complexity of a language processing task (see Section 3.2). These experimental approaches usually frame theoretical questions in terms of a small number of categories (e.g., familiar words versus obscure words to look at the processing associated with lexical retrieval), and in terms of tasks that try to "stretch" or "break" language to see how it functions (e.g., through the use of ill-formed sentences).

In this chapter, we present an additional stream of recent work that uses computational modeling of both language and brain activity to build on this research.

In Section 3.3, we describe studies that explore the breadth and depth of the human lexicon. Models from computational lexicography and the word vector space/embedding literature are employed to empirically model the various dimensions of meaning along which words can differ, which are common to many early (Katz and Fodor, 1963) and current theories (Barsalou and Wiemer-Hastings, 2005; Baroni and Lenci, 2010) of word meaning. By employing distributional semantic theory as a "prior", we can use computational models to separate those aspects of brain activity that are related to word meaning from those related to other aspects of the experimental task or extraneous noise.

In Section 3.4, we look above the level of the word, at how lexical units combine, in real time, to form short phrases. As in the work with single words, we can take advantage of existing theories of language to characterize the representations produced by compositional processes. This takes advantage of the fact that huge quantities of textual data are cheaply available to build and evaluate fine-grained models, and then these models can be tested against the expensive and limited collections of brain data. Again, we differentiate the brain activation attributed to the act of composition from the brain activity attributable to the composed semantics of a phrase.

Most recently, there has been a movement in cognitive neuroscience toward the use of naturalistic tasks, which are claimed to be more ecologically valid

(Willems, 2015). In Section 3.5, we describe some experiments that use natural reading and listening tasks to elicit holistic and realistic language processing, without resorting to constructed stimuli with hand-engineered syntactic or semantic errors. A combination of tools from computational linguistics and crowd annotation allows us to build detailed multilevel models of the perceptual, structural, and semantic work involved in understanding a real narrative. As with the word- and phrase-level work, the use of such a model brings several advantages. Computational modeling of the relationship between word features and brain areas differentiates the brain activity driven by language processing from irrelevant brain activity. Closer inspection of those sensitivities can tell us which brain areas and which parts of the time course are tied to particular subprocesses. And the generative nature of the models allows us to perform the "mind reading" trick of estimating (imperfectly, but at a level clearly above chance) what word, phrase, or sentence a person is processing at a given moment.

In this review we concentrate on research that uses recordings of brain activity and computational modeling as a tool to probe how language functions in the mind, rather than work that uses language as a probe to understand brain function. There is also a very large body of work that develops models of brain activity that build in it spatial, temporal, and network characteristics; that work is not covered here.

3.2 Grounding Language Architecture in the Brain

The earliest investigations of language in the brain began in the 1800s, when Paul Broca and Karl Wernicke studied patients with brain injuries that affected their ability to communicate (Bear et al., 2007). Broca's patients exhibited partial or complete loss of language abilities (aphasia), and their pattern of brain damage prompted him to conclude that language is controlled by a single hemisphere of a person's brain, almost always the left hemisphere (Dronkers et al., 2007).[1] Broca's work also led him to identify a region of the brain in the posterior inferior left frontal gyrus ("Broca's area," see Figure 3.1) associated with a particular variety of aphasia. Nonfluent aphasia, also called Broca's aphasia or expressive aphasia, is characterized by the impairment of language production, while comprehension and general cognition remain intact.

Wernicke found that lesions to a different left hemisphere region (posterior superior temporal gyrus, or "Wernicke's area," see Figure 3.1) led to a distinct

[1] Throughout this chapter, we refer to the typical left-lateralized localization of language areas in the brain of a right-handed person. Left-handed people (and, indeed, some right-handed people) have language centers located in the right hemisphere of the brain.

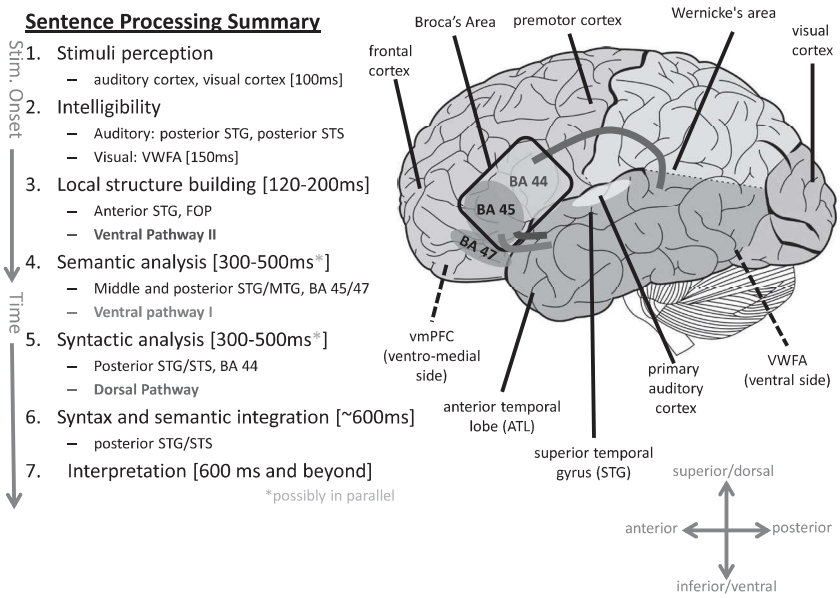


Figure 3.1 An overview of the location and timing of language processing (for sentences) in the brain. Posterior refers to areas toward the back of the head; anterior, toward the front; superior/dorsal, toward the top of the head; inferior/ventral, toward the bottom. Medial refers to locations toward the middle of the brain, where the two hemispheres meet. A gyrus refers to a ridge of the cortex, and a sulcus is the fold between gyri. BA, Brodmann area; STG, superior temporal gyrus; STS, superior temporal sulcus; VWFA, visual word form area; FOP, frontal operculum (just ventral to Broca's area); MTG, middle temporal gyrus (between superior and inferior gyri); vmPFC, ventromedial prefrontal cortex. Adapted from Friederici (2011).

pattern of language impairment. Fluent aphasia (also jargon aphasia or Wernicke's aphasia) is characterized by the easy production of language that is mostly nonsensical or wandering. Intonation and speed of speech are usually normal, and if one ignores the content of the utterance, the speech can seem quite typical. Patients often have difficulty following verbal instructions, indicating that language understanding is also affected. These symptoms have led to the theory that these areas are instrumental in mapping language sounds or written words to semantic content.

Since the 1970s the study of language in the brain has been transformed by brain-imaging technologies, most commonly electroencephalography (EEG), magnetoencephalography (MEG), and functional magnetic resonance imaging (fMRI).

EEG is the oldest of the three brain-imaging technologies discussed here. EEG measures the voltage fluctuations on the scalp, induced by the coordinated firing of groups of brain cells, colloquially referred to as brain waves. The propagation of these postsynaptic electrical fields to the scalp is instantaneous, but is distorted spatially and temporally through low-pass filtering, as it passes through varying densities of tissue and the skull. Scalp signals can be resolved back to underlying brain sources after making some physical assumptions, but even in ideal settings spatial resolution is poor, on the order of 7 mm (Im et al., 2007). On the other hand, EEG gives excellent temporal resolution, usually recorded at hundreds or thousands of samples per second (Hz), whereas the maximum firing rate of neurons is typically about 50 Hz. EEG has an additional advantage among the modalities discussed here in that it requires quite simple equipment (essentially sophisticated signal amplifiers) and so can be used in a range of lab and nonlab environments. Miniaturization of electronics has recently made wearable EEG a reality for both research and consumer uses.

MEG measures the minuscule magnetic field corresponding to the electrical fields detected by EEG. Like EEG, MEG measures the postsynaptic currents in apical dendrites (i.e., the arrival of a new electrical message to a neuron), particularly of cells in the sulci ("valleys" of the cortical folding) (Hansen et al., 2010). The propagation of these magnetic fields is not distorted by passage through the head, and so the signals, although similar in kind to EEG, are cleaner and contain more high-frequency content (a sample MEG recording appears in Figure 3.2). MEG signals can be resolved to locations in the brain with a much higher spatial resolution, on the order of 2–3 mm in ideal conditions (Hamalainen et al., 1993). However MEG is an expensive and complex technology, requiring a magnetically shielded room and supercooling to support superconducting magnetometers.

The imaging technique with the greatest spatial resolution is fMRI, which can achieve resolution as fine as 1 mm. A sample fMRI image appears in Figure 3.3. fMRI measures changes in blood oxygenation in response to increased neuronal activity, called the blood oxygen level–dependent (BOLD) response. Because fMRI depends on the transport of oxygen via blood to the brain, its time constant is governed by the rate at which blood can replenish oxygen in the brain. Although fMRI can acquire images at the rate of about one image per second, the BOLD response can take more than 5 s to reach its peak after a stimulus is shown. Thus, among the three modalities discussed here, fMRI has the worst time resolution and the best spatial resolution.

Neurolinguistic studies generally use carefully controlled comprehension tasks, such as rapid serial visual presentation (presenting a phrase or sentence, word by word at a fixed rate on a screen) or sentence reading followed by questions to ensure that the participant is attending to and processing the

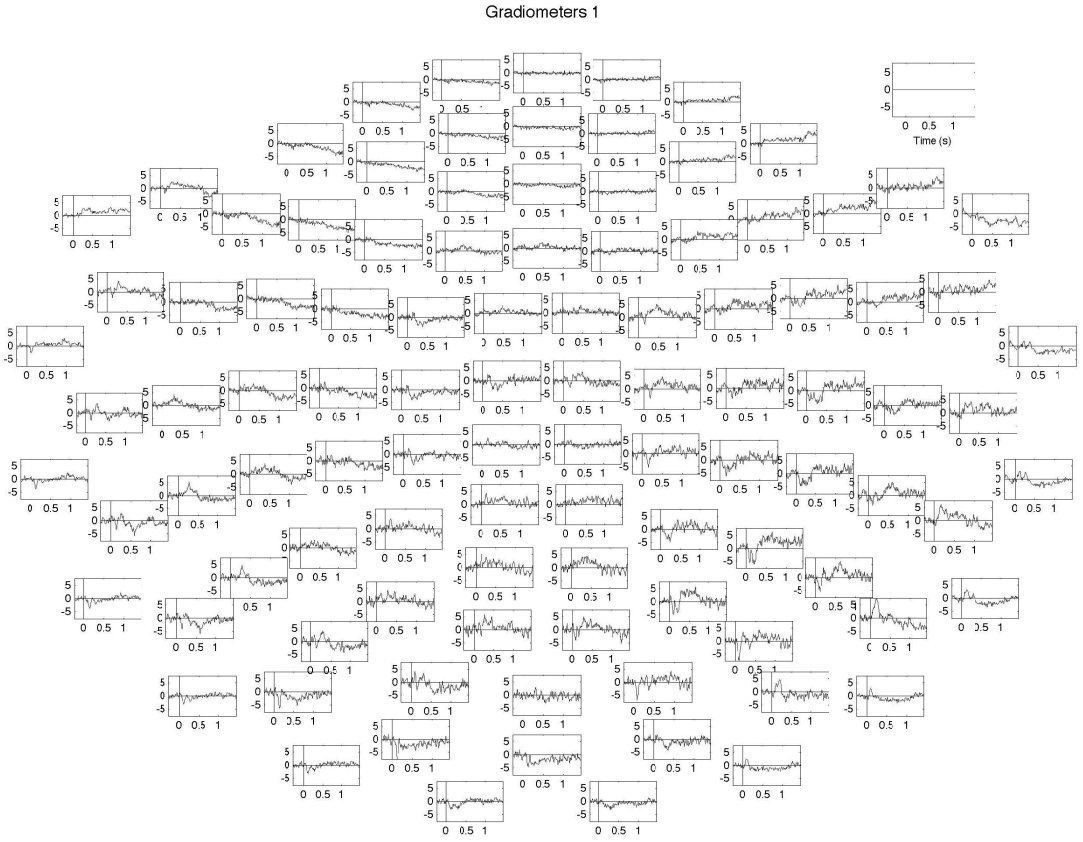


Figure 3.2 An example MEG recording averaged over twenty repetitions of a person reading the word *bear*. Each subplot represents the recording from the first gradiometer at one of the one hundred and two sensor positions on the MEG helmet. For simplicity, the other two hundred and four sensor recordings are not shown. In this diagram, the helmet is oriented as if we are looking down on it from above. The nose of the subject points to the top of the figure, and the left side of figure corresponds to the left hand side of the subject. Time is along the x-axis of each plot and the y-axis corresponds to the gradient of the magnetic field in 10^{-3} T/cm (Fyshe, 2015)

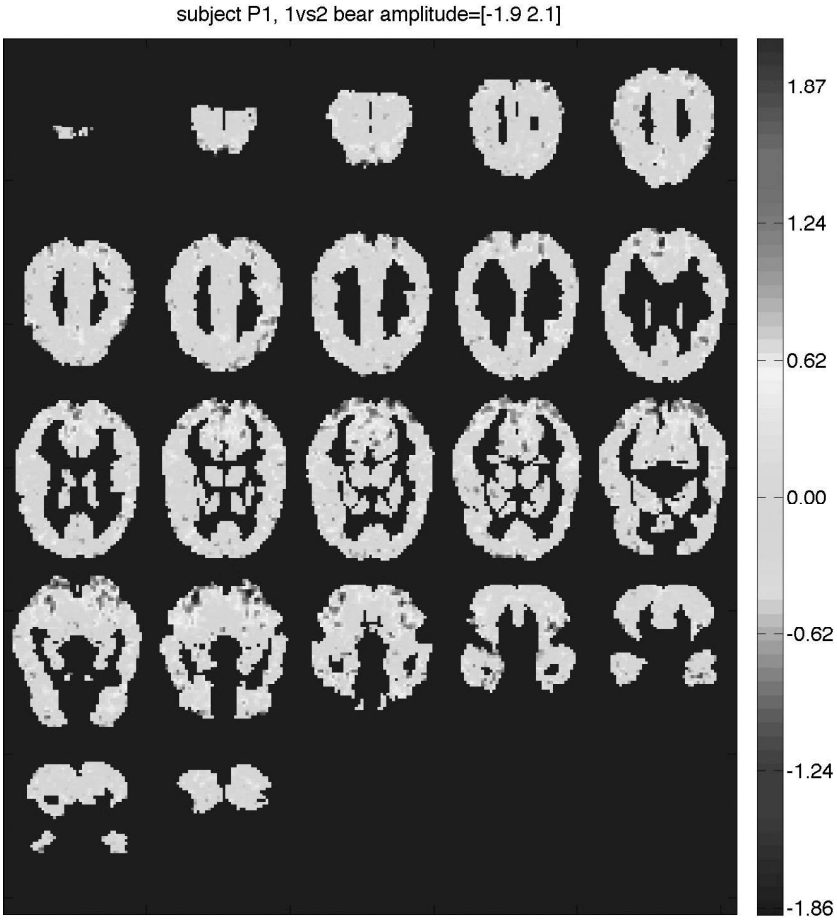


Figure 3.3 An fMRI image averaged over six repetition of a person reading the word *bear*. An fMRI image is three-dimensional, here shown in slices progressing from the top of the head (*top left*) to the bottom of the head (*bottom right*). In each slice, the front of the head points towards the bottom of the figure, and the right side of the subject is shown on the left side of the each image (as if we are viewing the brain of a subject laying face down, from the top of their head). The color of each voxel (pixel in brain space) represents the percent change over baseline of the BOLD response in that brain area (Fyshe, 2015).

materials. This limits the conditions to be analyzed, and avoids machine artifacts due to participant movement. Reading tasks are predominant, as they simplify the question of when (visual) word processing begins. With auditory presentation (Hagoort, 2008; Brennan and Pylkkänen, 2012) the articulation of a word can stretch over many hundreds of milliseconds and it is not self-evident

at which point during the word processing can begin, given anticipatory effects (Marslen-Wilson and Tyler, 1980; DeLong et al., 2005).

Neuroimaging work on language started with EEG. Kutas and Hillyard (1980) noted a more negative current in centroparietal sensors triggered by a semantically mismatched ending to a sentence (e.g., *He spread the warm bread with socks*), which became apparent from about 400 ms after the critical incongruous word appeared. This N400 was originally thought to be a reaction to semantically incompatible words, but it has since been shown that the N400 can be evoked by sentences with semantically less predictable words. For example, *Jenny put the sweet in her (mouth/pocket) after the lesson* elicits an N400 for the word *pocket*. Although *pocket* is a semantically sound word choice, it is judged less probable than the alternative (*mouth*). It has also been shown that the N400 can appear before the incongruent word if the indefinite article (*a/an*) does not match the predicted noun. For example, an N400 will occur for the word *an* in the sentence *It was breezy, so the boy went to fly an kite* because the word *kite* is so strongly predicted and *an* is the wrong indefinite article (DeLong et al., 2005). The N400 can also be evoked in the context of arbitrary lists of words, where the magnitude of the effect is larger for infrequent words, and smaller for words that have been semantically primed by previous words presented. This effect has led to the interpretation that the N400 indexes the cognitive load involved in retrieving a word from the mental lexicon (Kutas and Federmeier, 2011).

In contrast to the N400, the P600 is characterized by a positive-going current that peaks around 600 ms after stimulus onset, also in centroparietal sensors. Typically, the P600 is seen at a sentence position where there is a syntactic violation (e.g., word order mistakes, plural verb disagreement, grammatical gender mismatch) (Kuperberg, 2007). However, under certain circumstances a P600 can be evoked even when the syntax of a sentence is correct. For example, the sentence *Every morning the eggs would eat toast for breakfast* will induce a P600 for the underlined word "eat," although the sentence is syntactically sound, and elicits no N400, although the sentence is semantically incongruent. This phenomenon was called a "semantic illusion" because it was seen as "fooling" the reader into thinking that the word is semantically sound due to a strong conceptual link (Hoeks et al., 2004).

Because of their excellent time resolution, EEG and MEG have provided insights into the manner and order of processing in language comprehension. One major debate in linguistics is on the degree of modularity and serialization in language processing. For example, early models by Friederici (2002) posited a strictly serial model consistent with syntax-central and modular models of language (Fodor, 1983; Chomsky, 1995). The early timing of semantic effects evidenced in the N400 and the fact that it can appear also in nonsentence final positions suggest that syntax analysis does not always strictly precede

semantics, and later studies from the Hagoort lab demonstrated that both encyclopedic knowledge and discourse-specific facts are integrated online into computations of semantic correctness (Hagoort et al., 2004; Özyürek et al., 2007).

The fine spatial resolution of fMRI and MEG have allowed us to study specific regions of the brain during language processing. Levels of processing can be disentangled by comparing the brain activity elicited by different kinds of language materials. For instance, brain areas that are active for real words, but not for pseudowords, might be assumed to be involved in lexical retrieval (McCandliss et al., 2003; Salmelin, 2007). Those areas active for both but not active for a nonlinguistic symbols might be assumed to be involved in reading. And those areas active for phrases and sentences but not for single words presumably play a role in syntactic or semantic composition.

Following this analytical paradigm, large swathes of cortex are implicated as specialized for language processing (Fedorenko et al., 2012). The temporal lobe is broadly involved (including Wernicke's area), is almost always reported as having left-hemisphere predominance, and is thought by many to be primarily responsible for lexical processing. The inferior frontal gyrus (which includes Broca's area) and neighboring left-hemisphere areas are through to be involved in processing sentence structure and meaning, although this is an active area of research (Hagoort, 2005; Friederici, 2011).

Models of language processing have taken inspiration from vision research (Hickok and Poeppel, 2004, 2007; Friederici, 2011), which has broken visual processing into two streams: dorsal and ventral. In human vision, information passes from the low-level perceptual areas at the back of the brain via the ventral stream to posterior temporal areas, to determine *what* an object is, and via the dorsal stream to parietal cortex and motor areas, to determine *where* the object is and *how* it can be manipulated. Applying this same dual-stream hypothesis to language processing, the ventral stream through the temporal lobe is responsible for word semantics (i.e., what), whereas the dorsal stream links motor areas of the brain (including the articulatory network in the posterior inferior frontal gyrus) with auditory and sensorimotor areas of the brain.

Hagoort (2005) proposes a related model that focuses on the neural mechanisms for the unification of language (which covers both semantic composition and syntactic operations), which enables the generation of composed meaning, considered by many theories to be the central defining feature of human language competence. The MUC model consists of three functions:

Memory: Recalling the meaning of a word, lexical access. The temporal cortex and the inferior parietal cortex are involved in the memory process of Hagoort's model.

Unification: Integrating the retrieved meaning of a word with the meaning representation calculated with the context leading up to that word. This

includes extralinguistic sources of meaning like gesture and gender of speaker. This processing resides in the left inferior frontal cortex, with Brodmann areas BA 47 and BA 45 specialized for semantic unification, and neighboring areas BA 45 and BA 44 (Broca's area) specialized for syntax (see Figure 3.1) (Hagoort, 2014).

Control: Governing the actions required for language, like taking turns during a conversation. Control requires the dorsolateral prefrontal cortex, anterior cingulate cortex (ACC) and the parts of the parietal cortex that govern attention.

The model of Kuperberg (2007) is primarily based on EEG evidence and does not make strong claims on localization. Despite this, it does posit two parallel processes in language (semantic memory and semantic combination), which have similarities to the dual-stream model and mirror two of the three components of the MUC model. Under this model, the P600 is due to the continued analysis required if the output of this combinatorial stream is incongruent with the output of the predictions of the semantic stream. The combinatorial stream processes two types of constraints: morphosyntactic information, which is used to resolve syntax errors, and semantic-thematic information, which can influence the N400 because it operates in parallel with the semantic memory stream. Processing of this constraint may continue after the N400 window if more combinatory analysis is needed.

As we have seen in this section, neuroimaging studies have concentrated on seeking answers about the neural organization of the language faculty, and the brain's processing of linguistic input. See Figure 3.1 for a summary. Mostly absent is the consideration of more specific details of languages, such as the representation of words, phrases, sentences, and the structures that underlie them. In the following sections we describe recent studies that combine computational modeling of language with recordings of brain activities to examine the finer grain of languages.

3.3 Decoding Words in the Brain

As mentioned in the previous section, architectural theories of language processing see the temporal lobe as central to the generic retrieval and processing of words. Less work has been devoted to characterizing the representation and processing of particular words or word classes.

There is a great deal literature on the representation of classes of *objects*, represented using images. For instance there is well-documented specialization for semantic class in the inferior temporal cortex, with areas of the fusiform gyrus in particular being differentially sensitive to living and nonliving things

(Martin et al., 1996). In the vision science community this is considered "high-level vision," where the low-level visual input has been abstracted away to the extent that these parts of the brain encode something about the meaning and content of an image (Connolly et al., 2012; Carlson et al., 2013). An obvious question then is whether the representations in these brain areas are indeed visual, or are rather amodal. One study demonstrated similar patterns of activity, specific to semantic categories in these "visual" areas, in congenitally blind participants (Mahon et al., 2009), suggesting that this area (conveniently located next to language semantic areas) may be the locus of amodal (and perhaps symbolic) meaning. A series of subsequent studies found commonalities in brain activity patterns evoked by words, and by their corresponding pictures, with left posterior temporal areas emerging as key to amodal decoding (Shinkareva et al., 2011; Devereux et al., 2013; Fairhall and Caramazza, 2013; Simanova et al., 2014).[2]

In recent years, embodied theories of meaning (Barsalou and Wiemer-Hastings, 2005) have challenged this classical position. A particularly influential study looked for evidence of the sensory/motor coding of common concepts. Pulvermüller (2005) demonstrated that passive reading of physical action verbs caused increases in brain areas responsible for controlling the corresponding body parts. For instance, reading the verb *lick* caused activation in the vicinity of the face sensory/motor area, *pick* in the hand area, and *kick* in the leg area. In addition to the strong claim that meaning was inherently modal, this added to the evidence that conceptual information, whatever its content, was coded in a distributed fashion across the brain (Haxby et al., 2001; Marslen-Wilson et al., 2001; Martin, 2007), and is also consistent with the position that the brain uses both amodal and modality-specific representations when processing semantics (Fernandino et al., 2016; Handjaras et al., 2016).

Most studies of this type have an inherent practical limitation. Conventional experiment designs, and the cost involved in collecting brain data, both conspire to restrict us to small numbers of stereotyped concept types (such as the overstudied animals and tools [e.g., Murphy et al., 2011]). Because an adult vocabulary consists of tens of thousands of entries, it is hard to see how such an approach could attempt to provide a comprehensive account of our mental lexicon.

A pioneering study by Mitchell et al. (2008) provides an alternative paradigm for studying language and other types of higher cognition. Rather than characterizing a cognitive faculty (in this case, language) in terms of a small number of constructed conditions, it attempts to describe the fine grain and the breadth

[2] Simanova et al. (2014) stands out by using spoken words and associated sounds (e.g., a dog's bark) in addition to the pictorial and written word stimuli used in most other studies.

of language. It takes advantage of models of language from computational linguistics and uses these as an intermediate feature description to establish the mapping between brain recordings and the stimuli and tasks that evoke them.

Mitchell set out to computationally describe the brain activity associated with single words and concepts. This posited, possibly distributed, pattern of brain activity is termed a *neural signature*. Recordings of brain activity are noisy, and practical limitations prevent us from collecting the large amounts of data that would be needed to directly discover the neural signature for a single word. Instead, an intermediate model describes the similarities and dissimilarities among the concepts of interest in terms of semantic dimensions. Whereas linguistics and lexicography have historically relied upon binary and categorial features (Katz and Fodor, 1963; Fillmore, 1982; Miller et al., 1990), more recently word spaces and word vectors (embeddings) have become broadly used as an empirically grounded description of word meaning (Lund et al., 1995; Landauer and Dumais, 1997; Collobert and Weston, 2008; Baroni and Lenci, 2010). Mitchell used a simplified vector space model (VSM), based on

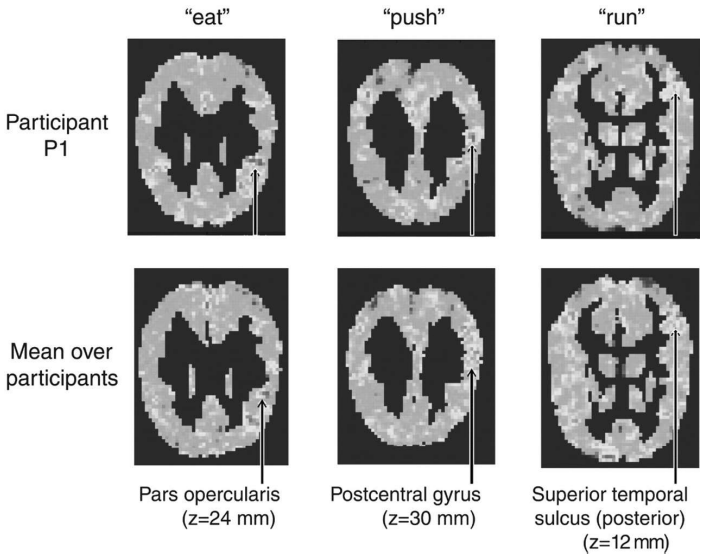


Figure 3.4 Several slices from fMRI images showing the *learned* proportion of brain activation that can be associated with a particular verb from the set of twenty-five verbs used in Mitchell et al. (2008). Note that verbs map onto areas of the brain previously shown to be associated with related semantic information. Figure courtesy of Mitchell et al. (2008).

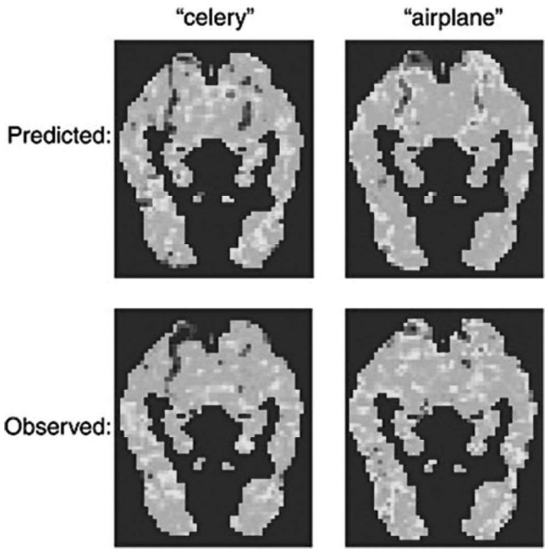


Figure 3.5 The predicted (*top row*) and observed (*bottom row*) brain activation for a particular person while reading the word "celery" or "airplane" (*left and right columns, respectively*). Although the brain images are not perfect matches, it is clear that the observed activity for celery is closer to the predicted activation for celery than airplane, and similarly the observed airplane activation is more similar to the predicted airplane activation. Figure adapted from Mitchell et al. (2008).

co-occurrence statistics with verbs encoding sensory-motor interaction, according to a large set of web *n*-gram statistics.

On the basis of this a regularized regression model was trained to find mappings between individual model features and the recordings of fMRI activity, in the process identifying components of brain activity with a semantic interpretation (Just et al., 2010). A side effect of discovering neural signatures for these semantic components was that such components can be recombined in novel ways to estimate the neural signature for unseen concepts (that is, words or images that were not part of the computational training set). Examples of semantic components discovered in brain activity are shown in Figure 3.4, whereas Figure 3.5 compares several observed neural signatures to the estimated reconstruction for those concepts (when they were unseen by the algorithm).

This made it possible to guess, based on a person's brain activity, which concept was being processed. Of course it is conceivable that such brain decoding performance could be driven by low-level confounds with semantics:

for instance artifacts tend to be more angular and less visually textured than natural kinds. But the use of a particular set of intermediate features, encoding semantics rather than perceptual characteristics, suggests that semantics is the driving factor. A follow-on study also demonstrated that decoding works across languages, e.g., training on fMRI data from English speakers, and decoding what concept a Portuguese speaker is thinking about (Buchweitz et al., 2012).

Follow-on studies established that this analytical paradigm worked with other semantic feature spaces, including crowdsourced semantic judgments (Palatucci, 2011), psychological feature norms (Chang et al., 2011), association norms (Akama et al., 2015), structured taxonomies (Jelodar et al., 2010), Wikipedia definitions (Pereira et al., 2010), and a broader range of word space embeddings (Devereux and Kelly, 2010; Murphy et al., 2012b; Bullinaria and Levy, 2013; Fyshe et al., 2014). Similar analyses have been performed with EEG (Murphy et al., 2009; Simanova et al., 2010) and MEG (Sudre et al., 2012); using videos as stimuli in place of words or images (Huth et al., 2012); for other classes of words beyond the small number of concrete categories used in the original experiments (Anderson et al., 2014); and reversing the direction of inference, so that semantic features can be "read out" of brain images in an approximate fashion (Pereira et al., 2011). Across all of these studies of words in the brain, it is noted that semantic representations are distributed throughout the cortex, and there is no singular locus of lexical meaning.

3.4 Phrases in the Brain

Once we can detect the neural signatures of individual words, we can study how they combine in the brain to form composed phrases. There have been several studies on adjective phrase processing by normal adults, including MEG studies that have implicated right and left anterior temporal lobes (RATL and LATL) as well as ventromedial prefrontal cortex (vmPFC). Adjective-noun pairs elicit increased neural activity when compared with word lists or nonwords paired with nouns, with activity significantly higher in LATL (∼200 ms), vmPFC (∼300–500 ms), and RATL (∼200 and 300–400 ms) (Bemis and Pylkkänen, 2011). When comparing a compositional picture-naming task with a noncompositional picture-naming task, Bemis and Pylkkänen (2013) found differences in the magnitude of activation in LATL. Bemis and Pylkkänen hypothesize that, due to the timing of these effects, the activity in vmPFC is related to semantic processes, and that LATL activity could be due to the either the syntactic or semantic demands of composition.

Semantic composition has also been studied using word vector space models. The computational linguistics community has proposed several methods of

combining word vectors to produce an estimate of the meaning of a phrase, including adding or multiplying the vectors together (Mitchell and Lapata, 2010; Erk, 2012). These studies have spurred several brain-imaging experiments that look for the composed semantic representation in the brain.

Adjective-noun composition in fMRI was explored by Chang et al. (2009) with twelve adjective-noun pairs and corpus-derived vectors composed of verb co-occurrence statistics, inspired by Mitchell et al. (2008). They showed that, in terms of R^2 (regression coefficient of determination, or variation explained), a multiplicative model of composition outperformed an additive composition model, and also the adjective or the noun's semantic vector. However, in terms of ranking the predicted brain activation under the learned model by distance to the true brain activation (as in Mitchell et al.), the additive, multiplicative, and noun-only model performed similarly. Fyshe et al. (2013) explored using these composed vector representations to decode the phrase from MEG data. They showed that a more complex function, "dilation," gave better decoding of phrases from MEG data. Dilation emphasizes the dimensions of the noun that are shared with the adjective, which is a plausible metaphor for the composition of some adjective-noun phrases.

Baron and Osherson (2011) studied the semantic composition of adjective-noun phrases using fMRI. Here, the stimuli was not linguistic, but rather was the faces of young or old males (boys and men) and young or old females (girls and women). In the fMRI scanner, the faces were presented in the same order for several minutes (time block). For each time block within the experiment, subjects were given a category (e.g., girl) and were asked to determine whether each of the stimuli faces was a member of that category. Thus, for each block, the face stimuli was constant, and only the concepts being matched (e.g., young and female) differed. Thus, any differences in activation can be attributed only to the matching task, and not to the stimuli. Baron and Osherson then created conceptual maps by learning models to predict brain activity based on the age (young or old) or gender of the matching task. They found that the brain activation of a composed concept (e.g., young male) could be estimated by the multiplication or addition of adjective (e.g., young) and noun (e.g., boy) brain activation maps. Areas of the brain that could be approximated well with an additive function were widespread and covered frontal, parietal, occipital, and temporal lobes, whereas the multiplicative function was useful for predicting just to the LATL.

How do the experimental results for semantic composition relate to the models of semantic unification previously discussed? If the syntactic form is held constant (as in adjective-noun phrases), the brain processes for syntactic combination are identical. However, when the semantics of the phrase changes, the semantic retrieval/memory and unification processes will also change, resulting in differential brain activity.

In Bemis and Pylkkänen's work, the semantic content of the words is constant, but the task differs (combining words into phrases or not). Their findings show increased activation in LATL, RATL, and vmPFC, which implies that the combinatorial processes of adjective-noun composition are at least partially handled in these areas. This is consistent with the local structure-building process in Figure 3.1, involving ventral pathway II. Hickok and Poeppel (2007) also hypothesize that the anterior temporal lobe is involved in composition, although they localize it to a slightly more medial temporal location.

Recall from our discussion of single words in the brain that semantics is distributed throughout the cortex. This finding implies that the semantic portion of semantic composition will likely occur in many places in the brain, even if composition is mediated by areas of the temporal lobe. This is congruent with the additive model of Baron and Osherson (2011). Perhaps the temporal lobe (indicated by multiple studies for the syntactic processes of composition) acts like the conductor of an orchestra, and each of the distributed semantic areas is an instrument. Signals are sent by the conductor to raise or lower particular elements of the orchestra, or to cause specific areas to play in synchrony. This is a metaphor for the way the brain could encode changes in semantics due to composition, bringing the activation of brain areas up or down, or causing areas to work in synchrony to encode meaning altered by context.

3.5 Stories in the Brain

Story processing is emerging as a new, more ecologically valid way to study the finer grain of language in the brain. Using authentic materials allows us to simultaneously examine the multiple levels of processing engaged in language comprehension. By moving to the narrative we can also investigate discourse factors (e.g., keeping track of events, characters' perspectives, reader response) beyond the levels of syntax and semantics seen in sentence studies.

Mason and Just (2006) review many studies involved in processing narrative, and identify networks involved in story understanding: a coarse semantic processing network (in the right temporal lobe), a topical coherence monitoring network in the bilateral frontal lobes, a text integration network in the anterior temporal and inferior frontal cortices. They also identify networks that process narrative information: a "protagonist's perspective interpreter network" in the right superior temporal cortex and the medial frontal lobes, and an imagery network in the bilateral intraparietal sulcus. In Speer et al. (2009), changes along different narrative dimensions were manually annotated (e.g., changes in goal, time, character identity, or location). Different regions in the brain had their activity correlated with these dimensions. Notably, the change in character identity also correlated with posterior lateral and medial frontal activity along with other temporal regions. This experiment and others mentioned in

Mason and Just (2006) mostly used hand-labeled features instead of computational language models used for text annotation. Story comprehension tasks have also been used to examine how the implied affect of individual words relates to that of phrases and larger passages (Hsu et al., 2015).

Other studies of story in fMRI focus on language processing instead of narrative structure, and many of them focus specifically on syntactic processing, and the cognitive load imposed by structures of varying complexity. In Bachrach (2008), tailored stories are used to auditorily present complex syntactic structures with higher frequency than average. Multiple measures of syntactic structure are computed and found to correlate with the activity of multiple brain regions in the temporal and inferior frontal cortices, including the left anterior temporal lobes. Theory-of-mind features were also used, and, once again, found to be correlated with the activity in the posterior temporal cortex. In Brennan et al. (2010), syntactic load is measured by building a parse tree of every sentence and computing the tree depth of each word. This syntactic feature also predicts the activity in the anterior temporal lobe, suggesting that the anterior temporal lobe is involved in the structural aspect of sentence composition as well as the semantic aspect, or that the two might be hard to disentangle. Mechanisms of memory, anticipation, and information structure are also being studied intensively (Hale et al., 2015; van Schijndel et al., 2015; Frank et al., 2015).

In Wehbe et al. (2014b), subjects read a chapter of J. K. Rowling's fantasy novel *Harry Potter and the Sorcerer's Stone* in the fMRI scanner in rapid serial visual presentation mode. Computational language models were used to label the words of the chapter. A semantic feature space of the words was constructed using non-negative sparse embedding (Murphy et al., 2012a), and syntactic features were obtained by identifying the part of speech and grammatical role of the words using the MaltParser (Nivre et al., 2007). Multiple regions spanning the bilateral temporal cortices were found to represent syntax or semantics, and sometimes both, hinting to the possibility that syntax and semantics might be nondissociated concepts. Other hand-labeled features were identified that characterized narrative components such as the presence of different story characters (which corresponded to activity in classical theory-of-mind areas) and their physical motions (which corresponded to activity in regions also activated during the perception of biological motion). The results are shown in Figure 3.6.

Huth et al. (2016) had subjects listen to hours of narrated stories (from *The Moth Radio Hour* series of podcasts). The authors built a VSM based on word co-occurrences with a set of nine hundred and eighty-five frequently used English words. A generative model was estimated that predicts brain activity as a function of this semantic VSM, and it was able to predict a considerable portion of the variance of activity across all regions typically referred to as the semantic system, spanning most of the temporal cortices, parts of the parietal

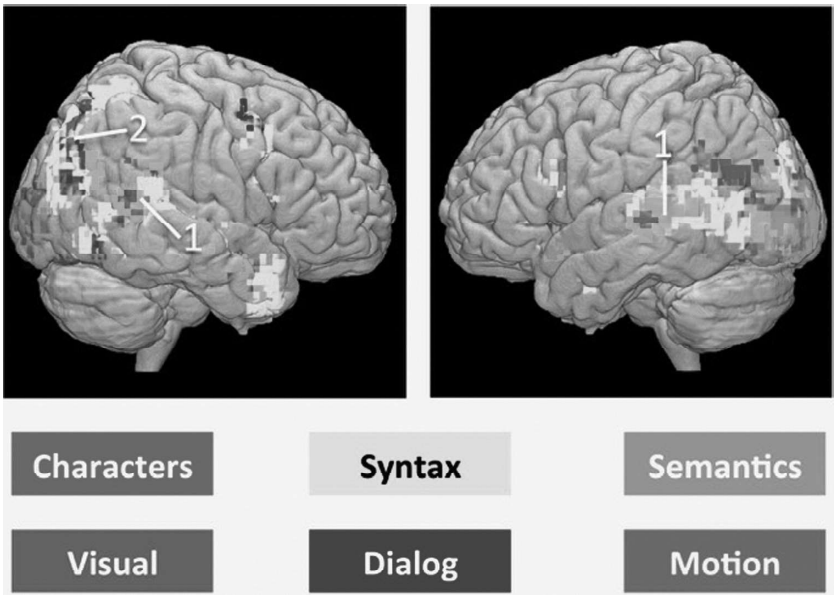


Figure 3.6 Story reading brain map, adapted from Wehbe et al. (2014b). "Results obtained by [...] a generative model, showing where semantic, discourse, and syntax information is encoded by neural activity. Note this model identifies not just where language processing generates neural activity, but also what types of information are encoded by that activity. Each voxel location represents the classification done using a cube of $5 \times 5 \times 5$ voxel coordinates, centered at that location, such that the union of voxels from all subjects whose coordinates are in that cube are used. Voxel locations are colored according to the feature set that can be used to yield significantly higher than chance accuracy. Light green regions, marked with (1), are regions in which using either semantic or syntactic features leads to high accuracy. Dark gray regions, marked with (2), are regions in which using either dialog or syntactic features leads to high accuracy."

cortices, and the frontal cortices. Importantly, and in conjunction with the previously mentioned studies, the uncovered semantic representation was highly bilateral. Furthermore, the authors constructed an atlas of semantic representations by combining data across all their subjects using a model that accounts for individual variations. They were able to identify a large set of semantically selective areas that each encode a constrained set of concepts. One of their findings is that the bilateral posterior temporal cortex is highly responsive to words related to social interaction. Previously we saw this region activated by tasks related to keeping track of the protagonist perspective.

Another method for studying language in a naturalistic scenario is to find voxels that are highly correlated across subjects in various conditions. Namely,

Lerner et al. (2011) contrasted intersubject voxel correlations when listening to paragraphs, scrambled sentences, scrambled words, and backward speech. The authors found the activity was consistent among subjects in each condition in a set of increasingly larger brain regions that are hierarchically organized: in the simplest condition (backward speech), the activity was consistent mostly in the primary auditory cortex. As the temporal integration window became longer when moving to words, sentences, and paragraphs, the consistent region became larger and encompassed a larger and larger part of the middle temporal cortex, eventually spreading to the posterior temporal cortex, the temporoparietal junction, and the inferior frontal gyrii. This study highlights the importance of studying the difference between the representation of the combined meaning of words and the representation of the meaning of these words in isolation.

To investigate this question, the same *Harry Potter* experiment was performed in MEG by Wehbe et al. (2014a). This experiment studied the representations of the properties of a word versus the properties of the context that preceded it, and tried to identify the different stages of continuous meaning construction when subjects read a text. The researchers used a recurrent neural network language model (Mikolov, 2012) to obtain feature representations for the context of a word (computed before the word appeared) and the features of

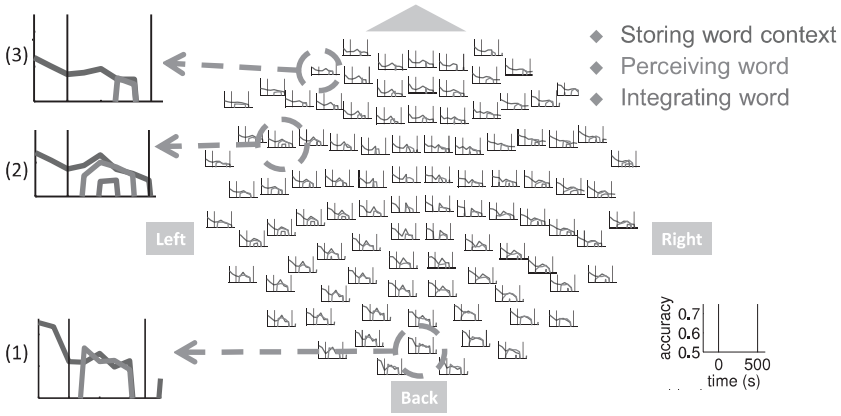


Figure 3.7 Adapted from Wehbe et al. (2014a). Time-line of word integration across the MEG helmet. For each of the one hundred and two locations the average accuracy at that location is plotted versus time. The axes are defined in the rightmost, empty plot. Accuracy should be seen as an indicator for the underlying process. Three plots have been magnified to show the increasing delay in processing the incoming word as information travels through the brain from the visual cortex (1) to the temporal (2) and frontal (3) cortices. Furthermore, the integration process is shown to occur in the left and right temporal cortices, after the word perception begins. The word context is maintained before the word is shown, and is gradually updated as the word is perceived.

that word. The recurrent neural network language model iteratively combines consecutive words and predicts the incoming word while maintaining a hidden vector representation of context. The model was run on the words of the chapter and the context vector, as well as the vector specifying the properties of the current word, and a measure of how surprising that word was given its context was extracted at each step. These vectors were then used to predict brain activity. The results, shown in Figure 3.7, reveal that context is more predictive of brain activity than the properties of the incoming word, hinting that more brain activity might be involved in representing context. Furthermore, the results include a suggested time line of how the brain updates its context representation. They also demonstrate the incremental perception of every new word starting early in the visual cortex, moving next to the temporal lobes, and finally to the frontal regions. Lastly, the results suggest that the integration process occurs in the temporal lobes after the new word has been perceived.

3.6 Summary

Brain activity data is the most direct record we have of the psychological states and processes that underlie language function. In this chapter, we reviewed earlier work that studied neuroimaging data using the tools of experimental psychology. This approach has provided new insights into the macrofunction and architecture of the language faculty, but can be limited in the generality and detail it can describe. Because brain data is difficult to obtain, statistical and machine learning methods that rely on large amounts of data have limited traction. This means that computational linguistics is an ideal and complementary tool, leveraging the huge quantities of naturalistic text available on the web to build detailed models and instantiate fine-grained theories. The use of computational features of text as an intermediate description enables models of greater generality that may draw conclusions beyond the limited set of stimuli that can be presented in a single experiment.

Naturalistic experimental tasks can also play a role, again letting computational models resolve the multifactorial complexity of human language comprehension. Authentic texts as stimuli do present considerable challenges to analysis (e.g., the systematic confounding among syntax and semantics, and Zipfian distribution of words and phrase categories [see, e.g., Hasson and Egidi, 2013]), for which methodological solutions continue to emerge. At the same time they are excellent from the point of view of engaging participants in authentic language processing, and are more representative of real-world language experience than many hand-tailored materials. This more holistic approach to understanding language processing is also consistent with an increasing tendency in neuroscience to understand brain activity in terms of networks and interactions among functional units.

Looking to the future, we expect to see continued progress, driven by larger and more varied datasets, and more powerful learning algorithms (as in computational linguistics, deep learning methods are beginning to impact neuroscience, e.g., Koyamada et al., 2014; Zheng et al., 2014). As methods and data improve, we hope to gain greater insight into the fundamental questions of linguistics, such as the universality of representations (Zinszer et al., 2015), the (in)dependence of syntax and semantics, and question of how language knowledge is encoded (Handjaras et al., 2016) and interacts with real-world and procedural knowledge.

Data sharing in neuroscience is increasing, driven both from grassroots (Yarkoni et al., 2010) and by the policies of governments and other funders. Another area of rapid change is analyses that make use of multiple modalities of data, combining recordings of brain activity with textual data (Fyshe et al., 2014), eye-gaze tracking (Desai et al., 2016; Henderson et al., 2016), and collections of natural images (Khaligh-Razavi and Kriegeskorte, 2014; Clarke et al., 2015; Anderson et al., 2015). And finally, advances in analysis may help our understanding of individual variation in language processing (see, e.g., Charest et al., 2014), outside the lab as well (Kidmose et al., 2013).

References

Akama, Hiroyuki, Miyake, Maki, Jung, Jaeyoung, and Murphy, Brian. 2015. Using graph components derived from an associative concept dictionary to predict fMRI neural activation patterns that represent the meaning of nouns. *PloS one*, **10**(4), e0125725.

Anderson, Andrew J., Murphy, Brian, and Poesio, Massimo. 2014. Discriminating taxonomic categories and domains in mental simulations of concepts of varying concreteness. *Journal of Cognitive Neuroscience*, **26**(3), 658–81.

Anderson, Andrew James, Bruni, Elia, Lopopolo, Alessandro, Poesio, Massimo, and Baroni, Marco. 2015. Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induced by written text. *NeuroImage*, **120**, 309–322.

Bachrach, Asaf. 2008. "Imaging Neural Correlates of Syntactic Complexity in a Naturalistic Context." PhD thesis, Massachusetts Institute of Technology.

Baron, Sean G., and Osherson, Daniel. 2011. Evidence for conceptual combination in the left anterior temporal lobe. *NeuroImage*, **55**(4), 1847–52.

Baroni, Marco, and Lenci, Alessandro. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, **36**(4), 673–721.

Barsalou, Lawrence W., and Wiemer-Hastings, Katja. 2005. Situating abstract concepts. Chap. 7, pages 129–163, in Pecher, D., and Zwaan, R. (eds), *Grounding Cognition: The Role of Perception and Action in Memory Language and Thinking*. Cambridge, UK: Cambridge University Press.

Bear, Mark F., Connors, Barry W., and Paradiso, Michael A. 2007. *Neuroscience: Exploring the Brain*. 3rd ed. Baltimore: Lippincott Williams & Wilkins.

Marslen-Wilson, William, Tyler, Lorraine Komisarjevsky, and Moss, Helen E. 2001. Towards a distributed account of conceptual knowledge. *Trends in Cognitive Sciences*, **5**(6), 244–252.

Martin, Alex. 2007. The representation of object concepts in the brain. *Annual Review of Psychology*, **58**(1), 25–45.

Martin, Alex, Wiggs, Cheri L, Ungerleider, Leslie G, and Haxby, James V. 1996. Neural correlates of category-specific knowledge. *Nature*, **379**(Feb), 649–652.

Mason, R. A., and Just, M. A. 2006. Neuroimaging contributions to the understanding of discourse processes. Pages 765–799 of: Traxler, M., and Gernsbacher, M. A. (eds), *Handbook of Neuropsychology*. Amsterdam: Elsevier.

McCandliss, Bruce D., Cohen, Laurent, and Dehaene, Stanislas. 2003. The visual word form area: Expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences*, **7**(7), 293–299.

Mikolov, Tomas. 2012. "Statistical Language Models Based on Neural Networks. PhD thesis, Czech Republic: Brno University of Technology".

Miller, George A., Beckwith, Richard, Fellbaum, Christiane, Gross, Derek, and Miller, Katherine. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, **3**(4), 235–244.

Mitchell, Jeff, and Lapata, Mirella. 2010. Composition in distributional models of semantics. *Cognitive Science*, **34**(8), 1388–429.

Mitchell, Tom M., Shinkareva, Svetlana V., Carlson, Andrew, Chang, Kai-Min, Malave, Vicente L., Mason, Robert A., and Just, Marcel Adam. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, **320**(5880), 1191–5.

Murphy, Brian, Baroni, Marco, and Poesio, Massimo. 2009. EEG responds to conceptual stimuli and corpus semantics. Pages 619–627 of: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*.

Murphy, Brian, Poesio, Massimo, Bovolo, Francesca, Bruzzone, Lorenzo, Dalponte, Michele, and Lakany, Heba. 2011. EEG decoding of semantic category reveals distributed representations for single concepts. *Brain and Language*, **117**(1), 12–22.

Murphy, B., Talukdar, P., and Mitchell, T. 2012a. Learning Effective and Interpretable Semantic Models using Non-Negative Sparse Embedding. In: *International Conference on Computational Linguistics (COLING 2012)*, Mumbai, India.

Murphy, Brian, Talukdar, Partha, and Mitchell, Tom. 2012b. Selecting Corpus-Semantic Models for Neurolinguistic Decoding. Pages 114–123 of: *First Joint Conference on Lexical and Computational Semantics (*SEM)*.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kubler, S., Marinov, S., and Marsi, E. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, **13**(2), 95.

Özyürek, Asli, Willems, Roel M., Kita, Sotaro, and Hagoort, Peter. 2007. On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience*, **19**(4), 605–616.

Palatucci, Mark M. 2011. "Thought Recognition: Predicting and Decoding Brain Activity Using the Zero-Shot Learning Model." PhD thesis, Carnegie Mellon University.

Pereira, Francisco, Botvinick, Matthew, and Detre, Greg. 2010. Learning semantic features for fMRI data from definitional text. In: Murphy, Brian, Korhonen, Anna, and Chang, Kevin Kai-Min (eds), *1st Workshop on Computational Neurolinguistics*.

Pereira, Francisco, Detre, Greg, and Botvinick, Matthew. 2011. Generating text from functional brain images. *Frontiers in Human Neuroscience*, **5**(August), 1–11.

Pulvermüller, Friedemann. 2005. Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, **6**, 576–582.

Rayner, Keith. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, **124**(3), 372–422.

Salmelin, Riitta. 2007. Clinical neurophysiology of language: The MEG approach. *Clinical Neurophysiology*, **118**(2), 237–54.

Shinkareva, Svetlana V., Malave, Vicente L., Mason, Robert A., Mitchell, Tom M., and Just, Marcel Adam. 2011. Commonality of neural representations of words and pictures. *NeuroImage*, **54**(3), 2418–25.

Simanova, Irina, van Gerven, Marcel, Oostenveld, Robert, and Hagoort, Peter. 2010. Identifying object categories from event-related EEG: Toward decoding of conceptual representations. *PloS one*, **5**(12), e14465.

Simanova, Irina, Hagoort, Peter, Oostenveld, Robert, and Van Gerven, Marcel A. J. 2014. Modality-independent decoding of semantic information from the human brain. *Cerebral Cortex*, **24**(2), 426–434.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. 2008. Cheap and fast – but is it good?: Evaluating non-expert annotations for natural language tasks. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 254–263.

Speer, N. K., Reynolds, J. R., Swallow, K. M., and Zacks, J. M. 2009. Reading stories activates neural representations of visual and motor experiences. *Psychological Science*, **20**(8), 989–999.

Sprouse, J., Schütze, C. T., and Almeida, D. 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001-2010. *Lingua*, **134**.

Sudre, Gustavo, Pomerleau, Dean, Palatucci, Mark, Wehbe, Leila, Fyshe, Alona, Salmelin, Riitta, and Mitchell, Tom. 2012. Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, **62**(1), 463–451.

van Schijndel, M., Murphy, B., and Schuler, William. 2015. Evidence of syntactic working memory usage in MEG data. In: *Proceedings of the Sixth Workshop on Cognitive Modeling and Computational Linguistics*.

Wehbe, Leila, Vaswani, Ashish, Knight, Kevin, and Mitchell, Tom. 2014a. Aligning context-based statistical models of language with brain activity during reading. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Wehbe, Leila, Murphy, Brian, Talukdar, Partha, Fyshe, Alona, Ramdas, Aaditya, and Mitchell, Tom. 2014b. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, **9**(11), e112575.

Willems, Roel M. (ed). 2015. *Cognitive Neuroscience of Natural Language Use*. Cambridge, UK: Cambridge University Press.

Yarkoni, Tal, Poldrack, Russell A., Van Essen, David C., and Wager, Tor D. 2010. Cognitive neuroscience 2.0: Building a cumulative science of human brain function. *Trends in Cognitive Sciences*, **14**(11), 489–496.

Zheng, Wei-Long, Zhu, Jia-Yi, Peng, Yong, and Lu, Bao-Liang. 2014. EEG-based emotion classification using deep belief networks. Pages 1–6 of: *IEEE International Conference on Multimedia and Expo*.

Zinszer, Benjamin D., Anderson, Andrew J., Kang, Olivia, and Wheatley, Thalia. 2015. How speakers of different languages share the same concept. Pages 2829–2834 of: *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

4 Graph Theory Applied to Speech: Insights on Cognitive Deficit Diagnosis and Dream Research

Natália Bezerra Mota, Mauro Copelli, and Sidarta Ribeiro

Abstract

In the past ten years, graph theory has been widely employed in the study of natural and technological phenomena. The representation of the relationships among the units of a network allow for a quantitative analysis of its overall structure, beyond what can be understood by considering only a few units. Here we discuss the application of graph theory to psychiatric diagnosis of psychoses and dementias. The aim is to quantify the flow of thoughts of psychiatric patients, as expressed by verbal reports of dream or waking events. This flow of thoughts is hard to measure but is at the roots of psychiatry as well as psychoanalysis. To this end, speech graphs were initially designed with nodes representing lexemes and edges representing the temporal sequence between consecutive words, leading to directed multigraphs. In a subsequent study, individual words were considered as nodes and their temporal sequence as edges; this simplification allowed for the automatization of the process, effected by the free software *SpeechGraphs*. Using this approach, one can calculate local and global attributes that characterize the network structure, such as the total number of nodes and edges, the number of nodes present in the largest connected and the largest strongly connected components, measures of recurrence such as loops of 1, 2, and 3 nodes, parallel and repeated edges, and global measures such as the average degree, density, diameter, average shortest path, and clustering coefficient. Using these network attributes we were able to automatically sort schizophrenia and bipolar patients undergoing psychosis, and also to separate these psychotic patients from subjects without psychosis, with more than 90% sensitivity and specificity. In addition to the use of the method for strictly clinical purposes, we found that differences in the content of the verbal reports correspond to structural differences at the graph level. When reporting a dream, healthy subjects without psychosis and psychotic subjects with bipolar disorder produced more complex

graphs than when reporting waking activities of the previous day; this difference was not observed in psychotic subjects with schizophrenia, which produced equally poor reports irrespective of the content. As a consequence, graphs of dream reports were more efficient for the differential diagnosis of psychosis than graphs of daily reports. Based on these results we can conclude that graphs from dream reports are more informative about mental states, echoing the psychoanalytic notion that dreams are a privileged window into thought. Overall these results highlight the potential use of this graph-theoretical method as an auxiliary tool in the psychiatric clinic. We also describe an application of the method to characterize cognitive deficits in dementia. In this regards, the *SpeechGraph* tools were able to sensitize a neuropsychological test widely used to characterize semantic memory, the verbal fluency test. Subjects diagnosed with Alzheimer's dementia were compared to subjects diagnosed with moderate cognitive impairment, either with amnestic symptoms only or with damage in multiple domains. Also studied were elderly individuals with no signs of dementia. The subjects were asked to report as many names of different animals as they could remember within one minute. The sequence of animal names was represented as a word graph. We found that subjects with Alzheimer's dementia produced graphs with fewer words and elements (nodes and edges), higher density, more loops of three nodes, and smaller distances (diameter and average shortest path) than subjects in the other groups; a similar trend was observed for subjects with moderate cognitive impairment, in comparison to elderly adults without dementia. Furthermore, subjects with moderate cognitive impairment with amnestic deficits only produced graphs more similar to the elderly without dementia, while those with impairments in multiple domains produced graphs more similar to the graphs from individuals with Alzheimer's dementia. Importantly, also in this case it was possible to automatically classify the different diagnoses only using graph attributes. We conclude by discussing the implications of the results, as well as some questions that remain open and the ongoing research to answer them.

4.1 Introduction

Every day when we wake up, before talking with other people, we talk with ourselves using inner speech to remember what day it is, where we are, to make plans about what to do in the next few minutes or hours, who we are going to meet, or what we are supposed to do. When we recognize this "inner speech"

as coming from ourselves, we may simply call it "thinking." However, sometimes this inner speech is not recognized as self, but rather as stimuli generated elsewhere; this is the basis of what we call psychosis. Sometimes past memories dominate this mental space, and we focus on past feelings of sadness, joy, fear, or anxiety. Past and future memories are mixed in these first moments even before any interaction with another person. This flow of memories and thoughts helps organize our actions and to soothe our anxiety and sadness, as we can plan future solutions to solve past problems. Organized, healthy mental activity allows old and new information to interact in order to support different actions that take experience into account in an integrated manner. But what happens with this flow of thoughts when we are unable to organize our inner space?

For centuries, psychiatry has described symptoms known as thought disorder that reflects disorganization of this flow of ideas, memories, and thoughts (Andreasen & Grove, 1986; Kaplan & Sadock, 2009). Those symptoms are related with psychosis, a syndrome characterized by hallucinations (when one perceives an object that does not exist; a sensorial perception without a real external object) and delusions (when one believes in realities that do not exist for other people; ideas or beliefs not real for their peers) (Kaplan & Sadock, 2009). There are many different causes for psychosis, such as the use of psychoactive substances or neurological conditions such as cerebral tumors or epilepsies. However, psychotic symptoms may occur without a clear cause, starting with a strange feeling or perception, getting worse, creating a confused reality hard to share even with the closest person, and causing major mental suffering.

In association with this strange reality, the patient can experience the feeling of fragmentation of thoughts, having difficulty to organize ideas or to follow a flow of memories, impacting the way to express what they are thinking or feeling, creating meaningless speech (symptoms known as "alogia," and "poor speech"). This frequently reflects a mental disorder known as schizophrenia. In other cases, the person may experience another aberrant organization of thought, with higher speed of mental activity, associating different memories and ideas (known as "flight of thoughts"), creating a speech with large amount of words (a symptom known as "logorrhea") that never reaches the main point. This pattern of thought disorder is common during the mania phase of bipolar disorder, a psychiatric condition mainly described by opposite mood cycles comprising depressive and manic phases. This speech pattern changes during depressive phases in the opposite direction (low speed of thought, fewer associations, fewer amount of words during speech). The speech content can reflect that strange psychotic reality on all those conditions with unlikely word association, but the organization of ideas reflected in the word trajectories reveals different directions of thought disorder, helping psychiatrists make differential

diagnosis between bipolar disorder and schizophrenia, predicting different life courses and cognitive impacts.

The description of these different patterns of thought organization perceived through language helped psychiatrists distinguish between two different pathological states and predict different life courses (with higher cognitive deficits for schizophrenia, first known as *Dementia Precox* [Bleuler, 1911]). However, recognizing these features subjectively requires a long-term professional training and adequate time with each patient to know each individual and avoid misjudgments. And even with the best evaluation conditions it is only possible to quantify those features subjectively, judging disease severity by grades on the psychometric scales such as BPRS and PANSS (Bech, Kastrup, & Rafaelsen, 1986; Kay, Fiszbein, & Opler, 1987). The differential diagnosis requires at least six months of observation during the first episode (First, Spitzer, Gibbon, & Williams, 1990), which means that the initial treatment may occur under considerable doubt regarding the diagnostic hypothesis. This lack of objective quantitative evaluation also negatively impacts the research strategies that aim to find biomarkers for complex psychiatric conditions (Insel, 2010).

Another condition that benefits from early diagnosis and correct interventions to prevent major cognitive damage is Alzheimer's Disease (AD) (Daviglus et al., 2010; Kaplan & Sadock, 2009; Riedel, 2014). Specific characterization of risk during preclinical AD requires specialized investigations and still challenges professionals in the field, due to a lack of a consensual description of each stage (Daviglus et al., 2010; Riedel, 2014). Failure to recognize AD early on can lead to a loss of opportunity to prevent cognitive decline (Daviglus et al., 2010; Riedel, 2014). In summary, the currently poor quantitative characterization of cognitive impairments related to pathological conditions such as psychosis or dementia hinders the early detection of these conditions. In this scenario, the new field of computational psychiatry has been proposing mathematical tools to better quantify behavior (Adams, Huys, & Roiser, 2015; Montague, Dolan, Friston, & Dayan, 2012; Wang, & Krystal, 2014).

To this end, natural language processing tools are particularly interesting. It is now possible to simulate the expert's subjective evaluation with better precision and reliability, either by quantifying specific content features such as semantic incoherence (Bedi et al., 2015; Cabana, Valle-Lisboa, Elvevag, & Mizraji, 2011; Elvevåg, Foltz, Weinberger, & Goldberg, 2007), or by analyzing the structural organization of word trajectories recorded from patients (Bertola et al., 2014; Mota et al., 2012; Mota et al., 2014).

4.2 Semantic Analysis for the Diagnosis of Psychosis

One useful tool used to characterize the incoherent speech characteristic of psychotic crises is called Latent Semantic Analysis (LSA) (Landauer & Dumais,

1997). The strange reality created during psychotic states impacts the coherence of the flow of words when patients express their thoughts freely, leading to improbable connections between semantically distant words within the same sentences.

LSA is based on a model that assumes that the meaning of each word is a function of its relationship with the other words in the lexicon (Landauer & Dumais, 1997). By this rationale, if two words are semantically similar, i.e., if their meanings are related, they must co-occur frequently in texts. It follows that if one has a large enough database of word co-occurrences in a large enough corpus of texts, it is possible to represent each word of that corpus as a vector in a semantic space, and their proximity in that space will be interpreted as semantic similarity (Landauer & Dumais, 1997).

When healthy subjects describe their normal reality, it is expected that they will use words that are semantically similar within the same text. However, when reality becomes bizarre, as typical of psychotic states, subjects are expected to use semantically distant words in sequence, thus building incoherent speech. That incoherence can be quantified as a measure of semantic distance between consecutive words or sets of words (for example, a set of words used in the same sentence). The more incoherent the speech, the larger the semantic distance between consecutive words or set of words. This was first shown for chronic patients with schizophrenia diagnosis (Elvevåg et al., 2007) and helped predict diagnosis in the prodrome phase, 2.5 years before the first psychotic crises (Bedi et al., 2015).

4.3 What Is a Speech Graph?

One way to quantify thought disorder is to represent the flow of ideas and memories reflected on the flow of words during a free speech as a trajectory and create a speech graph. A graph is a set of nodes linked by edges (formally defined as G=(N, E), being N={w1, w2, …, wn}and E={(wi, wj)} [Bollobas, 1998; Börner, Sanyal, & Vespignani, 2007]). The criteria determining how a link is established between two nodes define topological properties of these graphs that can be measured locally or globally. In the present case, each word is defined as a node, and the temporal sequence of words during a free speech is represented by directed edges (Mota et al., 2014) (Figure 4.1). From a speech graph we can objectively measure local and global features of the word trajectory that reflects the flow of thoughts during a free speech task (like when the subject reports a daily event, a past memory, or even a dream memory).

In the last decade, graph theory has been widely employed in the study of natural or technological phenomena (Boccaletti et al., 2006). By allowing the representation of the relationships among their units, the overall structure of a network can elucidate characteristics that could not be understood by

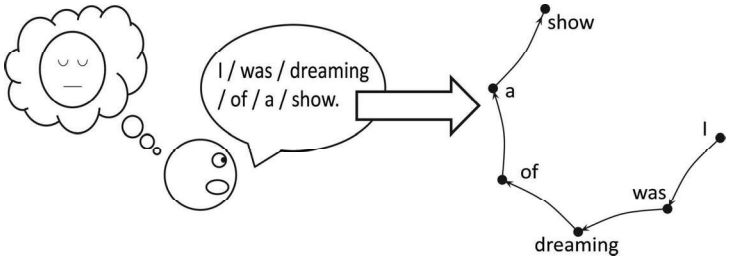


Figure 4.1 Examples of speech graphs from dream reports of schizophrenic, bipolar, and control subjects. Starting from transcribed verbal reports, graphs were generated using custom-made Java software (see the following text). Figure from Mota et al. (2014).

considering only a few units. The meaning of the represented structure basically depends on what is being considered as a node and on the definition of the presence and direction of edges (links between nodes). Graph theory as a tool not only may help tackle problems in the basic sciences but can also be applied to solve complex problems in everyday life, otherwise difficult to characterize and measure. An interesting strategy in scientific research is to keep both goals in focus: seek to understand a phenomenon at the fundamental level, while at the same time use the knowledge as a tool to solve practical problems (Stokes, 1997). With a simultaneous focus on basic and applied research, the application of graph theory to represent the relationship between spoken words helps understand how different psychiatric conditions differentially impact the flow of words during free speech, and how we can apply this knowledge to perform differential diagnosis.

Once reports are represented as graphs, one can calculate several attributes that quantify local and global characteristics. We calculated 14 attributes comprising 2 general graph attributes (Nodes and Edges), 5 recurrence attributes (Parallels – PE and Repeated Edges – RE; Loops of one – L1, two – L2 and three nodes – L3), 2 attributes of connectivity (Largest Connected Component – LCC and Largest Strongly Connected Component – LSC) and 5 global attributes (Average Total Degree – ATD, Density, Diameter, Average Shortest Path – ASP, Clustering Coefficient – CC) (Figure 4.2).

In order to compare graphs with different number of elements (controlling verbosity difference as measured by different amounts of words), two main strategies were used. First we divided each graph attribute by the amount of words in the report, assuming a linear relationship between graph attribute and verbosity. A pertinent critique is that the relationship between graph attributes and verbosity is not always linear, and for some attributes it is not clear if there is a direct relationship (Figure 4.3). A second strategy was to attribute a graph for

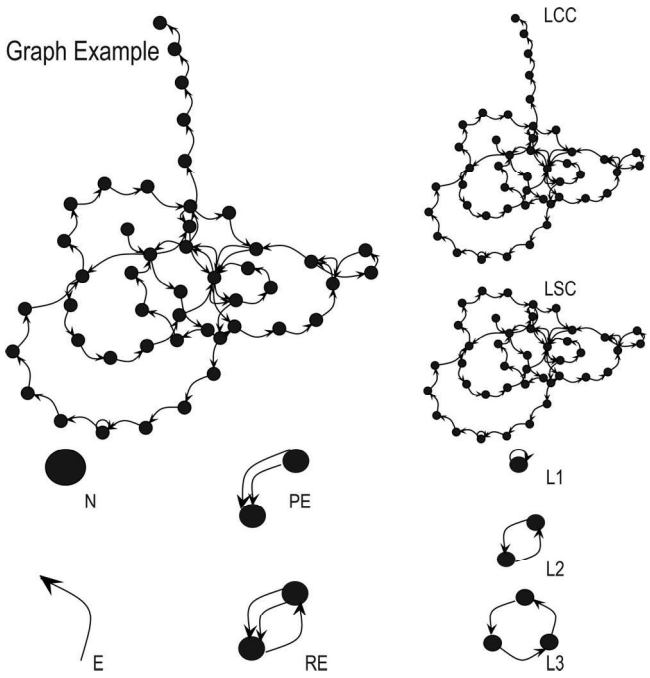


Figure 4.2 Examples of Speech Graph Attributes described earlier (figure from Mota et al., 2014).

Speech Graph Attributes:

1. **N:** Number of nodes.
2. **E:** Number of edges.
3. **RE (Repeated Edges):** sum of all edges linking the same pair of nodes.
4. **PE (Parallel Edges):** sum of all parallel edges linking the same pair of nodes given that the source node of an edge is the target node of the parallel edge.
5. **L1 (Loop of one node):** sum of all edges linking a node with itself, calculated as the trace of the adjacency matrix.
6. **L2 (Loop of two nodes):** sum of all loops containing two nodes, calculated by the trace of the squared adjacency matrix divided by two.
7. **L3 (Loop of three nodes):** sum of all loops containing three nodes (triangles), calculated by the trace of the cubed adjacency matrix divided by three.
8. **LCC (Largest Connected Component):** number of nodes in the maximal subgraph in which all pairs of nodes are reachable from one another in the underlying undirected subgraph. When you have all the words on one large connected component, LCC will be the same as N.

(*continued*)

each set of a fixed number of words, skipping an also fixed number of words to build the next graph, assuming a certain level of overlap between consecutive graphs. This "sliding window" approach allows calculating the average graph attributes of a graph with a fixed number of words. This enables the study of topological characteristics of graphs with different reports size (say, small, medium, and big graphs). A critique for this strategy is the arbitrary cut of word sequences that can change topological properties, mainly global attributes. This is an important discussion of ongoing research that needs to be addressed carefully, so as to enable a better interpretation of the results.

4.4 Speech Graphs as a Strategy to Quantify Symptoms on Psychosis

In an attempt to represent the flow of thoughts presented in a free speech, speech graphs were initially designed with nodes representing lexemes (a subject, object, or verb on the sentence), and their temporal sequence represented as directed edges, yielding directed multigraphs with self-loops and parallel edges (Mota et al., 2012). Analyzing dream reports represented as graphs from 24 subjects (8 subjects presenting psychotic symptoms with schizophrenia diagnosis, 8 subjects also with psychotic symptoms diagnosed as bipolar disorder in the mania phase and 8 control subjects without any psychotic symptom), it was possible to quantify psychiatric symptoms such as:

(Figure 4.2 caption continued)

9. **LSC (Largest Strongly Connected Component):** number of nodes in the maximal subgraph in which all pairs of nodes are reachable from one another in the directed subgraph (node a reaches node b, and b reaches a).
10. **ATD (Average Total Degree):** given a node n, its Total Degree is the sum of "in" and "out" edges. Average Total Degree is the sum of Total Degree of all nodes divided by the number of nodes.
11. **Density:** number of edges divided by possible edges (D = 2*E/N*(N-1)), where E is the number of edges and N is the number of nodes.
12. **Diameter:** length of the longest shortest path between the node pairs of a network.
13. **Average Shortest Path (ASP):** average length (number of steps along edges) of the shortest path between pairs of nodes of a network.
14. **CC (Average Clustering Coefficient):** given a node n, the Clustering Coefficient Map (CCMap) is the set of fractions of all n neighbors that are also neighbors of each other. Average CC is the sum of the Clustering Coefficients of all nodes in the CCMap divided by the number of elements in the CCMap.

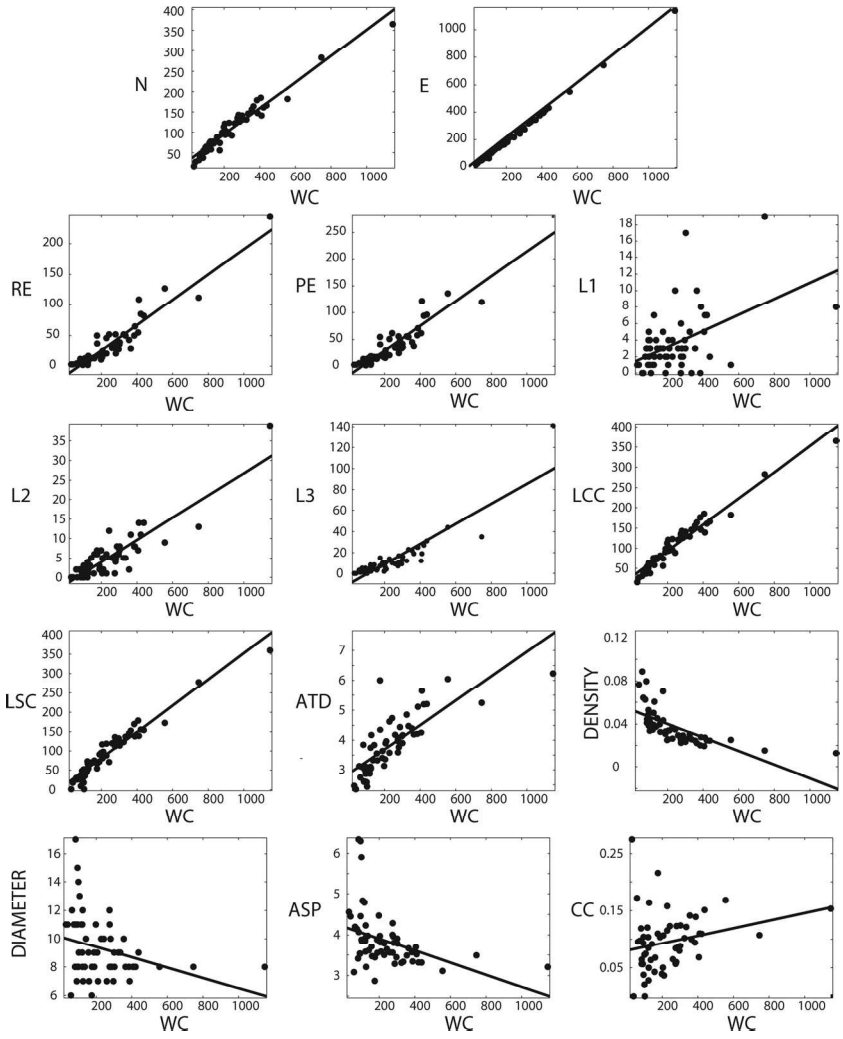


Figure 4.3 Linear correlation between SGA and word count (WC) (figure from Mota et al., 2014).

1. Logorrhea, described as the increase of verbosity characteristic of bipolar disorder on mania phase. This was quantified not only by counting more words in the bipolar group but also by more frequent recurrence (more parallel edges), even when controlling for differences in verbosity by dividing graph attributes by the amount of words in the speech. This means that the reports tend to return more often to the same topics.

2. Flight of thoughts, described as talking about other topics than the main topic asked, which is also characteristic of bipolar disorder. In the bipolar group, more nodes were used to talk about waking events upon request to report on a recent dream.
3. Poor speech, described as loss of meaning on the speech and perceived as a set of words that are poorly connected, characteristic of schizophrenia. This was quantified as more nodes per words, denoting reports that address the topics only once, neither branching nor recurring, so almost all the words used will count as a different node.

It was possible to automatically sort schizophrenia from bipolar group using a machine learning approach. A Naïve Bayes classifier was used to distinguish between both groups, and to distinguish between pathological groups and non-psychotic subjects (Kotsiantis, 2007). The classifier received as input either speech graph attributes or grades given from psychiatrists concerning psychiatric symptoms (using standard psychometric scales: PANSS [Kay et al., 1987] and BPRS [Bech et al., 1986]). Classification accuracy was assessed through the calculation of sensitivity, specificity, kappa statistics, and the area under the receiver operating characteristic curve (AUC), described as a plot of sensitivity (or true positive rate) on the y-axis versus false positive rate (or 1-specificity) on the x-axis. An AUC around 0.5 means a random classification, whereas AUC $= 1$ means a perfect classification (none of the possible errors were made). It was possible to classify the pathological groups against non-psychotic group using graph attributes and psychometric scales with high accuracy (AUC higher than 0.8) (Table 4.1). But to distinguish between schizophrenia and bipolar groups, graph attributes performed better than psychometric scales (AUC $= 0.88$ using graph attributes as input, while AUC $= 0.57$ when using psychometric scales as input) (Table 4.1).

This first study had some limitations concerning the low sample (only eight subjects per group) and the methodology. First, the transformation from a text to a graph was handmade, a process that is time consuming and has a higher risk of error. Second, the graph was not completely free of subject evaluation (a node

Table 4.1 *Classification metrics between diagnostic groups using SpeechGraph Attributes (Mota et al., 2012).*

	Sensitivity	Specificity	Kappa	AUC
S × B	93.8%	93.7%	0.88	0.88
S × C	87.5%	87.5%	0.75	0.90
B × C	68.8%	68.7%	0.37	0.80

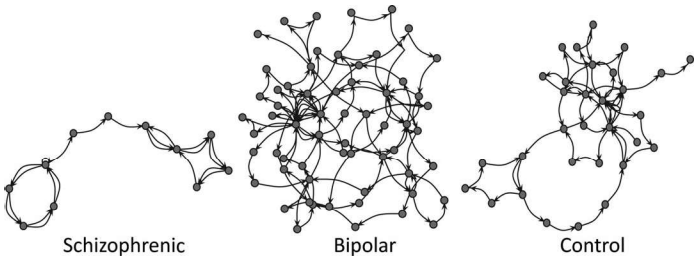


Figure 4.4 Representative speech graphs extracted from dream reports from a schizophrenic, a bipolar and a control subject (figure from Mota et al., 2014).

was considered as a subject, object, or verb on the sentence and, at a grammar level, it required a syntactic evaluation). So, in order to avoid these problems and to allow the study of a larger sample with larger texts, in a subsequent study we employed words as nodes and their temporal sequence as edges, a simplification that allowed the process to be automatized by the *SpeechGraphs* software (Mota et al., 2014). This custom-made Java software, available at http://neuro.ufrn.br/softwares/speechgraphs, receives as input a text file and returns the graph based on the text with all the 14 graph attributes described before. It is also possible to cut the text in consecutive graphs with a fixed number of words, controlling for verbosity and exploring different sizes of word windows to study cognitive phenomena.

To characterize distinct pathological phenomena in the speech of different types of psychosis, the *SpeechGraphs* tool was applied. Symptoms of Bipolar Disorder such as logorrhea could still be associated to the increase of the network size (Figure 4.4) (Mota et al., 2014; Mota et al., 2012). Also, symptoms of schizophrenia such as alogia and poor speech were measured as fewer edges (E), and smaller connected components (LCC) and strongly connected components (LSC) when compared to bipolar and control groups (Figure 4.4), producing less complex graphs in the schizophrenia group even after controlling for word count (comparing consecutive graphs of 10, 20, and 30 words with one word as step). In graphs from this group there are fewer edges between nodes and fewer nodes connected by some path or mutually reachable. This means that the schizophrenia group tends to talk only a few times about the same topic, not returning or associating past topics with consecutive ones, probably denoting cognitive deficits such as working memory deficits.

Using these network characteristics it was also possible to automatically sort the schizophrenia and bipolar groups, and those from subjects without psychosis, with AUC $= 0.94$ to classify schizophrenia and control groups, AUC $= 0.72$ to classify bipolar and control group, and AUC $= 0.77$ to classify schizophrenia and bipolar groups (Table 4.2). These results

Table 4.2 *Classification metrics between diagnostic groups using SpeechGraph Attributes (Mota et al., 2014).*

	AUC	Sensitivity	Specificity
S × B × C	0.77	0.62	0.81
S × B	0.77	0.69	0.68
S × C	0.94	0.85	0.85
B × C	0.72	0.74	0.75

highlight the potential use of this method as an auxiliary tool in the psychiatric clinic.

To better understand the relationship between these graph features and the symptomatology measured by psychometric scales, the correlation between those metrics was analyzed. Edges, LCC, and LSC were strongly negatively correlated with cognitive and negative symptoms (as measured by psychometric scales). In other words, when the subjects presented more severity on symptoms such as emotional retraction and flattened affect (loss of emotional reaction), poor eye contact (with the interviewer during psychiatric evaluation), loss of spontaneity or fluency on speech, and difficulty in abstract thinking (measured by the ability to interpret proverbs), their reported dreams generated graphs with fewer edges and fewer nodes on the largest connected and strongly connected component. Those psychiatric symptoms are more common in subjects with schizophrenia (Kaplan & Sadock, 2009), indicating how we can measure the impact on cognition and deficits in social interactions of these individuals through graphs of speech (Mota et al., 2014). Cognitive and psychological aspects that drive this pattern of speech such, as working memory, planning, and theory of mind abilities, may explain those deficits and help elucidate the pathophysiology of the different psychotic disorders. When the interviewer asks the subject to report a memory, the way the subject interacts socially with the interviewer and recalls what to report, planning the answer and the sequence of events to report, impacts the sequence of words spoken, reflecting their mental organization.

4.5 Differences in Speech Graphs due to Content (waking × dream reports)

We already understand that during pathological cognitive states there is an impact on the flow of thoughts or memories that we can track by the word trajectory. But what happens with physiologically altered consciousness states like dream mentation? Is it possible to characterize differences between dream

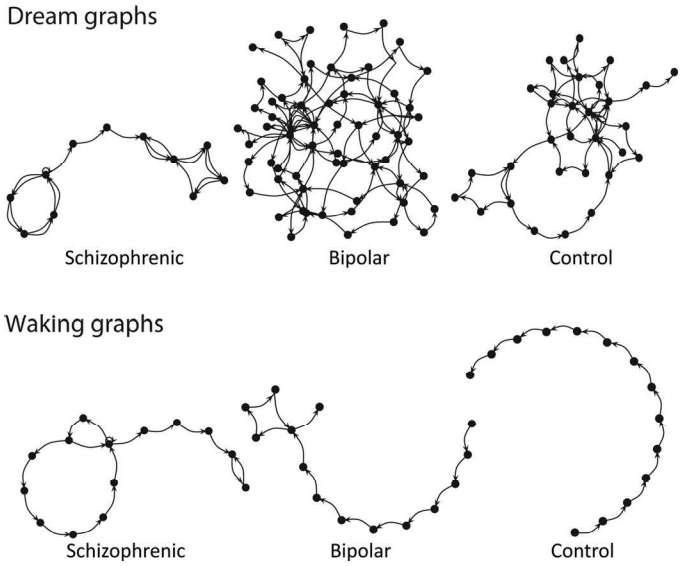


Figure 4.5 Representative speech graphs examples extracted from dream and waking reports from the same schizophrenic, bipolar, and control subject (figure from Mota et al., 2014).

and daily memories regarding word trajectories? Does it inform any additional features about general cognition?

A few minutes before waking up every day we can experiment an exclusively internal reality not shared with our friends or family: dreaming. This reality is internally built based on a set of memories with different affective valences, with different types of meaning only accessible by the dreamer. This confused mental state is phenomenologically similar to a psychotic state, as there is a lack of insight regarding the bizarreness of this strange reality (Dresler et al., 2015; Mota et al., 2014; Scarone et al., 2007). Thus it would not be surprising to expect that the flow of information regarding dream memories could better reveal thought disorganization characteristic of psychotic states.

During the studies with psychotic populations, there were differences in speech graphs depending on the speech content. When reporting a dream, subjects without psychosis and subjects with bipolar disorder produced more complex graphs (higher connectivity) than when reporting daily activities of the previous day, a difference that was not observed in subjects with schizophrenia (those subjects reported dreams or daily memories with the same few connected graphs) (Figure 4.5) (Mota et al., 2014). Therefore, graphs of dream reports were more efficient in group sorting than graphs of daily reports (Mota et al., 2014).

Another intriguing result was found in the correlations between speech graph attributes and clinical symptoms measured by psychometric scales PANSS (Kay et al., 1987) and BPRS (Bech et al., 1986). Only dream graphs connectivity attributes were strongly and negatively correlated with negative and cognitive symptoms (as measured by both scales) that are more common in schizophrenia. Waking report graphs showed negative correlations between general psychotic symptoms such as loss of insight (measured by PANSS) and incoherent speech (measured by BPRS) with LCC (also a connectivity attribute) (Mota et al., 2014). This emphasizes that reports of dream memories requires different cognitive functions and empathy abilities than reports of daily memories.

Based on these results we can conclude that graphs from dream reports are more informative about mental states than are graphs representing waking reports. This result echoes the psychoanalytic proposal that dreams are a privileged window into thought (Freud, 1900; Mota et al., 2014). This observation has started a new basic research approach to quantitatively understand what is going on when we remember a dream. The use of electrophysiological approaches (most notably, multichannel electroencephalography) to characterize different sleep stages in the laboratory allows the access to dream mentation by their reports at the same time that we access electrophysiological activity during sleep.

4.6 Speech Graphs Applied to Dementia

Considering the characterization of cognitive deficits in conditions such as dementia, the use of tests designed to characterize specific cognitive impacts on memory domain are useful in early evaluation. One example is the Verbal Fluency Test, which consists on verbal recall of different names of a specific category (usually animals) during a fixed time. This was first used to investigate the executive aspects of verbal recall, counting the capacity to produce an adequate quantity of words in a limited condition of recall, not repeating or recalling different categories (Lezak, Howieson, Bigler, & Tranel, 2012). The individual needs to access semantic memory correctly and to be flexible in order to quickly change the words (using temporal cortex structures), and to store the already mentioned words to avoid repetitions, which requires executive functions such as inhibitory control (using frontal cortex structures) (Henry & Crawford, 2004).

Different pathologies, such as dementia, can damage the performance on this task. As different structures are involved to correctly answer the task, different kinds of errors can help distinguish between different causes (damage in different locations). Different causes of dementia lead to different symptomatology

evolutions, which represent different location damages. The characterization of word trajectory with the application of the *SpeechGraph* tool complements this neuropsychological test (Bertola et al., 2014). A total of 100 individuals – 25 subjects diagnosed with Alzheimer's dementia, 50 diagnosed with Moderate Cognitive Impairment (25 of them with only amnestic symptoms and the others 25 with damage in multiple domains), and 25 elderly subjects with no signs of dementia – were asked to report as many names of different animals as they could remember in one minute (Nickles, 2001). The sequence of animal names was represented as a word graph.

It was observed that subjects with Alzheimer's dementia produced graphs with fewer words and elements (nodes and edges), higher density, more loops of three nodes and smaller distances (diameter and average shortest path) than did other groups, with the same trend for subjects with moderate cognitive impairment compared to elderly adults without dementia (Bertola et al., 2014). Furthermore, subjects with moderate cognitive impairment with only amnestic deficits produced graphs more similar to the elderly without dementia, while those with impairments in multiple domains produced graphs more similar to the graphs from individuals with Alzheimer's disease. Also in this case, it was possible to automatically classify the different diagnoses only from graph attributes (Bertola et al., 2014). There was also correlation between speech graph attributes and two important standard cognitive assessments widely used on geriatric population, denoting an important correlation between word trajectory on verbal fluency recall and general cognitive status (measured with MMSE – Mini Mental State Exam) and functional performance (measured with the Lawton Instrumental Activities of Daily Living Scale) (Bertola et al., 2014).

On one hand, the more cognitively preserved were the elderly, the more unique nodes were produced on less-dense graphs. On the other hand, the more functionally dependent the individuals were, the less words, nodes, and edges were produced on denser graphs with smaller diameter and average shortest paths (Bertola et al., 2014). Another differential impact was evident for three-node loops, a repetition of the same word with only two words in between (example: "lion," "cat," "dog," "lion"), found in higher frequency in the Alzheimer group compared with MCI and control groups (Bertola et al., 2014). This means an impairment in working memory since the early stages of the Alzheimer's disease (already recognized by other working memory assessments [Huntley & Howard, 2010]).

These results point to the additional information that the characterization of word trajectory brings to a well-established neuropsychological test. On this application example, as the test has restricted rules, we expect that the subject produces a certain type of graph, and different types of deviations from this expected pattern informs about cognitive impairments.

4.7 Future Perspectives

Word graphs are not the only tool to quantify psychiatric symptoms on speech analysis. As pointed out in the introduction, other approaches aim to quantify semantic similarities between words (Bedi et al., 2015; Elvevåg et al., 2007). The relationship between speech incoherence measured by LSA and speech structure measured by Speech Graphs is not clear yet. Both measures take into account word sequences and word co-occurrences, but with very different approaches (one compares with a semantic model based on a large corpus, and the other uses graph theory to characterize topological features of the speech sample). Understanding both approaches better can improve automated speech analysis for clinical purposes such as diagnosis and prognosis prediction, creating useful follow-up tools in a clinical set.

Other interesting perspective is to combine language analysis with prosody analysis. Semiautomated tools have characterized prosodic deficits related to schizophrenia diagnosis. The patients made more pauses, were slower, and showed less pitch variability and fewer variation in syllable timing, expressing a flat prosody when compared to matched controls (Martínez-Sánchez et al., 2015). The relationship between expressive prosody and language features during free speech can elucidate several cognitive characteristics subjectively perceived by well-trained psychiatrists (Berisha, Wang, LaCross, & Liss, 2015).

A better understanding of word trajectories in free speech can also be applied in settings other than the psychiatric clinic. As these tools show important correlations with cognitive deficits in psychosis and dementia, could it be useful to characterize cognitive development in a school setting? This kind of approach could help predict cognitive impairment early enough to allow quick intervention, preventing learning disabilities that later on would be harder to manage. This could also help quantitatively characterize cognitive development in a naturalistic manner.

Acknowledgments

The authors dedicate this chapter to the memory of Raimundo Furtado Neto, who made important contributions to the development of the SpeechGraphs software. This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), grants Universal 480053/2013-8 and Research Productivity 306604/2012-4 and 310712/2014-9; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) Projeto ACERTA; Fundação de Amparo à Ciência e Tecnologia do Estado de Pernambuco (FACEPE); FAPESP Center for Neuromathematics (grant # 2013/07699-0, S. Paulo Research Foundation FAPESP).

5 Putting Linguistics Back into Computational Linguistics

Martin Kay

Abstract

Almost all practitioners of natural language processing make a crucial error that also besets much of Chomsky's argument about the poverty of the stimulus in first language learning, namely that we can discover all we need to know about language by examining sufficiently large quantities of it. The error is to ignore the crucial function of language in referring to things in the real and imaginary worlds. Speakers and hearers appeal to this in many ways. Translators rely on it to provide information that they must supply in the target text but which is only implicit in the source. Reference is one of several properties of language that relate parts of a text or a discourse that may not be adjacent to one another. To the extent that linguists are concerned with how people use language for communication, they are interested in processes in people's heads and, to the extent that they are concerned with processes, they are interested in computation. It is this, rather that engineering feats like machine translation, that gives computational linguistics its special role.

5.1 Explicit and Implicit Information

Alfred Charles William Harmsworth, 1st Viscount Northcliffe (1865–1922), was the owner of two British newspapers, the *Daily Mail* and the *Daily Mirror*. He is credited with first pointing out that "Dog Bites Man" is an unlikely headline, whereas "Man Bites Dog" might herald a newsworthy story. Modern journalists are referring to the same phenomenon when they point out that no one ever writes a story about a plane because it did not crash. The point is obvious to journalists and to ordinary citizens seeking to avoid newsworthy flights. But it is missed by the adherents of Noam Chomsky's linguistic theories and by practitioners of a branch of linguistic engineering that has come to be known as *natural language processing* (NLP). This is curious because, but for an interest in language, these two groups have little in common.

Chomsky argues against the behaviorist view of language acquisition espoused, for example, by Skinner, on the grounds that even very large quantities of raw linguistic data contain neither the kind nor the quantity of information that would enable a general learning device to acquire everything that it would have to know to use the system as humans do. A device that could learn a language would have to come to the task with much knowledge of how language works already built in. Babies are clearly not born knowing English or Chinese, but they are born with brains that are specially adapted to the kinds of structure that are found in English and Chinese. This is the argument from what is called the *poverty of the stimulus*.

Children learn language almost entirely from positive examples – examples used appropriately and effectively in normal communication. Many of the very reasonable, but incorrect, hypotheses that they might entertain about the way the language works could only be rejected on the basis of specific negative information. By way of illustration, he points out that there are two ways in which an English learner might suppose the sentence *The man who is hungry is ordering dinner* might be turned into a question. One is to move the second occurrence of the auxiliary verb *is* to the front, giving the expected result: *Is the man who is hungry ordering dinner?* But, if questions are formed by moving auxiliaries to the front, then why not move the earlier occurrence of *is* in this sentence? This gives *Is the man who hungry is ordering dinner?* But this is a mistake that children do not make. The claim is that the raw data does not contain the information that would be needed to reject it, so that some, at least, of it must be acquired from another source. We take note, in passing, of the unquestioned assumption underlying this example, that the two sentences – the assertion and the question – are related by an operation that transforms one into the other.

The argument from the poverty of the stimulus has had many and persuasive critics. The details are beyond the scope of this essay. More interesting to us is another argument about how children acquire language that receives very little attention, in part, perhaps, because it is so obvious. For all that children learn language easily, they would presumably not do so if the effort did not pay off in some fairly direct way. To be worth learning, language needs to be good for something. An unfortunate child that was confined to a dark room where it heard English sentences through a loudspeaker would presumably learn little English. A child learns to say *doggie* when it sees a dog, and is indeed corrected if it says *kitty cat* instead. It says *cookie* when it wants a cookie, sometimes with a definite payoff.

As Ferdinand de Saussure (1906–1910) famously pointed out, the link between a word and the things it can be used to refer to is completely arbitrary in all but a few very special cases. The Portuguese word *puxe* is pronounced substantially like the English word *push*, but it means *pull*. There would be no way to learn this from a loudspeaker in a dark room. The importance of

this for Chomsky's linguistics is considerable, but it is also not central to our present concerns. However, it is also of crucial importance for natural language processing, a great deal of which is based precisely on the idea that the knowledge required for many language-processing tasks, such as machine translation, can be learned entirely automatically from studying sufficiently large amounts of text. That most important function of language, the referential function, is ignored entirely.

There is, of course, an obvious and very serious objection to this line of argument, namely that the referential function, though clearly central to our everyday use of language, does not need to be provided for explicitly in purely linguistic activities such as translation. A French sentence that contains the word *chien* is probably talking about a dog. This function is filled in English by the word *dog*. So, translating the French sentence into English involves at least three entities: the French word, the English word, and the dog that connects them. However, crucial though the dog is to the understanding of the sentences, it is not crucial to the translation process. If the two words are connected through the animal, then they are connected, and we can do as second language learners routinely do and say simply the *dog* means then same thing as *chien*. We are, of course simplifying the story greatly. Words often have several meanings and thus several translations, and we must consider the context in order to decide which is in play in any particular place. The underlying argument, however, remains strong and must be taken seriously. Our claim, however, will be that it is false.

In what follows, we concentrate mainly on translation, but similar considerations apply equally to other kinds of language processing. The claim that the referential function of language does not need to be addressed explicitly in translation rests on the notion that translation is an essentially *syntactic* phenomenon. In other words, it consists in (1) deciding what lexical items the words and phrases of the source text correspond to, (2) finding words and phrases in the target language with corresponding lexical items, and (3) putting these words and phrases in an appropriate order. For these purposes, we can think of a lexical item as being essentially a pair consisting of a word or phrase in each of the languages. This model is implicit in almost all second language teaching, at least in the early years, and all work on machine translation, whether by practitioners of the new NLP or by linguists of the older rule-based tradition. It is in contrast with what we can refer to as *pragmatic* tradition, in which the translator adopts a position similar to that of the original author and, having understood what that author is trying to say, seeks a way of saying it in the target language, without attempting to relate individual words and phrases in one language with words and phrases in the other.

The work of most modern journeyman translators is largely syntactic, but problems requiring a pragmatic solution are quite common, and pragmatic solutions are often adopted even when syntactic solutions are readily at hand. When

mainline trains arrive at Paris, the loudspeakers generally address the passengers with the words: *Assurez-vous que vous n'avez rien oublié dans le train*, which can be translated syntactically into English as *Make sure that you have not forgotten anything on the train*. A marginally more idiomatic rendering would probably be *Make sure that you have not left anything on the train*, which is a slight concession to pragmatic translation, because the English verbs *forget* and *leave* would not normally be said to mean the same thing. They are equivalent in the present context because what they *refer* to is the same, namely articles that remain on the train even though the passengers intended to take them off with them.

As a matter of fact, the English message on the loudspeakers in Paris is often neither of these, but rather *Make sure that you take all your belongings with you*, which is clearly pragmatic. There is nothing in the original about belongings or about passengers having anything with them. If a passenger was carrying something for someone else, it is not clear that it should be counted among that passenger's belongings. A passenger could have left nothing on the train, but left it somewhere on the platform or had it stolen by a pickpocket, in which case he would not have it with him but would not have forgotten or left it on the train.

Professional translators generally strive for a result that is smooth and idiomatic so that it reads like an original document. They often feel that this is possible only if they resort to pragmatic translations. Sometimes it is quite unavoidable, most notably where the target language requires substantive information to be made explicit that is only implicit in the source text. These situations are so common, and the crucial pieces of information often seem so unimportant, that they are not recognized for what they are.

Consider a text that describes one person saying to another, *Is that your cousin?* We do not know exactly who is being referred to, but we take it that it is clear to the people involved in the conversation. If we have to translate this into French, we are faced with a serious problem, because the only words for *cousin* that French makes available are specific to one gender or the other. The sentences *Est-ce que c'est ton cousin?* and *Est-ce que c'est ta cousine?* are both good French, but they crucially contain information about the sex of the person being referred to. Someone might suggest *Est-ce que c'est ton cousin ou ta cousine?*, but this will not do because the people whose conversation is being reported do know who they are referring to. For the sake of the unfortunate translator, we must hope that the surrounding context contains information that resolves the issue.

Consider the somewhat simpler situation where the original is French. Let us suppose that the two people having the conversation are observing another pair – a man and a woman. One says to the other, *Est-ce que c'est ta cousine?* Since the word *cousine* can be used only for female referents, we know which

one of the two people is being referred to. If this is translated into English in the most straightforward way, we get *Is that your cousin?* in which the sex of the referent is no longer explicit. A translator who thinks this important must adopt a creative, pragmatic, solution such as *Is that woman your cousin?* unless the person is clearly a child, in which case it must presumably be *Is that girl your cousin?* Suppose the reply to the question is *Non, je n'ai pas de cousine*. This cannot be translated as *No, I do not have a cousin*, because the French speaker is denying only that they have a female cousin, leaving open the question of whether they have male cousins. We leave the translation of this sentence to the creative reader.

In the NLP framework, machine translation systems must be the result of an automatic process, referred to as *training*, applied to a corpus of existing translations that constitutes the *training data*. At present, this results in a *translation model* and a *language model*. The translation model consists essentially of the word and phrase pairs essential to syntactic translation, and the system learns these essentially by computing the probability of seeing particular words or phrases in a target sentence, given the presence of a particular word or phrase in the source. The assumption is that the greater the number of training sentences considered, the better the estimate of these probabilities will become. But this is clearly true only if the overwhelming preponderance of the target sentences are the result of syntactic translation. Every pragmatically translated sentence is a red herring, and the more of them that there are in the training data, the more the training process will be led astray.

An examination of the Europarl corpus, which has constituted that training data for countless statistical machine translation systems reveals that the majority of sentences consisting of more than a very few words contain at least some material that was translated pragmatically. The second sentence of the first text contains the words *I would like once again to wish you a happy new year in the hope that you enjoyed a pleasant festive period* opposite the French *je vous renouvelle tous mes voeux en espérant que vous avez passé de bonnes vacances.* There is nothing in the French that corresponds to *I would . . . like to* and we can only take it that the English *once again* is represented by the first two letters of the French *renouvelle*. The French *en espérant que vous avez passé de bonnes vacances* translated more or less word for word into English, would be *hoping that you have had a pleasant holiday*. Shall we say that *holiday* means the same as *festive period*?

The following sentence begins *Although, as you will have seen . . .* in English, and *Comme vous avez pu le constater . . .* (*As you have been able to ascertain . . .*) in French. The sentence after that begins *You have requested a debate . . .* which appears in French as *Vous avez souhaité un débat* (*You wanted a debate*). So it goes on. Remember that the systems that are trained using these data will operate in a strictly syntactic manner, as we have characterized it.

The Russian linguist, Roman Jakobson, one of the founders of the "Prague school" of linguistics, famously wrote: *Languages differ essentially in what they must convey and not in what they may convey*. Some require the sex of the person being referred to to be made explicit. The Bantu languages do not, but they do require information about its shape, which, in the case of abstract objects, for example, is just as conventional as gender is in Indo-European Languages. English nouns must be singular or plural, and verbs present or past. Definite or indefinite articles are required, with different consequences for the message conveyed. When looking for a translation of the common verb *go* into German, we must make explicit whether the going is on foot or in some sort of conveyance. Airplanes are distinguished for this purpose, from other kinds of conveyance. In French, one cannot talk of books without making it clear whether they are of the kind intended primarily to be read, to be written in, or if they contain tickets to be torn out. To translate *to be*, that most common of English verbs, into Spanish, one must determine whether the property being ascribed to the subject is temporary or permanent.

Practitioners of NLP, as that term is generally understood, agree that all the information required for tasks like translation is implicit in the translations themselves so that a computer program can be made to learn how to translate from examples previously produced by human translators. Some of the information may be in very weak dilution so that considerable ingenuity and massive amounts of data may be necessary to recover it. Counterexamples are sufficiently rare to be negligible. The examples we have adduced, however, involve some of the commonest lexical and grammatical phenomena, and their resolution seems to require real experience of the real world by the translator. A rich source of further examples is to be found in what Hector Levesque has called *Winograd schemata* after the Stanford computer scientist Terry Winograd.[1] These consist of pairs of sentences, of which the following is the most well known:

> The city councilmen refused the demonstrators a permit because they feared violence.
>
> The city councilmen refused the demonstrators a permit because they advocated violence.

The question is: Who does the pronoun *they* refer to in each sentence – the councilmen or the demonstrators? Subjects agree overwhelmingly that it is the councilmen that fear, but the demonstrators that advocate. If one of these sentences is to be translated into a language in which the pronoun is required to manifest some kind of agreement, say in gender, with its antecedent, and if the

[1] Terry Winograd, *Understanding Natural Language*, Academic Press, 1972. See also Gary Marcus: "Why can't my computer understand me?" *The New Yorker*, August 23, 2013.

two potential referents differ in this property, then the translator must solve the riddle. In this case, what is presumably required is general knowledge about demonstrations, councilmen, and violence. In other words, successful translation is possible only by a person or a mechanism that, in a serious sense, understands the texts it works on.

There is a striking contrast between what we can learn from newspaper headlines about dogs and men biting one another on the one hand and Winograd schemata on the other. The headline conveys information about something that is, at least to some extent, surprising. The information is laid out as clearly and explicitly as possible. In the Winograd schemata, the information we are concerned with is secondary, unsurprising, and inexplicit. Not surprisingly, they present quite different problems to the designer of language-processing systems. A well-known statistical machine translation system was given the English sentence *He had eaten earlier that evening* and invited to translate it into Spanish. It delivered the result *No había comido esa misma tarde*, which means, of course, exactly the opposite. Presumably this is because a person's failure to eat is generally more newsworthy, and therefore more frequently remarked upon, than their eating, which people are generally expected to do regularly. The word *no* is introduced because the sequence *había comido* has been seen rarely, if ever, without it. The inescapable conclusions are that statistically based systems cannot extract information that is not explicit in the source text because it simply is not there to be extracted from data that has no referential component, and it will often treat explicit information incorrectly because it is precisely the divergence of a text from what is expected that makes it worth writing in the first place.

Consider the schema:

> They could not get the book in the box because it was too big.
> They could not get the book in the box because it was too small.

What was to big or two small? The book or the box? Most of us have had enough experience trying to put objects into containers to know how their relative sizes can influence the expectation of success. The problem does not have to have a particular grammatical form that immediately labels it as a Winograd schema. The sentences could have been: *Nobody thought there would be any problem packing the computer up, but, when they tried to get it into the box, . . .*

> they found that it was too big.
> they found that it was too small.

The sentences could be translated into French somewhat as follows: *Personne ne croyait que l'on auraient des problèmes à emballer l'ordinateur, mais quant ils essayaient de le mattre dans la boite . . .*

il s'est révélé trop grand.
elle s'est révélée trop petite.

Let us be clear about the nature of this argument. The claim is not that, no matter how much text a machine were to read, it would never encounter enough accounts of attempts to put objects into containers, with explicit information on how the outcome was affected by their relative sizes, to support the necessary inference. Let as suppose that the training data was there, and in sufficient quantity. The claim we are making is that, without explicit experience with using boxes as containers, successfully and unsuccessfully, one would have no way of even recognizing these accounts as relevant. More particularly, there would be no way of recognizing them as important for gender agreement in a French translation.

The force of Chomsky's argument from the poverty of the stimulus in human language acquisition is complex and doubtless will be discussed for a long time. What seems altogether less problematic is its relevance to natural language processing. The model of language that aspiring members of the NLP community must embrace is indistinguishable from that of the child listening to language coming through a loudspeaker in a dark room. There must be no chink in the wall that would allow some light to fall on the referential component that gives the system its entire value. The information thus abjured cannot be recovered by amassing greater amounts of text or bringing greater ingenuity to its processing. It simply is not there.

5.2 Features

Since the fourth century BC, when Pāṇini wrote his Sanskrit grammar, linguists have been at pains to locate the components of language – sounds, words and parts of words, phrases, and sentences – in a space, and thus to give substance to the appearance of similarity and difference among them, and to the observation that similar components tend to behave in similar ways.

Vowels are high, mid, or low. Independently, they are front or back and, on a third dimension, they are rounded or unrounded. The similarities in what the articulators must do in order to produce vowels correspond closely to those among their acoustic properties and the functions they fill in various language. Consonants arrange themselves on another set of dimensions. The initial consonants in the words *pot*, *tot*, and *cot* are *voiceless obstruents*, or *unvoiced stops*. They are all produced by briefly interrupting the flow of air through the mouth and then releasing it suddenly. They are voiceless because, unlike their voiced counterparts in *bot*, *dot*, and *got*, they are pronounced without moving the vocal cords. In English, there is release of air following the consonant when an unvoiced obstruent begins a word or a stressed syllable, except when

the obstruent is preceded by an *s* as in *spade*, *stake*, and *skate*. These are generalizations that are easy to make and to state given a prior classification of consonants in which unvoiced obstruents constitute a natural category. They also make it natural to observe and to describe other languages in which the realease of air, for example, is absent.

Linguists recognize multidimensional feature spaces on every level of linguistic analysis. Indo-European morphology has a number dimension with singular, plural, and sometimes dual, a case dimension with nominative, accusative, and up to six others in different languages, and a gender dimension with masculine, feminine, and neuter. In syntax, there are declarative and interrogative sentences, active and passive sentences, noun phrases and verb phrases, and so forth. A fundamental part of linguistics is thus a classification of linguistic phenomena that applies to all languages and that provides a set of *features* in terms of which to describe these phenomena in individual languages.

If a natural language-processing system can be constructed on the basis of a system of features, then one with the same capabilities can, in principle, be built without them. All occurrences of a term like *accusative noun* in the design of the first system would simply be replaced by a list of all accusative nouns in the design of the second. A system built in this way would embody none of the generalizations and insights that linguists see as central to their field but, if the object of the enterprise is to engineer a system to fill a practical need, this is a secondary issue at best.

However, that this might be done in principle by no means implies that it is what should be done in practice. At some point during the training of a system intended to translate between English and German, it might discover that the English word *dog* can be put together with the German word *Hund* to form a lexical unit. In a quite unrelated event, it might conclude that another lexical unit can be formed with *dogs* and *Hunde*. It might learn that *dogs* could also be paired with *Hunden* because, as we are quietly reminded by the linguist inside us, *Hunde* is the plural of *Hund* except in the dative case, when it is *Hunden*. A linguist would see the events in these three classes as different from one another, but he would see their similarities as more important than their differences. In particular, he would doubtless be led to postulate a different kind of lexical unit relating not two different English words and three different German words, but a single English lemma and a single German lemma.

A lemma belongs to a family of words, *dog* and *dogs* in English, and *Hund*, *Hunde*, and *Hunden* in German. The number of lemmas in a text is generally smaller than that of words because there are less of them. The amount of information that each provides is therefore greater. The concomitant advantage is greater for morphologically rich languages than morphologically poorer ones – greater for German than for English, greater for Finnish than for German, and greater for Turkish than for Finnish.

In the early stages of studying German, students learn that the objects of certain prepositions, such as *von*, are always in the dative case, and they can apply this rule to all the nouns they encounter, even if they have never actually seen them in the dative plural. Students of German learn very early how to recognize and construct the dative plural forms of regular nouns. They also learn that the plural forms for the other cases are all the same. Also, in the singular, most nouns have at most two forms. In this respect, the linguistically informed human language learner lives in an incomparably richer and more secure environment than the machine for whom different spellings imply different words.

Zipf's law is a mathematical codification of the observation that very few words and phrases occur very frequently in texts, and a large number occur extremely rarely. Indeed, as the sizes of the texts one considers increases, the number of words that occur only once increases also. If we think of a person's entire experience with a language as a single text, this means that there will also be a large number of words that that person has seen in only one of its possible forms. In a frequently cited paper, Eugene Charniak[2] describes automatically extracting rules of a context-free grammar from a corpus of 300,000 words. The grammar contained 10,605 rules, of which 3,943 occurred only once. Rules are subject to Zipf's law just as words are. As Charniak says, *At about eleven thousand rules, our grammar is rather large* and, of course, correspondingly uninformative.

The observations we are making here support the view that the generalizations linguists are at pains to formalize in the course of analyzing a language are also made by speakers in the course of learning the language. We have argued that there is great practical advantage in generalizations about vocabulary items and their morphological variants. But there is altogether greater advantage that comes with generalizations and features at the higher levels of linguistic organization. Mastery of sentence structure, for example, whether for the linguist or the language learner, rests crucially on grasping the notion of locality that is appropriate to the words in a sentence. This notion of locality is relative not to proximity in the string of words, but to proximity in a recursive structure which gives the sentence its coherence. Consider the following sentences:

1. She added all the things she had bought up the same street.
2. She added all the things she had bought up the same way.
3. She added all the things until she had bought up the whole street.

The first sentence claims that she added things and that the things she added were bought *up the same street*. The words *up the same street* constitute a single entity in the recursive structure of the sentence, as do the words *the same street*.

[2] Eugene Charniak. 1996. "Tree-bank Grammars" Technical Report CS-96-02, Department of Computer Science, Brown University.

So the word *up* is very close to *the same street* in this structure. In sentence 2, on the other hand, the word *up* is the second part of a two-part lexical item of which *added* is the first part. These two words are therefore more closely bound to one another than either of them is to the words *all the things she had bought* that separate them in the string. In sentence 3, *up* is also the second part of a complex lexical item. Here, however, the first part is *bought*. Up is closer to *bought* than it is to *the whole street*.

Together with the aforementioned three sentences, consider the following pair:

4. She put any member of the family that came down up for the night.
5. She put any member of the family that came up down as a scoundrel.

Sentence 4 contains the lexical items *put up* and *come down*. Sentence 5 contains *come up* and *put down*.

Generalizations like these go back at least as far as Pāṇini and their importance for ordinary language learners as well as linguistic theoreticians was taken for granted from the fourth century BC until the advent of natural language processing at the end of the 20th century. The machine translation systems developed within this framework place much of the burden on a component known as the *language model* that is charged with selecting the most promising of the candidate translations proposed by the so-called *translation model*. Language models are, according to the commonest conception, all about the proximity of words in the string. Roughly speaking, one candidate will be considered better than another if it contains more and longer substrings that were also found in the training texts. The more often they were found there, the better. On such criteria, the sequences *came down* and *came up* might well be recognized as essentially idiomatic, but it is hard to see how *put . . . up* and *put . . . down* might be so recognized.

As translations of the aforementioned sentences, Google Translate gives the following in French:

6. Elle a mis tout membre de la famille qui est descendu pour la nuit. (*She put each member of the family that descended for the night*)
7. Elle a mis tout membre de la famille qui est venu vers le bas comme un scélérat. *(She put each member of the family who came downwards like a scoundrel)*

As German translations, Google gives:

10. Sie legte ein Mitglied der Familie, die für die Nacht kam. (*She laid a member of the family who came for the night*)
11. Sie legte ein Mitglied der Familie, die nach oben nach unten als Schurke kam. (*She laid a member of the family who came from above to below as a scoundrel*)

The information in the recursive structure of sentences can easily conflict with what the string, considered simply as a linear sequence, seems to imply. The sentence *The man with the dog fought for his life* contains the sequences *the dog fought* and *fought for his life*. The second is aligned with the true structure of the sentence, but the first is not. However, it is reasonable to suppose that both sequences would be well represented in a training corpus of interesting size. A sentence that is very similar to this, but which lacks the ambiguity, is *The man with the dog fought for her life*. We still do not know whose life is being fought for – the life of the dog or that of some previously mentioned person or animal – but we do know that it was not the man. The ambiguity would have to be resolved if the sentence were to be translated into Russian where in the sentence *He fought for his life*, the pronoun used to translate *his* would be different depending on whether it referred to the same person as *he*.

One may counter this argument by pointing out that natural language processing is in its early stages. Many of its proponents would readily acknowledge the importance of syntax and of recursive structure. It is just that we have not yet discovered reliable ways of learning how to recognize that structure from raw data. If this situation persists, we will go back to the linguists, some of whom are among our best friends, and have them annotate texts so as to make their structure explicit, and will have our systems acquire this part of the knowledge they need from the resulting annotated corpora. By having them annotate texts rather than coming right out and explaining to us how it is done is that we will still be able to claim that the eventual system will be the product of machine learning, a position to which we are religiously committed.

The disadvantage of annotation is that it is subject to Zipf's law so that each annotation that is made contains less information than the one before it. But the approach cannot be dismissed casually. The best part-of-speech taggers, after all, make hardly any more errors on a given task than experienced humans, and they were learned from texts annotated by human taggers. Furthermore, the arguments we have been making about the importance of recursive structure for recognizing lexical units, and for translation in general, presumably apply equally to the task of part-of-speech assignment. So let us see how the Stanford tagger, widely acknowledged to represent that state of the art, behaves. For the input *She put any member of the family that came down up for the night*, it returns *She_PRP put_VBD any_DT member_NN of_IN the_DT family_NN that_WDT came_VBD down_RP up_RP for_IN the_DT night_NN*. Our main interest is in the words *up* and *down*, which are both correctly tagged as particles (RP). The same goes for *She put any member of the family that came up down as a scoundrel* which is tagged *She_PRP put_VBD any_DT member_NN of_IN the_DT family_NN that_WDT came_VBD up_RP down_RP as_IN a_DT*

scoundrel_NN. Once again, *up* and *down* are tagged as particles. Unfortunately, to say that something is a particle is to say no more than that is a one member of a pair. Without the other member of the pair, this information can engender nothing but frustration.

The levels of abstraction on which linguists study language make contact with the outside world on the lowest level where they face sounds, articulatory gestures, and character shapes, and at the top level, where they encounter references to concrete and abstract objects in the outside world. It is at the top level that the child in the dark room and the machine that sees only text suffer most spectacularly from their abandonment by science and understanding. Consider the following scenario, which is not fictional. A lady comes into a railroad car on a train that is about to leave Montpelier and asks one of the passengers, *Does this train go to Perpignon?* The passenger replies, *No, it stops in Béziers*. Two obvious alternatives would suggest themselves to someone who, against all odds, were in the position of translating this into German, namely *Nein, er hält in Béziers* and *Nein, er endet in Béziers*. They are incompatible because *hält* implies that the train stops briefly in Béziers, and then resumes its journey, possibly toward Perpignon, while *endet* implies that Béziers is the train's terminus. A translator that was familiar with the rail system in southwestern France would know that Béziers is indeed on the line from Montpelier to Perpignon and that, while that is the terminus for some trains, others continue on to Perpignon. What is in question is therefore almost certainly that Béziers is the train's terminus so that *hält* is the best choice for the German verb.

The alternative approach to the problem has the advantage that it would be open to translators without knowledge of the local geography or train schedules. Notice that the single word *No* would have been an entirely adequate answer to the lady's question. The rules of polite discourse, however, require some explanation to be given, and the remark about Béziers fills this function. If the train made a brief stop in Béziers before continuing toward Perpignon, then the reply would not constitute an explanation, so that a German translation using the word *hält* would make no sense.

The second of these translation strategies illustrates what has been referred to as the cooperative principle in language use. Language could presumably not be used with carefree abandon in everyday situations if each utterance had to be precisely crafted like a mathematical formula to fulfill its purpose. People who want to understand must therefore be understanding. They must construe the utterances they hear in the ways in which they think the speaker is most likely to have intended them. A good listener is therefore someone who brings to the task as much knowledge as possible of the subject matter in general, the potential referents of the words and phrases, and good judgment about the interlocutor.

5.3 Linguistic Computation and Computational Linguistics

We have been trying to promote the view that linguistic computation should be distinguished from computational linguistics. Proponents of natural language processing hold to the term *computational linguistics* presumably because it has provided them with a most effective Trojan horse in which to penetrate linguistics departments in universities as well as professional societies and conferences. Those involved, however, apparently either dismiss attempts to understand how language actually works as irrelevant or persuade themselves that, by achieving high scores in public competitions, they actually transform statistical ingenuity into linguistic insight. Some members of the fraternity were competent linguists in an earlier life. They may perhaps be clinging to the forlorn hope that these two almost unrelated enterprises may one day come together, their discoveries fusing and taking us to a level of understanding beyond anything we can now contemplate.

For the most part, proponents of natural language processing see themselves as engineers, and are happy to be judged by what they contribute to the solution of practical problems. They have given us Google Translate, many and varied information retrieval systems, and programs that can tell us if a consumer review was favorable to the product or not. They do not begrudge old-style linguists in general, and computational linguists in particular, their interest in what they call *science*. But it behooves the linguists to say how it is that computation plays such a central role in their endeavors. After all, physicists, psychologists, and accountants routinely use computers, and special techniques and procedures have been developed to fill their requirements. But, as has often been pointed out, we do not distinguish computational physicists, psychologists, and accountants. Why, then, do computational linguists cling to this term?

The historical origins of the term *computational linguists* may make its continued use today look incidental and perhaps a little cynical. People who were working on machine translation when the ALPAC report[3] appeared in 1966 foresaw the imminent disappearance of their livelihood if they could not rapidly recast themselves as the scientists that the report said would be required to give machine translation the foundation it desperately needed. If they were to be the practitioners of a new science, the most important thing that that science would need was a name. At least three names were proposed: *Automatic Natural Language Data Processing*, *Computational Linguistics*, and *Mechanolinguistics*. In the interest of good public order, we withhold the identities of the advocates for each of these proposals.

[3] John R. Pierce, John B. Carroll, et al., Language and Machines – Computers in Translation and Linguistics. ALPAC report, National Academy of Sciences, National Research Council, Washington, DC, 1966.

To say that we have come to be known as computational linguists by accident is not to say that the name is inappropriate or unwarranted. A language is not just a corpus. It is not simply a set of rules or organizational principles. In addition to these things, which characterize what is traditionally known as its *paradigmatic* component of language, it is a set of processes that take place in people's heads each time they produce or understand an utterance or write or read a sentence. These constitute the *syntagmatic* component. The paradigmatic component is at the service of the syntagmatic component giving, as it does, the structure shared among speakers of a language that enables them to understand one another.

There is a complex interplay between the paradigmatic and the syntagmatic components of language. The more permissive the rules and organizational principles, the more complex the syntagmatic processes, and the less psychologically plausible the resulting overall model. The study of processes, as abstract entities, is called *computer science*. The question of how a set of rules and organizational principles might be brought to bear on the problem of producing an utterance or understanding a sentence is in large measure one of algorithm design and computational complexity. We should by no means be taken as claiming that human language processors are essentially von Neuman computers or Turing machines. Computer science is about machines that exist and machines that are possible. In short, it is about the notion of process in a pure and abstract form, and this is at the very core of linguistics as it should and must be.

The approach to linguistics proposed by Zelig Harris (1900–1992) and developed by Chomsky and his followers is all about process. Sentences, for example, are described and explained in terms of the process that is assumed to result in their construction. Families of sentences, related typically by similarities in their meaning, start with a common form, which is then modified in a series of steps in which words and phrases are moved from one place to another in the structure. Movement is at the core of the system. An entity that would normally be expected to move from one place to another in the structure may be prevented from doing so because the location to which it would move is already occupied by another entity. The exact order of events is crucial, and everything turns on finding rules or, preferably, more general constraints on the sequences of events so that only the observed outcomes are possible.

The proponents of these systems usually refer to them as *generative grammar*, though they do not have an exclusive claim to this term. Systems that are similar to them in broad outline are also familiar to computer scientists. A *compiler*, for example, translates programs from the language in which the programmer writes to a minutely specified set of operations that the computer must carry out in exactly the order specified to compute the result the programmer desires. Programmers do not specify this sequence of events directly because the level of detail required is unnecessarily burdensome and obscures the

overall intent of the program. The language in which the programmer writes allows the sequence of events to be specified in much more general terms.

The models of language envisaged by Chomsky and his followers are similar to computers in requiring steps to be specified in minute detail. They differ to the extent that computer scientists recognize that there is a point beyond which detail obscures more than it reveals. Mathematicians do not attempt to explain the notion of quotients by describing the process of long division. Indeed, the only way a student can reasonably hope to understand long division is to first get a firm grasp of quotients. The possibility of bypassing long division that computers and calculators provided was seized immediately, and now we hear of it no more.

Computational linguistics has been, in large measure, an attempt to bring to linguistics some of the advantages that computers allowed in the teaching of arithmetic and that compilers brought to computer science. Given that the field is barely 60 years old, its successes have been remarkable. It has been successful in providing a number of models of linguistic processing whose computational power matches what the task seems to require extremely closely. It is powerful enough to allow all the necessary computation to be specified but constrained enough to capture important insights into the fundamental nature of linguistic processes.

Computational linguistics has given us a version of finite-state technology that is apparently a remarkably good fit for the requirements of morphology and morpho-phonology. It does not require the linguist who uses it to think in terms of states and transitions, but rather in terms of a sequence of representations of a word, each related in a straightforward way to the ones immediately preceding and following in the sequence, connecting an abstract representation at one end of the sequence with the observed representation at the other. This has become a common way of thinking among formal linguists, who generally find it very congenial.

More striking are the contributions that computational linguists have made in syntax, where such formalisms as Lexical Functional Grammar, Head Driven Phrase Structure Grammar, and Combinatory Categorial Grammar are widely acknowledged as major contributions. In these frameworks, Chomsky's problem of how to form a question from the sentence *The man who is hungry is ordering dinner* simply does not arise. The assertion and the question are indeed assumed to have similar underlying structures, but neither has a privileged position relative to those structures.

5.4 Conclusion

Nontrivial tasks, like translation, are AI-complete. In other words, any machine that can perform them would have to be able to mimic all aspects of human