Chapter 2

The Micro-Resource Model

1. Introduction

A linguistic theory consists of two sorts of information -- *resource* information (the static knowledge of language -- the lexicon, the grammar and semantics, and the mapping relations between them) and *process* information (concerning *how* the resource knowledge is utilised, e.g., how the static knowledge is used in the generation or parsing of text). Before discussing linguistic processes in part B, I will discuss the resources used in the WAG system.

While Systemic resources have been discussed in various places before (e.g., Halliday 1985; Berry 1975/77; Martin 1992, etc.), very little attention has been given to the inter-relation of the various resource modules, and to the inter-relation between resources and processing. The discussion in these chapters of the resource model is thus not wasted -- I am discussing a Systemic resource model oriented towards machine processing. These chapters also function as background for the later discussion of Systemic processing: the processing of Systemic resources cannot be understood unless the resources themselves are first understood.

This chapter provides an over-view of the resource model, the Systemic formalism, and a short description of two of the strata of the resource model: lexico-grammar and graphology. A description of the lexicon -- a resource which maps between strata -- is also provided.

The remaining chapters of part A outline in more detail the three components of the semantics -- ideational, interactional and textual meaning. The final chapter describes the resource which maps between the semantics and the lexico-grammar -- the semantico-grammar mapping resource.

2. Components of the Resource Model

This section provides a broad overview of the components of the resource model.

2.1 Linguistic Strata

A Hallidayan model of language posits four levels of linguistic representation: context, semantics, lexico-grammar, and phonology/graphology. These strata, and their stratal relation are shown in figure 2.1. This thesis deals with three of these levels of representation:

- **Micro-Semantic representation**: representation of the sentence in terms of content (ideational meaning), exchange (interactional meaning) and message (textual meaning);
- **Lexico-grammatical representation**: the representation of the sentence in terms of its lexical and syntactic structuring;

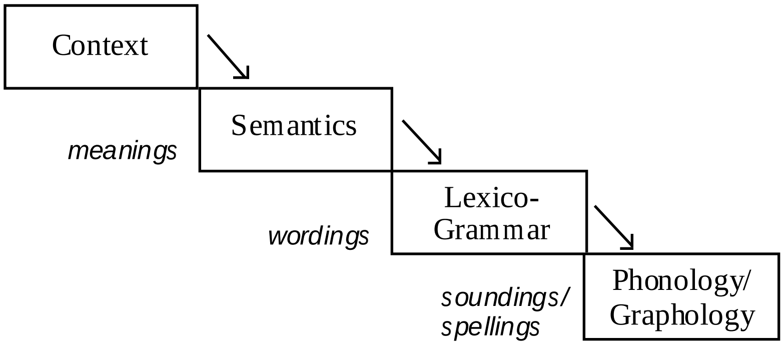


Figure 2.1: The Strata of a Systemic Model

- **Graphological representation**: the representation of the sentence as it appears on the page, a construction of words, characters, and punctuation marks.

Context is beyond the scope of this thesis. See Gregory & Carroll (1978), or Halliday & Hasan (1985) for discussion of context and its role in relation to language. I also will not deal with phonology, even though one target application of the WAG system is a dialogue system. Speech can be obtained from graphological output using existing text-to-speech technology. For instance, since the system runs on Macintosh computers, the Macintosh Speech Manager can be used to provide reasonable quality voices, converting punctuation to pauses (commas, full stops, colons), or intonation (although poorly -- from question marks, full stops). Although the quality of speech from text is not as good as produced using a full phonology, it is sufficient for this thesis since the focus is on the higher strata.

2.2 Micro- vs. Macro-Resources

Following Matthiessen (personal communication), I divide language resources along the axis of **micro-resources**, and a **macro-resources**. Micro-resources concern the representations of linguistic units which are co-extensive with the sentence. This includes the graphological sentence, but also the clause or clause-complex (lexico-grammatical stratum), which tend to be realised as sentences. It also includes the semantic representation of a sentence (a *micro-semantic representation*), which includes the specification of the speech-act, ideational content, and textual structure of the sentence.

The macro-resources concern the representations of *multi-sentential text*, whether at the graphological stratum (e.g., paragraphs, sections, etc.), or semantics (the ideational, interactional and textual structuring of multi-sentential text)[1].

This division of the resources into micro- and macro- cuts across the stratal boundaries, as shown in figure 2.2. Micro-resources -- the resources for sentence-size units -- will be our concern in this thesis.

[1] It is not clear what form macro-lexico-grammatical representation would take, but possible candidates are lexical collocation across sentences; and parallelism (the phenomena where adjacent sentences are provided with identical or similar grammatical structure - see Halliday & Hasan 1985).

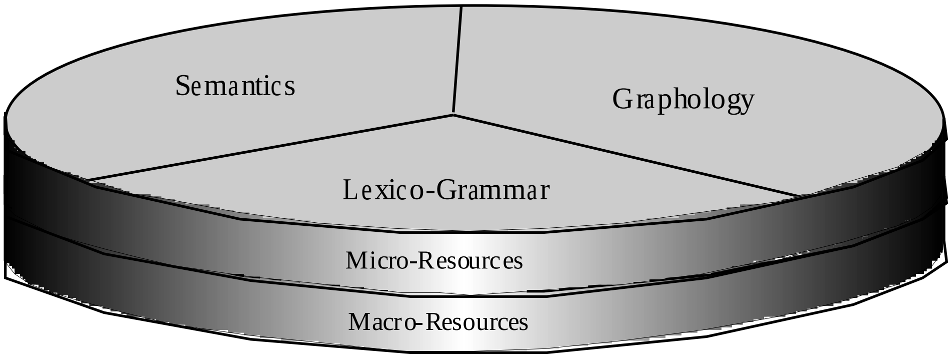


Figure 2.2: Micro- and Macro-: Cutting across the Strata

2.3 Semantic Metafunctions

Most of my attention in this resource model discussion will be on the semantics. The semantic resources describe the potential meanings for the language. Following Halliday (1978), meaning resources are split into three types:

- **Ideational Meaning**: the propositional content of the sentence, structured in terms of processes (mental, verbal, material, etc.), the participants in the process (Actor, Actee, etc.), and the circumstances surrounding the process (Location, Manner, Cause, etc.).
- **Interactional Meaning**: meanings which concern the speaker and hearer and their inter-relation. Interactional meaning includes the participant's attitudes, social roles, illocutionary goals, etc.
- **Textual Meaning**: How the text (or, in the case of micro-semantics, the sentence), is constructed as a message conveying information. This concerns, for instance, the inclusion or exclusion of information in the message, the prominence of information that is included, the projection of information as recoverable or not, and the thematic structuring of the message.

For details of these meaning resources, see chapter 3 (ideational resources), chapter 4 (interactional resources) and chapter 5 (textual resources).

2.4 Interstratal Mapping

An important part of any multi-stratal theory is the component which maps between the strata -- the interstratal mapping resources. These resource show how representations on each stratum corresponds to each other, for instance, the relationship between semantic representation and lexico-grammatical representation. Chapter 6 will explore this inter-stratal resource.

The lexicon forms another inter-stratal resource. It maps between the three strata, associating grammatical and ideational features with graphological forms. This will be discussed in section 6 below.

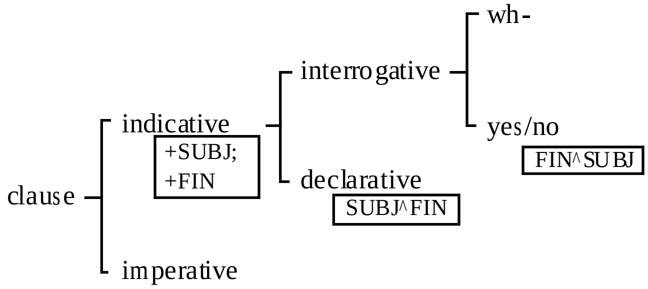


Figure 2.3: A Partial System Network with Realisations

3. The Systemic Formalism

One of the aspects of Systemic-Functional Linguistics (SFL) which makes it attractive is that the same formalism can be used for representation at all strata[2]. This approach has been taken very strongly in this thesis, using system networks and structures to represent not only grammatical structures, but graphological potential, ideational knowledge, speech-acts, etc. Systemics can be used as a general Knowledge Representation Language (KRL). While the examples in this section are mostly drawn from the lexico-grammar, later discussion will demonstrate the use of the formalism on other strata.

The Systemic formalism (cf. Halliday 1961; Hudson 1971; Matthiessen 1985; Bateman 1989b; Fawcett *et al.* 1993) is unique in the emphasis it gives to the paradigmatic (choice) axis of language. A Systemic grammar extracts all options out of the structure rules and represents these options as a separate resource. There are thus two components to the Systemic formalism -- a **system network** -- representing the linguistic options (paradigmatic axis -- see figure 2.3), and a set of **realisation statements** (structure templates) -- representing the potential linguistic forms (syntagmatic axis). The realisation statements are explicitly related to the features which they realise (shown in the boxes under the features). The linguistic options are termed **features.**

3.1 System Networks

A system network represents the options available to the language user. The network describes the mutual exclusivity or compatibility of the various options (in Gazdar *et al.* (1985)'s terms, *feature co-occurrence restrictions*, although the concept has been in use in Systemics since the early sixties). A system network consists of a set of **systems**, each system representing choices in paradigmatic opposition (mutually exclusive choices).

The ordering of systems from left to right in the network is read as more **delicate** specification of the options available. For instance, in figure 2.3, the first choice concerns the **mood** of the unit: *indicative* or *imperative*. The further structural options available for the unit are developed by the systems to right of this basic choice. *[indicative]* clauses for instance may be either *declarative* or *interrogative*.

In computational terms, a system network is an inheritance tree, each feature inheriting the properties of the features to its left. It differs from other types of inheritance trees by

[2] The claim here is not that all work by systemicists uses the system network formalism. Rather, I claim that most phenomena can be modelled in the formalism, although expansion of the formalism may be needed for some phenomena (e.g., a dependency interpretation of the formalism for modeling ideational structures).

requiring the features to be organised in terms of *disjoint coverings* (the systems). A system represents a *disjunction* of features -- only one feature must be chosen from an entered system. The system is also a *covering* -- it covers the total paradigmatic space -- if the system is entered, one of the features must be applicable .

Networks make use of several types of systems. Table 2.1 describes the most common system types. A gate is a system with only one feature choice. Gates were introduced by the Penman system to handle cases where a realisation statement (or statements) is conditioned by several features. Since Penman associates realisation statements only with single features, the multiple features are gated to produce a single, artificial, feature, to which the realisations are associated. An alternative approach (sometimes used by James Martin (lecture notes)) allows feature complexes to be associated with realisation statements e.g., *[a:b] -> +Role*. A third approach, used by Fawcett, associates the realisations with a single feature, but allows a set of additional features to condition the realisation, e.g., *[a] -> +Role if [b]*.According to Matthiessen (personal communication), this approach was also followed in the earliest version of Penman. WAG follows Penman’s present approach, using gates.

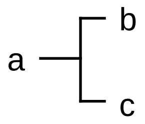
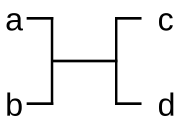
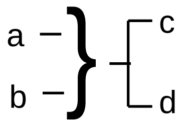
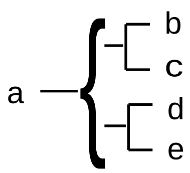
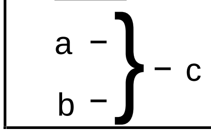
Simple entry condition 	If *a* is selected, then there is a choice between *b* and *c* (*b* and *c* are mutually exclusive).
Disjunctive entry condition 	If either of *a* or *b* is selected, then there is a choice between *c* and *d* .
Conjunctive entry condition 	If both *a* and *b* are selected, then there is a choice between *c and d* .
Simultaneous systems 	If *a* is selected, then there are two choices to be made -- between *b* and *c,* and between *d* and *e.*
Gates 	The combination of choices *a* and *b* can be represented by the single choice *c.*

Table 2.1: Types of Systems

Computational Representation of Systems: In WAG, systems are defined using forms like the following (derived from the Penman form):

```
(defsystem
   :name indicative-type
   :entry-condition indicative
   :features
     (  (declarative
          :selection-constraint
             (:type *speech-act* (:and (:or elicit propose)
                                       information-negotiating))
          :realisation   (:order Subject Finite)
        (interrogative)) )
```

The fields of this definition are as follows:

name: a unique identifier for the system, in this case, *indicative-type.*

entry-condition: the entry-condition of the system -- the feature context under which this system is activated. Note that the entry condition can include **and**s and **or**s, and can get as complex as necessary. For instance, from another system...

```
(:and not-auxed
      (:or unmarked-positive
           (:and negative be-intensive))
      (:or assertive be-intensive))
```

features: a list of entries, one per feature in the system. The first item in an entry is the feature itself. For each feature, two sorts of information may be provided:

selection-constraint: a structural condition which must be met for the feature to be selected. This field is used in the inter-stratal mapping (see chapter 6). In the above instance, the grammatical feature *indicative* can only be selected if the speech-act (which this clause is realising) is of a particular type.

realisation: the realisations statements associated with the feature.

3.2 Realisation Statements

A realisation statement consists of an *operator*, followed by a set of arguments (normally roles and features). Table 2.2 lists the realisation operators used in the WAG system, which are largely identical to those of Penman (see Matthiessen 1985). Under the Operator column, there is occasionally a second operator in brackets. This is an alternative labeling which WAG allows, replacing the direction-biased Systemic terms. The examples column shows two forms: the internal representation, and the notation usually used by Systemicists.

Operator	Example	Description
Insert (Require)	(:insert Pred) *+Pred*	The nominated role is required to be present in the structure.
Conflate (Same)	(:conflate Modal Finite) *Modal/Finite*	The nominated roles are filled by the same element, e.g., in a modal clause, both the Modal and the Finite role point to the same modal-verb.
Order	(:order Subj Fin) *Subj ^ Fin*	The filler of the first role is sequenced directly before the second. Any number of elements can be sequenced in a single statement.
Partition	(:partition Process Manner) *Process ... Manner*	The second element appears somewhere after the first, but not necessarily immediately adjacent.
Preselect (Type)	(:preselect Subject: nominal-group) *Subject: [nominal-group]*	The nominated role must be filled by a unit with the specified feature. Note that the preselection can be logically complex, allowing any combination of *and, or* or *not* in the feature specification. For instance: Subject : (:and nominal-group (:or nominative accusative) (:not wh-head))
Lexify	(:lexify Deictic the-det) *Deict : {the-det}*	The lexical item is assigned directly to the element of structure. Lexify overrides any preselect which may apply to the same element of structure[3].
Presume	(:presume Subject) *-Subject*	The specified role, while present in the structure for ordering purposes, is for other purposes not present in the structure. Used for phenomena such as grammatical ellipsis.

Table 2.2 WAG's Realisation Operators

3.2.1 A Note on Ordering

Penman's formalism employs two additional concatenation operators: *OrderAtFront* and *OrderAtEnd*, to specify that the filler of a role is to appear as either first or last element in a structure. Rather than introduce specific operators for these purposes, WAG uses only the *order* operation, but introduces two pseudo-roles to indicate the beginning ('Front') or end ('End') of the structure, e.g.,*Front ^ Subj ^ Obj ^ End.* This operator will become more prominent in the parser, where we need explicit statement of what can start a unit, and when a unit is finished.

The WAG implementation extends on the Penman formalism by allowing optional elements in *order* or *partition* statements. Optional elements do not necessarily appear in the final structure. Optionality is indicated by surrounding the element in parentheses, e.g.,

```
        (:order Subject  Finite  (Negator) )
```

By use of partition, optional elements, and the pseudo-roles, WAG can constrain unit sequencing in generation without resort to the default ordering rules which Penman relies on to determine order when the realisation rules under-constrain it. This step was necessary in the move to parsing, since default orderings make no sense in the parsing environment, where elements may occur in orders other than the default.

[3] This is the only case of non-monotonic logic in the WAG system. A non-monotonic system allows prior assertions to be over-ridden by later assertions, while a monotonic system does not.

3.2.2 Non-implemented Operators

There are several realisation operators used in Penman which I have found unnecessary. These are listed below.

Classify, Inflectify, e.g., *Deict = [det]*

Penman’s formalism makes a distinction between lexical, inflectional and grammatical features, and so distinct operators are used to classify a lexical item, specify its inflections or to preselect a grammatical item. The distinction has been dropped in favor of a generalised preselection operator which operates on any feature type (this includes units at all strata, e.g., ideational features).

OutClassify, e.g., *Deict != [det]*

Specifies that the lexeme filling the role cannot have the specified feature (i.e., lexical restriction). Since ‘preselect’ in WAG allows negation of features, this operator is unnecessary.

Expand, e.g., *Mood(Subj)*

'Expand' is a realisation operator which is occasionally used in Halliday's work, and is included in the Nigel grammar (Penman’s lexico-grammar). It specifies that the nominated role (in this example *Mood*) has a constituent role: *Subj*. It allows the grammar to specify role structure without introducing corresponding class structure (Matthiessen & Martin 1991). The WAG system is at present built on the assumption of a correspondence between class structure and role structure, so the Expand operator is not used.

3.3 Structures

So far I have discussed the Systemic formalism as potential: the resource for constraining the possible representations. I will now look at the instances drawn from this potential, in other words, Systemic representations.

The realisation statements of the Systemic formalism can be interpreted from either a *constituency* or a *dependency* perspective. In the constituency perspective, as is used in the grammar, the *insertion* of an element is interpreted as the requirement of a constituent. The inserted element is *part* of the inserting element. Figure 2.4 shows the constituency structure for a clause. At each level of constituency, the unit is assigned both role structure and a feature description (the *selection expression* of the unit).

This structure conforms to the realisation constraints of the partial system network of figure 2.3. Since it is *indicative*, both Subject and Finite must be present. Since it is a *yes-no* clause, the Finite is ordered before the Subject. Lexico-grammatical representation will be further explained in section 4 of this chapter.

The Systemic formalism can also be used to represent dependency structures. The insertion of an element is interpreted to mean the requirement of a *sister* element. I use this dependency interpretation for semantic structures. Figure 2.5 shows a simplified ideational analysis of the sentence in figure 2.4. The representation shows a material process, with associated dependent roles: Actor and Event-Time. The Actor also has a Name role specified.

Is	*Mary*	*coming*	*tomorrow*
[clause:modal:indicative:interrogative:yes-no:temporally-located]			
Finite/ Progressive	Subject	Pred/ ProgC	Circumstance
	[nominal-group:proper-group]		[adverbial-group]
	Head		Head
[be-aux]	[proper-noun]	[lexverb: ing-verb]	[adverb]

Figure 2.4: A Lexico-Grammatical Representation

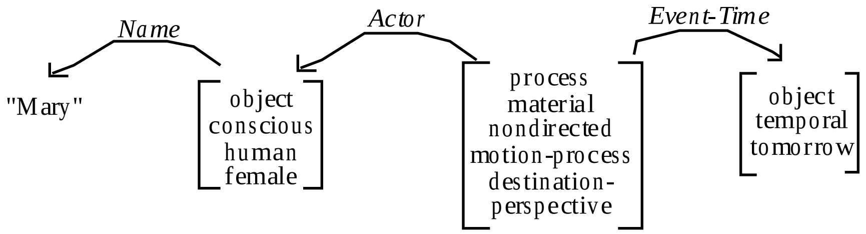


Figure 2.5: An Ideational Representation

I have discussed the Systemic formalism in general, and the remainder of Part A will look at specific applications of the formalism, detailing its use in the various modules of the resource model. I start below with a description of the lexico-grammatical stratum.

4. Lexico-Grammatical Resources

The lexico-grammatical resources specifies the possible lexico-grammatical structures of a language. This resource is used to guide the construction of a lexico-grammatical representation, in both analysis and generation. In common with the resources on other strata, it consists of a system network with associated realisation constraints. This network represents the lexico-grammatical potential, each lexico-grammatical representation being an instance derived from this potential. Halliday uses the term ‘lexico-grammar’ because:

> “Within this stratum there is no hard and fast division between vocabulary and grammar; the guiding principle in language is that the more general meanings are expressed through the grammar, and the more specific meanings through the vocabulary.” (Halliday & Hasan 1976, p5).

Only a brief description of the lexico-grammatical stratum will be provided here, since this stratum is well described elsewhere (cf. Halliday 1985; Hudson 1971; Berry 1975/77), and these descriptions are in general fairly consistent. I focus on the semantic stratum, since this area is less well described in the computational setting, and demonstrates more divergence of approaches. Inter-stratal mapping is also a focus, for the same reasons.

The descriptions of English used in examples throughout this thesis are, unless otherwise noted, derived from my own Systemic grammar of English, which varies in some ways from that of Halliday (1985). Since the focus on this thesis is not at all on providing a description of English, I will not usually point out where my descriptions vary from Halliday, or justify my variations. Any examples are shown merely to demonstrate the formalism.

Lexico-grammatical representations consist of two parts:

(i) **Role Structure**: a list of the component roles of the unit. Two or more roles may conflate, and thus have the same filler;

(ii) **Selection Expression**: each unit is assigned a set of features, being a valid path[4] through the system network.

[4] One can generate all possible combinations of features allowable from the network. Each of these combinations is termed a ‘path’ or ‘selection expression’. The term ‘path’ is used because to produce a path,one can start at the root of the system network (the left-most feature) and traverse to the leaves of the network (the right-most features), selecting one feature in each entered system. One forms a ‘path’ through the system network.

	Will	*the*	*man*	*come?*
clause rank	[clause: modal: interrogative: yes-no]			
	Finite/ Mod	Subject		Pred/ ModC
group rank		[nominal-group]		
		Deict	Head	
word rank	[modal-verb]	[determiner]	[noun]	[lexverb: infinite-verb]

Figure 2.6: Lexico-Grammatical Analysis showing Rank Structure

The sample analysis in figure 2.6 is based on a small fragment of a Systemic grammar, which will be introduced below. Each unit is assigned both a selection expression (in gray), and also its internal role structure (except for word-rank units).

Rank Structure: This means of diagramming lexico-grammatical structure brings out the *rank structure* of the clause. Halliday's 'rank hypothesis' states that all grammatical units can be analysed in terms of three 'ranks' of structure -- clauses, groups (= phrases), and words. He suggests that the typical pattern in English is to have clauses made up of groups, and groups of words. However, the hypothesis includes the notion of *rank-shift*, where a group can function within a group, e.g., 'the used car salesman', or clauses within groups, e.g., 'the fact that he was coming is certain'.

Multiple Layers of Structure: Note how some units have multiple roles assigned to them (e.g., the modal verb fills both the Mod and the Finite function). One of the features of a Systemic grammar is that it allows multiple layers of role structuring within each unit. More on this presently.

I will now show how this analysis relates to the lexico-grammatical resources: how the instantial relates to the potential. The main lexico-grammatical resource consists of a system network (a 'taxonomy' of valid grammatical units), with associated structural constraints (realisation statements). A sample network is shown in figure 2.7, describing only a small fragment of English sentences. This network describes finite clauses only, and allows clauses to be either declarative or interrogative. Clauses are also either modalised (contain a modal verb) or unmodalised (future tense is here treated as a type of modalisation). Other aspects of English clause structure, such as transitivity, presence of participants other than subject, perfective or progressive aspect, etc. are not considered. Nor is there any description of the nominal-group or word rank systems.

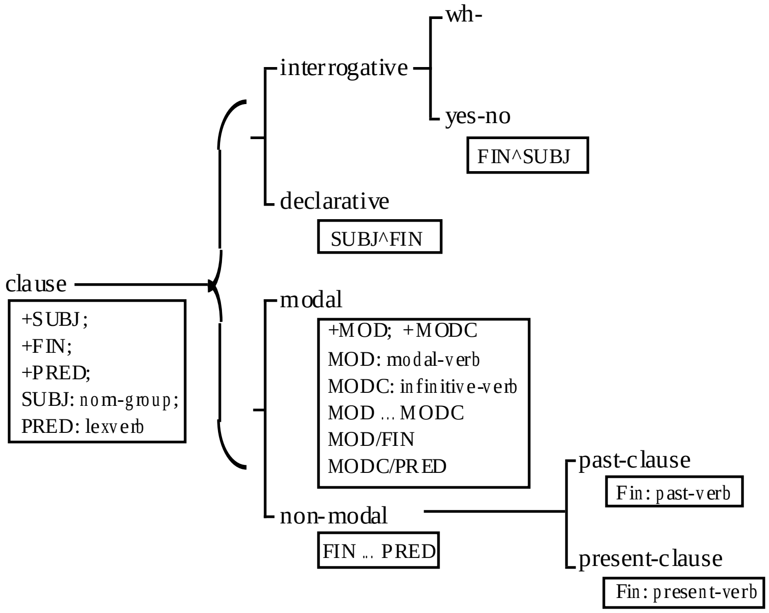


Figure 2.7: A Fragment of a Small Systemic Lexico-Grammar

The realisation constraints were described in section 3 of this chapter. I will briefly summarise those operators which appear in this example:

+Role   Role must appear in the role structure.

Role: feature   The role must be filled by a unit of the designated type.

Role1/Role2   The two roles conflate

Role1 ^ Role2   Role1 is immediately prior to Role2

Role1 ... Role2   Role1 occurs somewhere before (but not necessarily adjacent to) Role2.

Systemic processing mostly consists of combining the realisations of a unit's features (its *selection expression)*. In the example of figure 2.6, the clause is assigned four features: *[clause: modal: interrogative: yes-no]*. These features, in combination, place the following structural constraints on the unit:

clause: +Subj; +Fin; +Pred; Subj: nom-group; Pred: lexverb

yes-no: Fin ^ Subj

modal: +Mod;   Mod: modal-verb
+ModC; ModC: infinitive-verb
Mod ... ModC; Mod/Fin; ModC/Pred

I will show how these constraints combine in a simple generation example. Assume a two stage process, where we firstly apply the ordering constraints, and secondly apply the preselection realisations to determine the filler of each unit[5].

1) **Constraints on Role Placement**: the ordering, partition and conflation constraints are combined:

Fin^Subj + Mod/Fin => Mod/Fin ^ Subj
+ Mod...ModC => Mod/Fin ^ Subj ... ModC

[5] We would then repeat the same process for each of the constituents of the clause. See further discussion in chapter 11 on generation.

	Will	*the man*	*eat*	*the apple*	*today?*
Transitive		Actor	Process	Goal	Temporal-Location
Ergative		Agent	Process	Medium	Circumstance
Mood$_1$	Mood		Residue		
Mood$_2$	Finite	Subject	Pred	Object	Adjunct
Theme	Rhe-	Theme	-me		

Figure 2.8: Analysis showing Functional Layering

+ ModC/Pred => Mod/Fin ^ Subj ... ModC/Pred

Since this accounts for all of the inserted items, it can be assumed that there are no items between Subj and Pred. The final role ordering is thus:

Mod/Fin ^ Subj ^ Pred/ModC

...which is the same ordering as in figure 2.6.

It becomes obvious from this example why Mod and ModC are partitioned with respect to each other rather than ordered -- there is a possibility that they are not adjacent, that the Subject will fall between.

2) **Constraints on Role Fillers:** the preselection statements for each *role-bundle* are combined. A role-bundle is a set of conflated roles, e.g., the set of roles which a single item serves. Below we see the feature preselections for each role or role-bundle:

SUBJ: nom-group

MOD: modal-verb

PRED: lexverb + MODC: infinitive-verb

=> PRED/MODC: [lexverb:infinitive-verb]

A full Systemic-functional grammar (such as in Halliday (1985), or the Nigel grammar) assigns many more layers of structure at clause level. Figure 2.8 shows a typical analysis of a clause with five layers of structure.

My own approach has been to simplify the lexico-grammatical analysis, to make the parsing task easier. Lexico-grammatically, I deal only with the ergative and Mood$_2$ analysis. The information revealed in the other layers of representation has been shifted to the semantic stratum. For instance, Theme is part of the textual layer of the micro-semantic structure, not a role assigned to a grammatical unit (see chapter 5). The transitive layer is dealt with in the ideational representation, also part of the micro-semantics (see chapter 3). These patterns have, in short, been raised to a higher stratum, thus avoiding a certain redundancy which the Penman resources exhibit, in that the categories of the transitivity and theme analyses are repeated in the micro-semantic representation, e.g., the following roles are represented on both strata: Actor, Senser, Phenomenon, Theme, etc.

5. Graphological Resources

As phonology represents the meaningful structuring of speech, graphology represents the meaningful structuring of writing. These are two alternative media for the realisation of the lexico-grammar. Graphological resources tell how a text can be constructed “on the page”. They tell us that the first character of a sentence is capitalised; that a particular class of characters (punctuation) must end a sentence; that words consist of alphabetical characters and hyphens, etc. In generation, graphology is important to ensure that the text-output is properly formatted. In analysis, we need to analyse the text into its graphological constituents and prosodies before lexico-grammatical and semantic analysis can begin. For a fuller view of Systemic graphology, see Sefton (1990, 1992).

A graphological representation is structured in two ways, which I label segmental and supra-segmental graphology, following Sefton (1992) (alluding to the parallel terms in phonology):

1) **Segmental Graphology**: the breaking-down of the text in terms of parts and wholes, e.g., section into paragraphs, paragraphs into sentences, sentences into graphological-words and graphological-words into characters.

2) **Supra-Segmental Graphology**: Various graphological prosodies apply across multiple segments, potentially ignoring constituent boundaries (e.g., font style and size).

5.1 Segmental Graphology

Figure 2.9 shows a potential rank-scale for the graphological stratum, and figure 2.10 demonstrates a graphological analysis based on this scale. Note from this diagram that the graphological structure is distinct from the text as it appears on the page. The graphological representation is a structure of units, each node being a bundle of features only. The text that appears on the page is the equivalent of the acoustic form in phonology. While we might visually see 'a', in terms of the graphology we have *[a: lowercase: plain: times: 12]* [6].

Rank Scale	character	graphological-word	sentence	paragraph	section

Figure 2.9: The Graphological Rank Scale

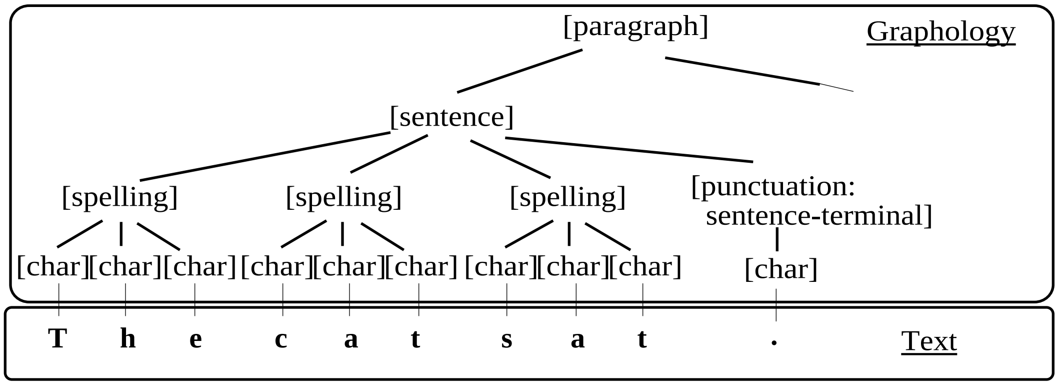


Figure 2.10: Graphological and Text Representations

5.1.1 Characters

The core of a character-rank network might look like that in figure 2.11. The network should be extended until the leaves are all actual characters. It is helpful to include classification of alphabetical characters in terms of their effect on morphological rules. For instance, some consonants double when adding a suffix, e.g., “trotting”, while others don’t, e.g., “paying”.

[6] Sefton (1992) distinguishes also graphetics, the writing equivalent of phonetics, which deals with non-meaning bearing differences between text instances.

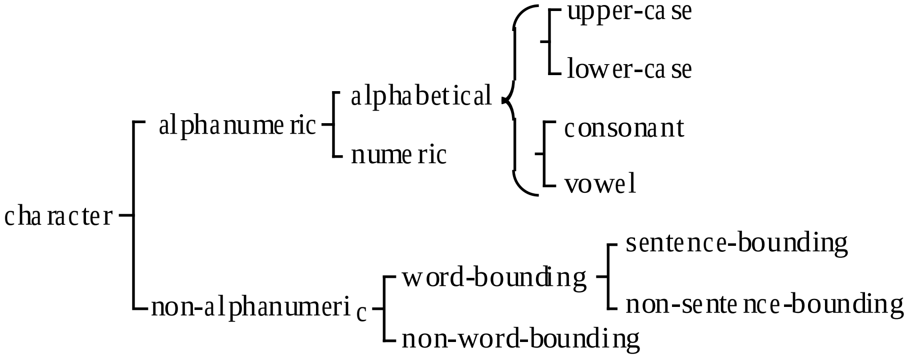


Figure 2.11: A Partial Character System Network

5.1.2 Graphological-words

A group of characters forms a graphological-word. A distinction needs to be made here between lexical items and graphological-words. “can” is a graphological-word. Two lexical-items (at least) use this spelling. These lexical-items can be given unique identifiers, e.g., *can-noun* (“the tin can”) and *can-modal* (“I can run”). These lexical-items have distinct semantic and grammatical patterning, but share the same spelling. This issue will be discussed further in section 6 in this chapter. Lexical-items may consist of multiple graphological-words, e.g., *New Zealand* is a single lexeme spelt using two graphological-words.

Punctuation: Although punctuation corresponds to prosodies in the phonological representation (e.g., tone-contours), punctuation marks are not prosodies in a graphological analysis. They are constituents of the sentence, and are treated here as a type of graphological-word.

Punctuation units are in general made up of single characters, e.g., commas, full-stops and quotation marks. Some are made up of several characters, for instance, “...”, indicating missing text in a quote, is a single punctuation mark. Note however that a sequence of punctuation characters does not usually make up just one punctuation-token. For instance “?).” occurring at the end of “the cat died (had it been fed?).” is three punctuation-tokens.

5.1.3 Sentences

The next graphological unit is the sentence, which is composed of graphological-words . As stated in the introduction chapter, ‘sentence’ refers to a graphological unit, which typically realises a grammatical clause or clause-complex. The graphological resources also needs to have knowledge about the layout of some special sentences, for instance, bulleted lists.

5.1.4 Paragraph, Section, Chapter

The text can also be analysed in terms of paragraphs and sections. These constituents are important for identifying macro-semantic relations, such as rhetorical structuring of the text. However, since this thesis focuses on single-sentence analysis, they will not be discussed here.

5.1.5 Graphological analysis

Graphological analysis is the process of segmenting a string of characters into graphological-words, sentences, etc. For instance, the following is a typical graphological analysis of the first sentence of a text, producing a list of graphological-words:

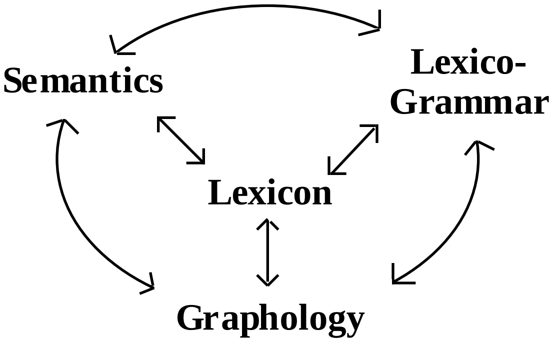


Figure 2.12: The Lexicon at the Inter-Stratal Hub

("a" "DASD" "dataset" "can" "be" "created" "by" "the" “specification" "of" "NEW" "in" "the" "DISP" "parameter" "of" "a" "DD" "statement" "in" "a" "job" "control" "statement" <period>)

A graphological anlayser also needs to treat capitalisation: a graphological-word at the beginning of the sentence is de-capitalised, unless it is available as a proper noun in the lexicon.

The WAG graphological analyser also performs partial lexical analysis, at least in regards to grouping together multi-word lexical items. For instance, if there is an item in the lexicon with spelling "job control statement", then the final three words of this analysis would be returned as a single graphological word. This approach also allows us to use a phrasal lexicon (e.g., treating "to kick the bucket" as a single lexical item), although this facility has not yet been used much within the WAG system.

5.2 Supra-segmental Graphology

There are also some non-constituent graphological patterns, such as ±bold, ±italic, ±underline, font style, font size, etc. Both in generation and analysis these prosodies can be important since they provide additional meaning about the text (focus, importance, indicating a technical term being introduced, etc.). I will not discuss these patterns further.

6. The Lexicon

The lexicon is an interstratal resource -- it relates graphology (graphological-words) to lexico-grammar (through lexical features) and semantics (through semantic features). The lexicon should not be seen as a resource of any one stratum, but rather the hub through which the various representations voice their constraints on other strata (see figure 2.12).

6.1 Terminology

Before describing the structure of the lexicon, and associated data structures, some clarification of terminology is useful:

graphological-word: A graphological unit -- constructed from a set of characters. Graphological-words need to be distinguished from grammatical words, which are lexico-grammatical objects.

lexical item: an abstraction over a set of graphological-words which share the same core ideational features, but differ in lexico-grammatical distribution (a lexical-item has a number of inflectional forms). In my usage, the term *lexeme* and lexical-item are interchangeable. The set of graphological-words which make up a lexical-item tend to be closely related in their character composition (e.g., “shoot”, “shooting”), but exceptions exist (e.g., “go”, “went”).

lexical feature: a classification of lexemes based on the grammatical distributional patternings of lexical items (e.g., verb, noun, common-noun, transitive-verb).

inflectional (or morphological) features: a subset of lexical features which are often realised graphologically (phonologically), e.g., by recurrent character strings “-ing”, “-s”, etc. (by recurrent phonemic patterns). Note that in this implementation, while inflectional features are realised graphologically, they are lexical features first, included only because they reflect differences in grammatical distribution.

6.2 The Lexicon and Lexical Entries

The lexicon is just a list of descriptions of lexical items. Each lexical item has one entry in the lexicon. The WAG system uses the basic Penman form for lexical entries, excepting that the WAG entry shows the semantic features of the lexeme (Penman allows only one semantic feature per lexeme, and this is represented in a separate resource). An example lexical-entry is shown below, a description of its fields follows:

```
(lexical-item
   :name ADDENDUM
   :spelling “addendum”
   :exception-spellings  ((PLURALFORM “addenda” ))
   :sample-sentence  “Here's an addendum”
   :grammatical-features (noun not-nominalisation common
                          countable nonsubstitute)
   :semantic-features (object decomposable nonconscious))
```

name: The lexical-identifier of the lexeme -- a unique key.

grammatical-features: the grammatical distributional features of the lexical-item.

semantic-features: The experiential concept(s) which this lexical-item realises. Since there may be multiple senses of the lexical item, this could be a list of lists. In the Penman system, each lexeme sense realises only a single concept (“man” must realises a single concept for *man*). In the WAG system, a lexeme realises concept-bundles (“man” realises -- as one sense -- [human: male: adult]).

spelling: the graphological form (spelling) of the unmarked form of the lexical item (the ‘root ‘spelling).

exception-spellings: for each inflectional form whose spelling is not predictable from the root spelling, using WAG’s morphological generator, this field provides the spelling.

sample-sentence: an example of how the lexical item is used. This field helps the user identify which sense of the word was intended. Note: any of the inflectional forms of the lexical item can be used in the example.

6.3 WAG’s Morphological Generator

The lexicon entries show only the root form of the lexeme. Morphological (Spelling) rules are used to generate the graphological spelling of the inflectional variants of the lexeme. Note that these rules are at present implemented procedurally -- as a morphological generating function. The rules embedded in this process are shown in table 2.3.

Class	Root Spelling	Inflection Class	Inflection Spelling
verb	-Cy	3rd/present/sing	-y => -ies
		past	-y => -ied
		v-en	-y => -ied
		v-ing	+ -ing
	-sh/-ch/-s/-z/-x	3rd/present/sing	+ -es
		past	+ -ed
		v-en	+ -ed
		v-ing	+ -ing
	-e	3rd/present/sing	+ -s
		past	+ -d
		v-en	+ -d
		v-ing	-e => -ing
	-VC$_1$	3rd/present/sing	+ -s
		past	+ -Ced*
		v-en	+ -Ced*
		v-ing	+ -ing
	otherwise	3rd/present/sing	+ -s
		past	+ -ed
		v-en	+ -en
		v-ing	+ -ing
noun	-Cy	plural	-y => -ies
	-sh/-ch/-s/-x/-y	plural	+ -es
	otherwise	plural	+ -s

Table 2.3: The Morphology Table

Key:

C	Consonant
C$_1$	Consonants except w, x and y
V	Vowels
*	consonant duplication takes place.

There are however many irregular forms, such as is/be/was. Where inflection forms cannot be predicted from the spelling of the root form, there is a field in the lexicon entry recording the spelling for that inflection. For an example, refer to the *:exception-spelling* field in the example above.

7. Summary of Micro-Resource Model

In this chapter I have looked at linguistic representations ranging over single sentences -- micro-linguistic representations, and the resources behind these representations. A model using three strata of linguistic representation was described: graphology, lexico-grammar and semantics. Two of these, lexico-grammar and graphology, were described, and also the lexicon, an inter-stratal resource.

The next chapters will look at the three strands of the micro-semantics -- micro-ideational, micro-interaction (move analysis) and micro-textual. I will then provide a description of the interstratal mapping formalism.