Chapter 4

Parsimonious Vole grammar

Now that the main theoretical foundations have been covered, I describe the structure of grammatical units and system networks adopted in this thesis and implemented in Parsimonious Vole parser. Some of them are from Sydney and others from Cardiff grammars. There are many common parts but also differences in parts of their paradigmatic and syntagmatic descriptions.

First I discuss on the structural differences between main units in Sydney and Cardiff grammars: the clause, the verbal group, nominal group, the adjectival and adverbial groups. Then focus on two important system networks: TRANSITIVITY and MOOD. The first is adopted from Cardiff grammar and the second belongs to Sydney grammar.

4.1 Grammatical units

The general principle for selection is that some unit structures are closer to the traditional syntactic analysis and thus possible to connect the elements of the Dependecy grammar. There are some units in both Sydney and Cardiff grammars that fit this purpose and some others that are semantically grounded and are more difficult to capture in structural variance and require additional lexical-semantic resources. This section discusses choices made for the current work.

4.1.1 Verbal group and clause boundaries

In Sydney Grammar the verbal group is described as an expansion of a verb just like the nominal group is the expansion of the noun (Halliday & Matthiessen 2013: 396). There are certainly words that are closely related and syntactically dependent on the

verb all together forming a unit that functions as a whole. For example the auxiliary verbs, adverbs or the negation particles are words that are directly linked to a lexical verb. The verb group functions as Finite + Predicator elements of the clause in Mood structure and as Process in Transitivity structure.

In Cardiff Grammar the verb group is dissolved moving the Main Verb as the pivotal element of the Clause unit. All the elements that form the clause structure and those that form the verb group structure are brought up together to the same level as elements of a clause. The clause structure in Cardiff Grammar comprises elements with clause related functions (like Subject, Adjunct, Complement etc.) and other elements with Main Verb related functions(Auxiliary, Negation particle, Finite operator etc.).

Regarded from the Hallidayan rank scale perspective, merging the elements of the verb group into clause structure is not permitted because the units are at different ranks. However it is not a problem for the relaxed rank scale version presented in Section 3.4.1. The reason for adopting such an approach is best illustrated via complex verb groups with more than one non-auxiliary verb such as in Example 34–36.

I begin by addressing the impact of this merger on (a) the clause structure (b) the clause boundaries and (c) semantic role distribution within the clause.

(34) (The commission **started to investigate** two cases of overfishing in Norway.)

(35) (The commission **started** (**to investigate** two cases of overfishing in Norway.))

(36) (The commission **started** (**to finish** (**investigating** two cases of overfishing in Norway.)))

In Sydney Grammar "started to investigate" (in Example 34) is considered a single predicate of investigation which has specified the aspect of event incipiency despite the fact that there are two lexical verbs within the same verbal group. The "starting" doesn't constitute any kind of process in semantic terms but rather specifies aspectual information about the investigation process. This is argued by looking at the conditions on participants and it is equivalent in a formal approach to looking at where the selection restrictions for complements come from. The boundaries of the clause governed by this predicate stretch to the entire sentence.

Semantically it is a sound approach. Despite the presence of two lexical verbs there is only one event. However, allowing such compositions leads to unwanted syntactic analysis for multiple lexical verb cases in examples such as 36. To solve this kind of problem Fawcett dismisses the verb groups and merges their elements into clause structure. He proposes the syntactically elegant principle of *one main verb per clause* (Fawcett 2008). Applying this principle to the same sentence yields a structure of two

Applying this structure to the previous example yields analysis such as in Table 3.12. The nominal group has the Epithet element filled by a coordination group formed of two Conjuncts and a Linker.

the	*immediate*	*and*	*not*	*so*	*far*	*distant*	*future*
Determiner	Epithet						Head
	Initiating Conjunct	Linker	Conjunct				

Table 3.12 Example analysis with coordination unit complex structure

Adopting the unit complex and in particular coordination unit requires two more clarifications: (1) does the complex unit carry a syntactic class, and if so according to which criteria is it established? (2) Does it have any intrinsic features or all of them are inherited from the conjuncts?

Zhang states in her thesis that the coordinating constructions do not have any categorial features thus there is no need to provide a new unit type. Instead the categorial properties of the conjuncts are transferred upwards (Zhang 2010). For example if two nominal groups are conjuncted then the complex receives the nominal class.

This principle holds for most of the cases however there are rare cases when the units are of different classes. Consider 31 analysed in Table 3.13 where the conjuncts are a nominal group "last Monday" and a prepositional group "during the previous weekend".

(31) I lost it (either last Monday or during the previous weekend).

either	*last*	*Monday*	*or*	*during*	*the*	*previous*	*weekend*
Pre-Linker	Initiating Conjunct		Linker	Conjunct			

Table 3.13 Coordination group with mixed class conjuncts

In this case there are two unit types that can be raised and it is not clear how to resolve this case. Options are (a) to leave the generic class *coordination complex*, (b) transfer the class of the first unit upwards, or (c) semantically resolve the class as both represent temporal circumstances even if they are realised through two different syntactic categories. This means that if no sub-classification is provided based on the constituent units below than there is no need to project/transfer upward the class of

the conjunct units. In this work I decided to leave the class generic and leave for the future an extensive unit complex classification.

I turn now to the last issue of this discussion, specifically whether the complex unit may have intrinsic features emerging from the conjunct elements.

In the example 32 the conjunction of two singular noun groups requires plural agreement with the verb. Even though semantic interpretation that only one item is selected at a time, syntactically both items are listed in the clause and attempting third person singular verb forms in 33 is grammatically incorrect. That leads to the conclusion that the coordination complex can have categorial features which none of the constituting units has.

(32) A pencil or a pen **are** equally good as a smart-phone.

(33) * A pencil or a pen **is** equally good as a smart-phone.

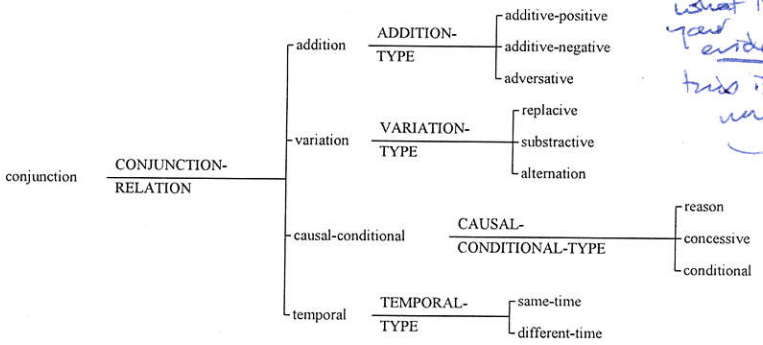


Fig. 3.11 Systemic network of coordination types

In the case of nominal group conjunction we can see that the plural feature emerges even if each individual unit is singular. For other unit classes it is not so obvious whether there are any linguistic features that emerge at the conjunction level. The meaning variation is semantic as for example conjunction of two verbs or clauses might mean very different things such as consecutive actions, concomitant actions or presence of two states at the same time and so on. This brings us to another feature of the coordination complex - the type of the relationship it constructs. The lexical items expounding the Linker and Pre-Linker (e.g. *and*, *or*, *but*, *yet*, *for*, *nor* or *so*) are indicator of relationship among the conjuncts and together can be systematised as the relationship types in the systemic network in Figure 3.11.

In this work I adopt the principles for establishing the logical structure of Sydney Grammar. It resonates closely with the traditional "semantically blinded" grammars because and it always provides a Head element even if it differs from the syntactically motivated pivotal element in Cardiff Grammar. Moreover these logical Heads correspond to dependency heads established in the Stanford dependency parse. Chapter 5 provides the grounds for cross-theoretical mappings and the empirical evaluation in Chapter 10 validates it.

In language is not unusual to have nominal groups with the Thing missing or elliptic clauses with the Main verb missing, therefore no rigid correspondence can be established between the logical Head and unit class. In this thesis the structure creation is performed in two steps first establishing the group boundaries and the unambiguous unit elements through a top down perspective (that is Sydney approach to unit creation) and second for each established group evaluate the internal structure in order to establish the group class (that is Cardiff approach to group formation). This process is detailed in the Chapter 8.

The evaluation in the second step, besides finalising the syntactically motivated unit structure, can as well assign semantically motivated unit structure. This part however is left out in the current thesis for the groups and only the clauses receive semantic role labels and process types described in Chapter ??.

3.4.6 Coordination as unit complexing

In Sydney Grammar unit complexes fill an important part of the grammar along with the *taxis relations* (Definition 3.2.14) which express the interdependency relations in unit complexes. *Parataxis* relations bind units of equal status while the *Hypotaxis* ones bind the dominant and the dependant units. Fawcett bypasses the taxis relations replacing them with coordination and embedding (Fawcett 2000: 271) leading to abandonment of unit complexing entirely. While embedding elegantly accounts for the depth and complexity of syntax, this approach to coordination is problematic.

Hereafter I discuss the utility and even necessity of keeping unit complexes in parsing. In particular I address the treatment of group and clause *coordination* but the same principle applies to fixed idiomatic structures such as *comparatives*, *conditionals* or *appositions*.

Treatment of the coordination phenomena is a challenge not only for SFL but for other linguistic theories as well. Sydney Grammar approaches it through unit complexing and taxis relations while Cardiff Grammar treats this phenomena as multiple distinct units filling or expounding the same element.

Table 3.8 illustrates an example analysis where the Complement is filled by a homogeneous nominal group complex held together through *paratactic extension* where the first element is a nominal group and the second is nominal group together with the conjunction which is not part of the experiential structure but remains accounted only in the logical structure of the nexus.

Ike	*washed*	*his*	*shirt*	*and*	*his*	*jeans*
Subject	Predicate/Finite	Complement				
		1			+2	
		Deictic	Thing		Deictic	Thing

Table 3.8 Clause with nominal group complex

In Table 3.9 the Epithet is filled by a nexus of paratactic extension. The first element of the nexus is the word "immediate" and the second element is the set of words "and not so far distant". The "not so far distant" is an adverbial group with a logical structure of sub-modifiers already discussed in Section 3.4.1 and the conjunction "and" is left implicitly part of the logical structure of the nexus creating a gap in the structure that is addressed in this discussion. Also note that, in Sydney Grammar the coordination is accounted as unit complex ensuring that only one unit fills an element of the parent, in contrast. as we will see below, to Cardiff Grammar.

the	*immediate*	*and*	*not*	*so*	*far*	*distant*	*future*
Modifier							Head
γ	β						α
Deictic	Epithet						Thing
	1	+2					
		Sub-Modifier				Sub-Head	
		δ	γ	β		α	

Table 3.9 Nominal group with word complex from (Halliday & Matthiessen 2013: 564)

In Table 3.10 is presented an example of analysis with Cardiff Grammar. The Complement is filled by two *sibling* nominal groups "his shirt" and "and his jeans" that are both of them fill the same element in accordance to Definition refdef:coordination. The conjunction "and" is accounted directly as part of the nominal group structure.

Opinions are divided (between Sydney and Cardiff schools) whether to invite the notion of a complex unit to handle coordination or not. If we side with Cardiff grammar and dismiss the unit complex then we allow an element to be filled by more than

3.4.4 Unit classes

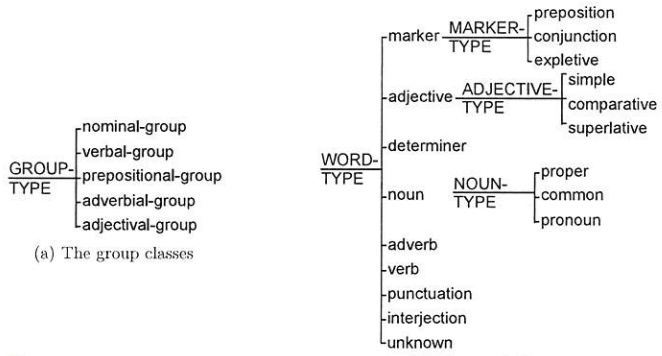
In SFL at large there is the consensus that linguistic forms and meanings are intertwined and mutually determined just like for any sign in a Saussurean approach to language. Both Halliday (quote below) and Fawcett (Definition 3.3.1) adopt this position.

> ...something that is distinctly non-arbitrary [in language] is the way different kinds of meaning in language are expressed by different kinds of grammatical structure, as appears when linguistic structure is interpreted in functional terms (Halliday 2003a).

When it comes to establishing the lexicogrammatical classes the two schools diverge. Halliday adopts the traditional grammar *word classes* or *parts of speech*: noun, verb, adjective etc. He then derives a set of groups (e.g. nominal group, verbal group, adverbial group etc.) that share properties of the word classes. In fact the class, in Halliday's words, "indicates the in a general way its potential range of grammatical functions" (Halliday & Matthiessen 2013: 76). For example the nominal group is a formation that functions as a noun may do and expresses same kind of meaning.

Following the idea that major semantic classes of entities (situations, things, qualities and quantities) correspond to the major syntactic units, Fawcett decided to mirror them into the lexicogrammar. This lead to a semantically based classification of syntactic units: clause, nominal group, prepositional group, quality group and quantity group (Fawcett 2000: 193–194) along with a set of minor classes such as genitive and proper name clusters. This is, in a way, a tight coupling of the grammatical units with an ontology which may be subject to change in the future. The converse may also be stated that the traditional part of speech are disconnected from the semantics in the sense that there is no one to one correspondence (as Fawcett attempts) but rather complex set of mappings. Establishing the exact interface of syntax and semantics is a hot ongoing theoretical exploration across the entire linguistic discipline a difficult task in practice. This discussion however is beyond the scope here.

In the current work I side with the Sydney classification of syntactic units that is close in line with traditional syntactic classifications (Quirk et al. 1985). I adopt the clause as a unit plus the four group classes of the Sidney grammar depicted in Figure 3.10.



(a) The group classes

(b) The word classes

Fig. 3.10 The group and word classes

The word classes or part of speech tags that I adopt here are the ones employed to annotate Penn Treebank corpora called the Penn tag set (Marcus et al. 1993) which, like Sidney unit classes, are also in line with the traditional grammar. This tag set has become a widely accepted standard in mainstream computational linguistics and there are multiple implementation of the part of speech taggers. The Stanford Parser which plays an important role in the software implementation of this thesis and is described in the Chapter 5, employs precisely the Penn tag set.

The Penn tag set was developed to annotate the Penn Treebank corpora (Marcus et al. 1993). It is a large, richly articulated tag set that provides distinct codings for classes of words that have distinct grammatical behaviour.

The Penn tag set is based on the Brown Corpus tag set (Kucera & Francis 1968) but differs in several ways. First, the authors reduced the lexical and syntactic redundancy. In the Brown corpus there are many unique tags to a lexical item. In the Penn tag set the intention is to reduce this phenomenon to a minimum. Also distinctions that are recoverable from lexical variation of the same word such as verb or adjective forms or distinctions recoverable from syntactic structure are reduced to a single tag.

Second the Penn Corpus takes into consideration the syntactic context. Thus the Penn tags, to a degree, encodes syntactic functions when possible. For example, *one* is tagged as NN (singular common noun) when it is the head of the noun phrases rather than CD (cardinal number).

modifiers and heads leading to a logical structure analysis as the one in Table 3.3. In such a structure the modifier is further broken down into a Sub-Head and Sub-Modifiers.

some	*very*	*small*	*wooden*	*ones*
Modifier				Head
δ	γ		β	α
	Sub-Modifier	Sub-Head		
	$\gamma\beta$	$\gamma\alpha$		

Table 3.3 Sydney logical structure analysis of Example 22

The corresponding experiential structure analysis is provided in the Table 3.4 Halliday & Matthiessen (2013: 391). Accordingly, the Epithet "very small" is composed of a quality adjective "small" and an enhancer modifier "very".

some	*very*	*small*	*wooden*	ones
Deictic	Epithet		Classifier	Thing
	Sub-Modifier	Sub-Head		

Table 3.4 Sydney experiential analysis of Example 22

As you can see the elements are further broken down into sub-elements composing in a way a structure of their own. This is possible because of the poly-structural and multi functional approach to text analysis which in this case leads to a complex structure of a nominal group. This kind of intricate cases can be simplified through the permission that elements of a group to be filled by other groups or expounded by words. This way, instead of having a sub-modifier construction simply consider that the Epithet is filled by an adjectival or nominal group which in turn has its own structure. Please note that I mention adjectival or nominal group because in Sydney grammar the adjectival group is considered as a nominal group with covert Thing where the Epithet acts as Head; this however is a discussion beyond the point I make here.

The same example analysed with Cardiff grammar would look like in Table 3.5. It follows precisely the above suggestion of filling the Epithet with another unit, in this case a Quality Group which in turn has its own internal structure.

some	*very*	*small*	*wooden*	*ones*
Quantifying Determiner	Modifier		Modifier	Head
	Quality Group			
	Degree Tamperer	Apex		

Table 3.5 Cardiff analysis of Example 22

(23) Indians had originally planned to present the document to President Fernando Henrique Cardoso.

Indians	*had*	*originally*	*planned to present*	*the document*	*to President Fernando Henrique Cardoso*
Mood		Residue			
Subject	Finite	Adjunct	Predicator	Complement	Adjunct
nominal group		adverbial group	verbal group	nominal group	prepositional phrase

Table 3.6 Sydney grammar Mood analysis of Example 23

Another case that deems the rank scale constraints too strict for the present work is in the case of Finite element in the Clause. Consider example 23 where the Finite and Predicator elements are filled by a single unit which is the verbal group which is against the constituency principles which restricts the composition relation to engage only with whole units.

Alternatively, if the unit filling the Finite element is considered separate from the verbal group filling the Predicator then it is always a single word, a modal verb, and never a verbal group. This again is a breach in the rank scale constraints which postulates that a unit may be composed of units of equal rank or a rank higher and cannot be composed of units that are more than one rank lower thus it is not permitted to have clause elements expounded by words directly.

The two cases above I use to demonstrate how the ranks scale construct as defined by Sydney grammar is too rigid and thus unsuitable for the current work. I drop the constituency constraints hence allowing the flexibility for elements to be filled by other units or, in other words, allow unit *embedding*. This approach removes the need of sub-structures in the unit elements reducing thus the structural complexity as seen in Table 3.5.

The weakening of constituency constraints makes embedding a normal (broadly defined in Definition 3.3.9) rather than an exceptional phenomena (strictly defined in Definition 3.2.5).

An approach to describe units outside the rank-scale was suggested by Fawcett (2000) and Butler (1985). Fawcett proposes replacing it with the filling probabilities to guide the unit composition simply mapping elements to a set of legal unit classes that may fill it. Units are carriers of a grammatical patterns they can be described in terms of their internal structure instead of their potential for operation in the unit above. Nonetheless I do not abandon the rank scale completely and I use it as the top level classifier of grammatical units (see Figure 3.10) falling in line with more traditional syntactic classes.

If in Sydney model exponence (Definition 3.2.12) is a relation that links abstract grammatical categorise to the data, in Cardiff model it has a restricted meaning referring to relation between items and elements only.

3.3.4 Componence and obscured dependency

Definition 3.3.5 (Componence). Componence is the part-whole relationship between a unit and the elements it is composed of (Fawcett 2000: 244).

Note that componence is not a relationship between a unit and its places; the latter, as discussed in Section 3.3.2, simply locationally relate elements of a unit to each other.

Componence intuitively implies a part-whole constituency relationship between the unit and its elements. But this is not the only view. Another perspective is the concept of *dependency* (which I will address in Chapter 5) or strictly speaking the *sister* or *sibling dependency* (not parent-daughter). It is suitable for describing relations between elements of structure within a unit.

(17) the man with a stick

For example the componence of nominal group in Example 17 is (*dd h q*) which are symbols for (*determiner head qualifier*). The same can be expressed in terms of sibling dependency relations depicted in Figure 3.7. The relations from *stick* to *with a* are not depicted because they belong in description of prepositional group *with a stick*.

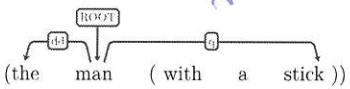


Fig. 3.7 Sibling dependency representation for "the man with a stick"

In both SFL theories, the sister dependency relations is considered a by-product or second order concept that can be deduced from the constituency structure thus unnecessary in the grammar model. I will come back to this point because current work relies on this dual view on elements of structure and relation to the whole unit.

The (supposed) dependency relation between a modifier and the head in the framework of SFG is not a direct one. Simply assume that what modifier modifies is the head. Here, however, the general function of the modifiers is to contribute to the meaning of the whole unit which is anchored by the head.

In the nominal group from Example 17, the *determiner* and *qualifier* are modifiers that contributes to the description of the referent stated by the *head*. So the head

realises one type of meaning that relates the *referent* while modifier realises another one. Both of them describe the referent via different kinds of meaning, therefore, according to Fawcett, they are related indirectly to each other because the modifier does not modify the head but the referent denoted by the head. From this point of view, whether the element is dependent on a sibling element such as the head or on the parent unit is beside the point because in syntax we can observe its realization in system networks (Fawcett 2000: 216–217). Next I move towards last concept in Cardiff model, *filling*, which is a relation between the elements of structure and the units below.

3.3.5 Filling and the role of probabilities

Definition 3.3.6 (Filling). Filling is the probabilistic relationship between a element and the unit lower in the tree that operates at that element (Fawcett 2000: 238, 251).

Fawcett replaces the rank scale with the concept of *filling probabilities*. The probabilistic predictions are made in terms of filling relationship between a unit and an element of structure in a higher unit in the tree rather than being a relationship between units of different ranks. This moves focus away from the fact that a unit is for example a group, and towards what group class it is.

In this line of thought, some elements of a clause are frequently filled by groups, but some other elements are rather *expounded* by items. The frequency varies greatly and is an important factor for predicting or recognizing either the unit class or the element type in the filling relationship.

Filling may add a *single unit* to the element of structure or it can introduce *multiple coordinated units*. Coordination (Example 18) is usually marked by an overt *Linker* such as *and*, *or*, *but*, etc. and sometimes is enforced by another linker that introduces the first unit such as *both*.

Definition 3.3.7 (Coordination). Coordination is the relation between units that fill the same element of structure (Fawcett 2000: 263).

(18) she is (friendly, nice and polite)

(19) she is (very very) nice!

Coordination is through by Fawcett as being not between syntactic units but between mental referents. It always introduces more than one unit which are syntactically and semantically similar (somehow) resulting in a *syntactic parallelism* which often leads to *ellipsis*.

morphemes and phonemes. I will also refer to it when describing the Cardiff theory of grammar and also briefly in the discussion of dependency relations in Section 5.6.

The elements of logical paratactic structure are notated left to right with numbers (1 2) while those of hypotactic structure with Greek letters (β α) right to left. The tactic relations can be of two types: that of expansion which relates phenomena of the same order of experience and that of projection which relates phenomena of one order of experience (usually saying or thinking) to an order of experience higher (what is said or thought). Projection can be of two types: *idea* (' single quote) and *locution* (" double quotes).

Expansion is further divided into three: *elaborating* (= equals), *extending* (+ is added to) and *enhancing* (× is multiplied by). Elaboration is a way to restate the same thing, exemplify, comment or specify in detail. Extending is the way to add new element give an exception or offer an alternative. And finally enhancing is the way to qualify something with some circumstantial feature of time, place, cause, intensity or condition.

3.2.6 Lexis and lexicogrammar

In SFL the terms *word* and *lexical item* are not really synonymous. They are related but they refer to different things. The term *word* is reserved (in early Halliday) for the grammatical unit of the lowest rank whose *exponents* are lexical items.

Definition 3.2.15 (Lexical Item). In English, a lexical item may be a *morpheme*, *word* (in traditional sense) or *group (of words)* and it is assigned to no rank (Halliday 2002: 60).

Examples of lexical items are the following: "'s" (the possessive morpheme), "house", "walk", "on" (words in traditional sense) and "in front of", "according to", "ask around", "add up to", "break down" (multi word prepositions and phrasal verbs).

If some theories treat grammar and lexis as discrete phenomena, Halliday brings them together as opposite poles of the same cline. He refers to this merge as *lexicogrammar* where they are paradigmatically related through delicacy relation. Hasan (2014) explores the feasibility of what would it mean to turn the "whole linguistic form into grammar". This then implies a assumption that lexis is not form and that its relation to semantics is unique which in turn is challenging the problems of polysemy.

3.3 Cardiff theory of grammar

As presented in the introduction and explained by Bateman (2008), the accounts along the syntagmatic axis had gone missing in the Sydney grammar leaving unresolved how to best represent the structure of language at the level of form. This section present the theory of systemic functional grammar as conceived by Robin Fawcett at the University of Cardiff. His book "A theory of syntax for Systemic Functional Linguistics" (Fawcett 2000) presented a proposal for a *unified syntactic model* for SFL that contrasts several aspects of Hallidayan grammar but share the same set of fundamental assumptions about the language; it is an extension and a simplification in a way.

Fawcett questions the status of multiple structures in the theory and whether they can finally be integrated into a simpler sole representation. A big difference to Hallidayan theory is renouncing the concept of rank scale which has an impact on the whole theory. Another is the bottom-up approach to unit definition as opposed to top-down one advocated by Halliday. These two and a few other differences have important implications for the overall theory of grammar and consequently for the grammar itself. As a consequence, to accommodate the lack of rank-scale, Fawcett adapts the definitions of the fundamental concepts and changes his choice of words (for example "class" and "unit" turn into "class of unit" treated as one concept rather than two distinct ones).

Fawcett (2000) proposes three fundamental categories in the theory of grammar: *class of unit*, *element of structure* and *item*. Constituency is a relation accounting for the prominent compositional dimension of language. However a unit does not function directly as a constituent of another unit but via a specialised relation which Fawcett breaks down into three sub-relations: *componence*, *filling* and *exponence*. Informally it is said that a unit is composed of elements which are either filled by another unit or expounded by an item. He also proposes three secondary relations of *coordination*, *embedding* and *reiteration* to account for a more complete range of syntactic phenomena.

3.3.1 Class of units

Fawcett's theory of language assumes a model with two levels of *meaning* and *form* corresponding to *semantic units* and *syntactic units* which are mutually determined (which is the case for any sign in a Saussurean approach to language).

Definition 3.3.1 (Class of Unit). The class of unit [...] expresses a specific array of meanings that are associated with each one of the major classes of entity in semantics

is a part of language description but it is only a syntagmatic manifestation of the systemic choices and one needs to account for both (Halliday & Matthiessen 2013: 23).

Definition 3.2.10 (System). A system is a set of mutually exclusive set of terms referring to meaning potentials in language and are mutually defining. The system is considered self-contained, closed and complete with the following characteristics:

1. the number of terms is finite,
2. each term is exclusive of all others,
3. if a new term is added to the system it changes the meaning of all the other terms (Halliday 2002: 41).

The concept of a system as presented in Definition 3.2.10 has its roots in the works of Saussure (1959 [1915]) and Hjelmslev (1953) and Halliday only cements it in SFL architecture of grammar.

Going back to the notion of class previously defined as a grouping of items identified by functions in the structure, it needs stressed here that class is not a list of formal items but an abstraction from them. By increase in *delicacy* a class is broken into secondary classes.

Definition 3.2.11 (Delicacy). Delicacy is the scale of differentiation or depth of detail whose limit at one end is the primary degree of categories of structure and class and on the other end, theoretically, is the point beyond which no further grammatical relations obtain. (Halliday 2002: 58)

We say that a category is refined into more subtle distinctions of subcategories which form a system as define above. Subsequently those distinctions to subcategories can be further refined in other systems. This relationship between these two systems is one of delicacy where the second one is more delicate than the first one and together they form a *system network*.

The graphical notations introduced by Halliday & Matthiessen (2013) are useful in reading and writing system networks in this thesis. Below is a system network with a simple *entry condition* (Figure 3.3), a *system network grouping* that share the same entry condition (Figure 3.4), a system network with a *disjunctive* and *conjunctive* entry conditions (Figure 3.5 and 3.6).

Fig. 3.3 A system with a single entry condition: if a then either x or y

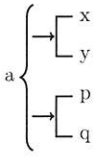


Fig. 3.4 Two systems grouped under the same entry condition: if a then both either x or y and, independently, either p or q

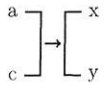


Fig. 3.5 A system network with a disjunctive entry condition: if either a or c (or both), then either x or y

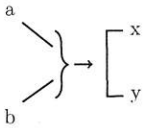


Fig. 3.6 A system with a conjunctive entry condition: if both a and b then, either x or y

It is worth noting that when a piece of language is analysed, it can be approached at various levels of delicacy. We say that delicacy is variable in description, and one may choose to provide coarse grained analysis without going beyond primary grammatical categories or it can dive into fine grained categorial distinctions, still being comprehensive with regards to the rank, *exponence* and grammatical categories.

Definition 3.2.12 (Exponence). Exponence is the scale which relates the categories of theory which are with high degree of abstraction to formal items on its low end. Each exponent can be linked directly to the formal item or by taking successive steps on the exponence scale and changing rank where necessary. (Halliday 2002: 57)

2. Each unit consists of *one or more* units of rank next below.
3. Units of every rank may form complexes.
4. There is potential for rank shift, whereby a unit of one rank my be down-ranked to function in a structure of a unit of its own rank or of a rank below.
5. Under certain circumstances it is possible for one unit to be enclosed within another, not as a constituent but simply in such a way as to split the other into two discrete parts (Halliday & Matthiessen 2013: 9–10).

For example, the down-ranking (Point 4) can be observed in nominal groups that incorporate a relative clause functioning as qualifier. In example 10 *that I got for Christmas* is a relative clause specifying which books are being referred. The unit split (Point 5) can be encountered in the instances of Wh-interrogative clauses containing a preposition at the end which in fact belongs to the Wh-group. In example 11 the prepositional phrase *Who ... about* is gapped and has an inverted order of constituents.

(10) I haven't read any books *that I got for Christmas*.

(11) *Who* are you talking *about*?

(12) I am talking about George.

The relation between units is that of consistency for which we say that a unit *consists of* other units. The scale on which the units are ranged is the *rank scale*. The rank scale is a levelling system of units supporting unit composition regulating how units are organised at different granularity levels from clause, to groups/phrases to words and the units of a higher rank scale consist of units of the rank next below. Table 3.1 presents a schematic representation of the rank scale and its derived complexes.

Rank scale ↓	**Complexing**
	Clause complex
Clause	
	Group(/phrase) complex
Group(/phrase)	
	Word complex
Word	
	(Morpheme complex)
(Morpheme)	

Table 3.1 Rank scale of the (English) lexicogrammatical constituency

Generalization 3.2.2 (Rank scale constraints). The rank relations are constrained as follows:

1. in general elements of clauses are filled by groups, the elements of groups by words and the elements of words by morphemes;
2. downward *rankshift* is allowed, i.e. the transfer of a given unit to a lower rank;
3. upward rankshift is not allowed;
4. only whole units can enter into higher units (Halliday 2002: 44).

The Generalization 3.2.2 taken as a whole means that a unit can include, in what it consists of, a unit of rank higher than or equal to itself but not a unit of rank more than one degree lower than itself; and not in any case a part of any unit (Halliday 2002: 42).

Following the rank scale constraints above, the concept of embedding can be defined as follows.

Definition 3.2.5 (Embedding). Embedding is the mechanism whereby a clause or phrase comes to function as a constituent within the structure of a group, which is itself a constituent of a clause (Halliday & Matthiessen 2013: 242).

Halliday states that embedding is a phenomena that occurs only when a phrase/-group or clause function within the structure of a group which is itself a constituent of a clause (Halliday 1994: 242). The above definition of embedding permits the only for a clause and groups that function as elements of groups which means that a clause cannot fill the elements of another clause (Fawcett 2000: 237).

3.2.2 Structure

Definition 3.2.6 (Structure). The structure (of a given unit) is the arrangement of *elements* that take places distinguished by the order relationship (Halliday 2002: 46).

Definition 3.2.7 (Element). Element is defined by the place stated as absolute or relative position in sequence and with reference to the unit next below (Halliday 2002: 47).

We say that a unit is composed of elements located in places and that its internal structure is accounted for elements in terms of functions and places taken by the lower (constituting) units or lexical items. The graphic representation of the unit structure

Interestingly enough, the Definition 3.1.3 makes no mention at all to the lexicon. This is because the formal grammars focus primarily on unit classes and how they are accommodated in various structures and so in formal linguistics the lexicon is often disconnected from the grammar. The systemic grammar, on the other hand, along with formal descriptions of grammatical categories and structures, includes the lexicon as part of grammar to form a *lexicogrammar*. At this point I have to mention that systemic functional grammar is not the only lexicalised one and there are others taking the same approach such as Lexical Functional Grammar (LFG), Head Phrase Structure Grammar (HPSG), Combinatory Categorial Grammar (CCG) and others.

Another important aspect to notice is that the grammar is defined as a field of study rather than a set of rules. De divergence in perspective on the subject led Halliday, since his early papers, to become conscious the difference between a study of a phenomenon with the phenomenon itself. By analogy to language as phenomenon and linguistics as the study of the phenomenon, discussed in (Halliday 1997), Halliday adopts the same wording for *grammar* as phenomenon and *grammatics* as the study of grammar; the same distinction holds for *syntax* and *syntactics*.

Definition 3.1.4 (Grammatics (Halliday)). Grammatics is a theory for explaining grammar (Halliday 2002: 369)

Moravcsik, another generative linguist, stresses the same distinction in her "An introduction to syntax" (Moravcsik 2006), and presents two ways in which the word *syntax* is used in the literature: (a) in reference to a particular aspect of grammatical structure and (b) in reference to a sub-field of descriptive linguistics that describes this aspect of grammar. In her words:

> ... syntax describes the selection and order of words that make well-formed sentences and it does so in as general a manner as possible so as to bring out similarities among different sentences of the same language and different languages and render them explainable. ... syntax rules also need to account for the relationship between string of word meanings and the entire sentence meaning, on one hand, and relationship between strings of word forms and the entire sentential phonetic form, on the other hand. (Moravcsik 2006: 25)

In her definition of grammar she includes the lexicon and semantics which is a somewhat more explicit statement than Radford's *interpretation*. She is also getting, in Definition 3.1.5, somewhat closer to what grammar stands for in SFL - Definition 3.1.6.

Definition 3.1.5 (Grammar (Moravcsik)). ... maximally general analytic descriptions, provided by descriptive linguistics, [are] called grammars. A grammar has five components: phonology (or, depending on the medium, its correspondent e.g. morphology), lexicon, syntax and semantics(Moravcsik 2006: 24–25).

Definition 3.1.6 (Grammar (Halliday)). To Halliday, lexico-grammar, or for short,simply grammar, is a part of language and it means the wording system - the "lexical-grammatical stratum of natural language as traditionally understood, comprising its syntax, vocabulary together with any morphology the language may display [...]" (Halliday 2002: 369).

The last point I want to mention is the approach to semantics. Formal grammars aim to account for the realisation variations, that is formation of words, phrases and sentences along with their arrangements and mention of semantics is often restricted to what may be termed the *formal aspect of meaning*.

By contrast, a systemic grammar is a functional grammar, which means (among other things) that it is semantically motivated, i.e. "natural". So the fundamental distinctions between formal and functional grammars is the semantic basis for explanations of structure.

Also, in SFL, the meaning is being approached from a semiotic perspective, placing the linguistic semantics in perspective with the linguistic expression and the real world situation. In this respect, Lemke (1993) offers a well formulated theoretical foundation that "human communities are eco-social systems that persist in time through ongoing exchange with their environment; and the same holds true for any of their sub-subsystems [...]" including language. The social practices constituting such systems are both material and semiotic, with a constant dynamic interplay between the two (Halliday 2002: 387).

To Halliday, the term *semiotic* accounts for an orientation towards meaning rather than sign. In other words, the interaction is between *the practice of doing and the practice of meaning*. As the two sets of practices are strongly coupled, Lemke points out that there is a high degree of redundancy between the *material-semiotic interplay*. And it perfectly resonates with Firth's idea of *mutual expectancy* between the text and the situation. This idea of interplay is incorporated in SFL as *language stratification* and is graphically represented in Figure 3.1.

Having that said, the stratification axis is a useful dimension to relate the formal and the systemic functional grammars. This is also an instrument employed by Hjelmslev (Tavernier 2011).

also created a lexical resource indicating for each word which elements can expanded it.

The parsing procedure is a simple lookup of words in the lexical resource selecting all possible elements it can expound and then selecting possible strips starting with the elements expounded by the word. Advancing from left to right, for each sentence word more strips compatible with the previously selected ones are selected within the collocation network constrains. The parser finds all possible combinations of strips composing parse trees representing possible output parses.

The corpus from which the vertical strips were extracted is 100,000 sentences large and was generated with Fawcett's natural language generation system and was tested on the same corpus leaving unclear how would the parser behave on a real corpus. In 98% of cases the parser returns a set of trees (between 0 and 56) that included the correct one with an average of 6.6 trees per parse.

Actually, using a larger corpus could potentially lead to a combinatorial explosion in the step that looks for vertical strips. It would decrease the accuracy of the parse because of the higher number of possible trees per parse.

2.4 Honnibal

Honnibal (2004; 2007) describes how Penn Treebank can be converted into a SFG Treebank. Before assigning to parse tree nodes synthetic features such as mood, tense, voice and negation he first transforms the parse trees into a form that facilitates the feature extraction.

The scope of SFG corpus was limited to a few Mood and Textual systems leaving aside Transitivity because of its inherently lexico-semantic nature. He briefly describes how he structurally deals with verb groups, complexes and ellipses as functional structures are much flatter than those exhibited in the original Treebank. Then he describes how are identified metafunctional features of unit class, mood function, clause status, mood type, polarity, tense, voice and textual functions.

The drawback of his approach is that the Python script performing the transformation does not derive any grammar but rather implements directly these transformations as functions falling into the same class of problems like Winograd's SHRDLU. By doing so the program is non-scalable for example in accommodation of larger grammars and knowledge bodies and unmaintainable on the long term as it becomes increasingly difficult to make changes.

2.5 Discussion

In this chapter were presented several relevant attempts to parse with Systemic Functional Grammars. All of them needed to reduce the grammar size or use toy grammars in order to compute results in a reasonable amount of time. The main problem in using SFGs for parsing is that they are much more complex than post-Chomskian grammars: the grammar contains both semantic and syntagmatic aspects of language, the system networks represent large numbers of simultaneous combinations of features, and multiple layers of function structure are conflated together.

Some parsing approaches use a syntactic backbone which is then flashed out with SFG description. Other ones use a reduced set or a single layer of SFG representation the third ones use an annotated corpus as the source of a probabilistic grammar. Regardless of the approach each limits the SFG in a one way or another balancing the depth of description with language coverage: that is either deep description but a domain specific language or shallow description but broad language coverage.

Current approach is aligned with works of Honnibal. Kasper and and O'Donnell with respect to using a backbone structure and enriching it with syntactic and semantic features. Current method employs rules for graph traversal in order to build a parallel backbone constituency tree and rules for graph matching to enrich it with systemic features.

Parsing Transitivity system is a task similar to Semantic Role Labelling and requires large lexicogrammatical resource describing verb meanings in terms of their process type and participant roles. O'Donnell approaches it by providing possible process types directly to the verb by employing self constructed lexicon where each word has syntactic and semantic features. Current approach uses PTDB (Neale 2002), which provides entire process configurations (semantic frames) for each verb sense and the feature assignment is simultaneous, if matched, to the entire configuration of process and its participants.

One major advantage, as compared to Honnibal's approach is that the grammar and the program are carefully disconnected so that the code is maintainable and scalable with the respect to size of the grammar.

Next chapter will introduce the structure of Systemic Functional Grammars and draw some parallels between Sydney and Cardiff schools.

Chapter 5 introduces the Dependency Grammar 1959, starting with its origins and foundations, evolution into its modern form, its applications in computational contexts particularly highlighting the Stanford grammatical model and parser. The usage of dependency grammar and dependency parse graphs is motivated in 1.7.2 as the primary input into the current parsing pipeline for creating the constituency structure. The last part of the chapter provides a set of principles and generalizations to establish cross theoretical bridge from DG towards SFG which are implemented into the Parsimonious Vole parser.

Next chapter starts with an introduction of Government and Binding Theory (GBT) explaining where the empty constituents occur in sentences. These constituents were motivated in 1.6 and are a part of solution for parsing with TRANSITIVITY system network. The second section of the chapter provides an inventory of different null elements and the last section provides, just like in the previous chapter, a cross theoretical overview, this time from GBT phrase parse structures into Stanford dependency grammar. It provides a theoretical translation of the principles from GBT into DG constituting the theoretical foundations for the technical solutions, in Section 9.3, for how to create null elements in DG and SFG graphs.

Chapter 7 provides the building blocks of the algorithms of this thesis. It makes the transition from linguistic theoretic presentations towards the computer science foundations introducing necessary typed sets, feature structures and graphs. These concepts are employed in the chapters that follow to represent linguistic constructs described in the previous chapters. An important role, in the current work, play the pattern graphs and the operations enabled by using them presented in Sections 7.3 – 7.5. The pattern graphs, as will be presented latter, constitutes a flexible and expressive method to represent systemic feature realisation rules. Also in this chapter, the system networks are defined in a simplified form corresponding to how they are currently used along with a simple strategy for choice propagation.

The first phase of the parsing pipeline (see Figure 1.8) concerning the constituency graph building is entirely covered by the Chapter 8. It presents how the input dependency graphs are first corrected, normalised and then rewritten into constituency graphs. The implementation of Parsimonious Vole also contains a full set of mapping rules between Stanford Dependency v3.5 to SFG constituency structure enumerated in Appendix .2.5.

The second phase of the pipeline (see Figure 1.8) concerning the enrichment of the constituency graph with increasingly more semantic features is described in the Chapter 9. It addresses two main system networks, that of MOOD and TRANSITIVITY

introduced in Chapter 4. The MOOD features are close to syntactic variation of text and can be addressed via graph patterns alone in the first part of the chapter. The TRANSITIVITY features are semantic in nature and require additional lexical-semantic resources from which graph patterns are generated first and then applied to enrich the constituency graph. The work presented in this chapter comprises a set of syntactically grounded graph patterns covering Mood and a few other small system networks. It provides with a clean machine-readable version of the PTDB along with a method to automatically transform PTDB records into semantically oriented Transitivity graph patterns. Also, graph patterns and alogithms have been developed to capture several principles and mechanisms for detecting null elements in texts.

Chapter 10 describes how the Parsimonious Vole parser was evaluated. This evaluation was constituted on two corpora. One was created by Ela Oren and K with the purpose of evaluating the syntactic features of this parser while the other was provided by Anke Schultz covering Cardiff Transitivity annotations. The chapter describes evaluation settings and results for syntactic and semantic parsing. Chapter 11 concludes this work by providing a thesis summary overview, indications for practical applications of this work and future directions to follow.

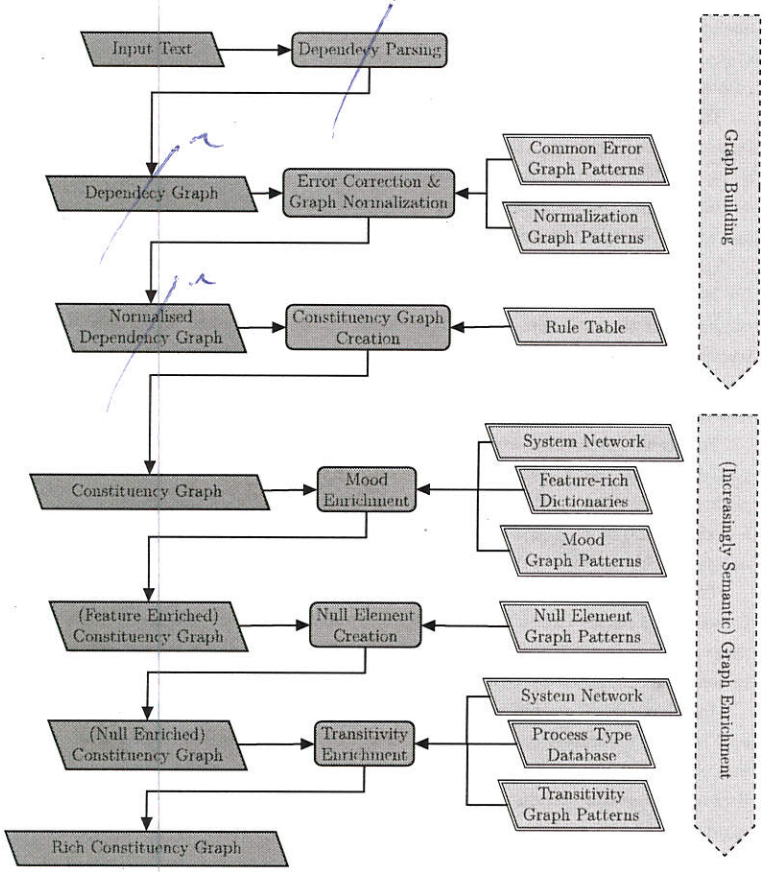


Fig. 1.8 The parsing process pipeline

graph nodes. The structural mappings are accompanied by operation specifications to perform during each traversal step. The output of this step constitutes the syntactic backbone on which the subsequent enrichment phases are performed.

Next follows the phase where each constituent node of the syntactic backbone is *enriched* with features. Some of them are syntactic in nature and others are lexical-

semantic. In between these enrichment phases there is an additional construction process adding where needed *empty constituents* that play an important role in semantic enrichment. The enrichment steps use *system networks*, *feature rich lexicons*, *graph patterns* and PTDB *semantic database* as additional resources. The *null element creation* process also needs a collection of graph patterns for identifying where and what kind of null elements occur (motivated in Section 1.6 and explained in detail in Chapter 6). The final result of the process is a *Rich Constituency Graph* of the original text comprising a substantial set of systemic feature selections associated with constituting units of structure.

The the detailed parser implementation choices and developed algorithms are presented in Chapters 8 and 9. Next section lays out the thesis structure indicating the important contributions that every chapter provides.

1.8 Thesis overview

Chapter 1 has provided an introduction to the work described in this thesis. It has indicated the are to which it seeks to contribute, and described the motivation of work from an applied and a theoretical perspective. In Chapter 2 a list of selected works on parsing with SFG is presented and briefly discussed.

Chapter 3 provides and overview of the SFL theoretical foundations. There are two outstanding traditions in SFL each providing a theory of grammar. First is developed in Sydney by Halliday. Matthiessen, Hassan, Martin, Rose and others. The second is developed in Cardiff by Fawcett. Tucker. Tench and others. I present both schools in first two sections of the chapter and then, in the third section, I provide a comparative critical discussion on both theories of grammar motivating relaxation of the rank scale approach to structure formation, unit classes and few other concepts relevant to current work.

In the next chapter I provide a description of the grammar implemented in the Parsimonious Vole which. It is a selection of unit classes from both Sydney and Cardiff Grammars following the theoretical motivation from the previous chapter. Here is also presented a selection of two system networks: MOOD and TRANSITIVITY that were selected to demonstrate how the current parsing method works. The former system network is tightly linked to the syntagmatic variations in the structure whereas latter describes ideational choices of the semantic structures and, thus, is farther from the surface variations. In order to integrate this system network I use a lexical database of verb meanings called Process Type Database Neale (2002).

Research question 3 (Compatibility of GBT and SFG). How can Government and Binding Theory be used for detecting places of null elements in the context of SFL constituency structure?

The problem of accounting for the *null elements*, mentioned above, is not addressed either in SFL or in Dependency Grammar. It is, however, addressed in detail in the Government and Binding Theory (GBT) (Chomsky 1981; Haegeman 1991), which is one of Chomsky's Transformational Grammars (Chomsky 1957a). One other goal in this thesis is to investigate, as formulated in Research question 3, to which degree GBT accounts of null elements can be reused as DG or SFG structures to undergo a cross-theoretic transformation enabling those accounts in DG or SFG contexts. Chapter 6 introduces GBT and investigate this hypothesis providing some of the cross-theoretic and inter-grammatical links to Dependency and SFL grammars that, as we will see in Chapter 9 benefits the Transitivity analysis.

1.7.2 Towards the syntagmatic account

The problem of structure construction can be outsourced as parsing with other grammars. This is done in the works of Kasper Kasper (1988) and Honnibal (2004); Honnibal & Curran (2007) who used phrase parse structures of the Chomskian style grammars. This approach is knows in SFL literature as *parsing with a syntactic backbone*. In this case, the problem changes into creating a transformation mechanism to obtain the SFL constituency structure rather than build it from scratch.

Research question 4 (Suitability of Stanford DG). How compatible are the grammatical categories and practices in the Stanford Dependency grammar with the ones in Parsimonious Volo grammar?

This thesis addresses the problem of constituency structure building by parsing the text with Stanford Dependencies parser version 3.5 (Marneffe & Manning 2008b,a; Marneffe et al. 2014) and then transforming the parse result into SFG constituency tree. The degree to which Stanford dependecies are suitable to serve as a syntactic backbone is one of the questiosn adressed in this thesis (Research question 4). An account of correspondence between linguistic primitives or configurations of primitives in the dependecy grammar to SFG primitives is provided in the end of Chapter 5 along with an analisys of Stanford dependecy grammar. The detailed description of the structure generation process is provided in the Chapter 8.

1.7.3 Towards the paradigmatic account

Once the constituency structure is in place it serves as foundation and informs the following process of feature enrichment. The configurations of units carrying grammatical categories and functions into *structural patterns* serve as "hooks" to guide the traversal of system networks in a way resembling the realization rules.

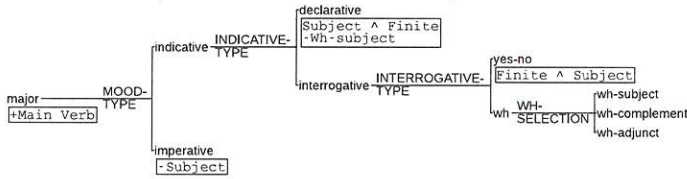


Fig. 1.7 A fragment of mood system from Halliday & Matthiessen (2013: 366)

The system network fragment in Figure 1.7 contains the realisation rules inscribed into rectangular boxes positioned below some features. These realisation rules indicate what shall be reflected in the structure when a feature is selected (discussed already in Section 1.6). The converse is also true: if structure contains a certain pattern then it is a (potential) manifestation of a given feature.

For example the structure of a *major* clause needs to have a predicate or Main Verb element realised. In the parsing process, testing whether there is a unit functioning as Main Verb below in the clause node suffices to assign the *major* feature to that clause. Next, if the clause has no unit functioning as Subject then it shall be assigned *imperative* feature, otherwise the *indicative* one. Further the INDICATIVE-TYPE system is enabled. Here the test is whether a Subject node is positioned in front of the Finite node and whether the Subject contain the preposition "who". This sort of queries on the structure can be formulated as *structure patterns* (see Section 7.3) and associated to features in the system network in the same manner the realisation rules are.

Such paterns can be identified in the instance structure and therefore the specific feature. In current parsing method, pattern recognition plays an essential role for fleshing out the constituent backbone with systemic features. The structural patterns are tested whether they *match* (see Section ??) anywhere in the constituency structure and if so then the matched nodes are enriched with the features proved in the pattern (described in Section 7.5). This process is detailed in Section 9.2.

The structure patterns in this work are expressed as *graph patterns* (described in Section 7.3). Note that I employ the concept of *graph* and not that of a tree because

missing the river boat". In the case of *worry about*$_1$ and *miss*$_1$ the first roles provided by Cardiff grammar are *compound* i.e. composed of two simple ones. In the example above, *worry about*$_1$ distributes the Phenomenon to the Complement "about missing the river boat" and the Agent-Cognizant role to an empty Subject that is said to be *non-realised*, *covert* or *null element*. A similar situation is for *miss*$_1$ that assigns an Affected-Carrier role to the empty Subject and the Possessed role to the Complement "the river boat".

Verb meaning	Semantic configuration	Participant role distribution
seem$_1$	Attributive	Carrier + Attribute
worry about$_1$	Two Role Cognition	Agent-Cognizant + Phenomenon
miss$_1$	Possessive	Affected-Carrier + Possessed (thing)

Table 1.3 Semantic role configurations according to Neale (2002); Fawcett (forthcoming)

Those unrealised Subjects in the embedded clauses are recoverable from the immediate syntactic context (no need for discourse) and correspond, in this case, to the Subject in the higher clause. This is annotated in Examples 7 and 8 therefore we can just mark the places of the null Subjects in the embedded clause in order to be able to assign the semantic labels this way ensuring that the minimal completeness constraint is fulfilled; otherwise the frame cannot be assigned to the constituents and another one shall be searched for instead. Notice also an index i to highlight that the null elements correspond to the higher clause Subject "She".

(7) *She* worried about missing the river boat.

(8) *She* missed the river boat.

(9) She$_i$ seemed [*null-Subject*$_i$ to worry [about *null-Subject*$_i$ missing the river boat]].

Now that the places of the covert constituents are explicitly marked in Example 9 and the recoverable constituents coindexed, we can distribute the semantic role configurations from Table 1.3 as provided in Table 1.4 below.

In language there are cases where constituents are empty but recoverable from the immediate vicinity by relying in most cases on syntactic means and in a few others additional lexical-semantic resources are needed. In SFL, Fawcett describes these elements in the context of Cardiff grammar (Fawcett 2008: 115,135,194) but provides no means to recover them. The Government and Binding Theory (GBT) developed in (Chomsky 1981, 1982, 1986) and based on phrase structure grammar, provides a

She$_i$	*seemed*	$\emptyset_i$	*to*	*worry*	*about*	$\emptyset_i$	*missing*	*the*	*river*	*boat.*
Attributive configuration										
Agent	Attribute									
Two role cognition configuration										
Agent & Cognizant				Phenomenon						
Possessive configuration										
Affected & Carrier							Possessed			

Table 1.4 Transitivity analysis in Cardiff grammar style (Neale 2002; Fawcett forthcoming) of Example 6

detailed account of mechanisms to detect and resolve the empty constituents. GBT explains how some constituents can *move* from one place to another, where are the places of *non-overt constituents* and what constituents do they refer to i.e. what are their *antecedents*. Such accounts of empty elements are useful in determining the correct distribution of participant roles to the clause constituents.

1.6.5 Problem summary

This section has shown the main challenges related to parsing with SFG which can be summarised as follows. First, the parsing task cannot be treated as a reversible generation task because the methods that have been shown to work for generation are not usable for parsing as such due to a high computational complexity. Second, the parsing task, regardless of the grammar, should first and foremost account for the sentence structure on the syntagmatic axis and only afterwards for the (semantic) features selected on the paradigmatic axis. Such syntagmatic account in SFL is insufficient for the parsing task. Third, syntagmatic account alone does not provide enough clues for assignment of semantic features and requires a lexical-semantic account within the grammar or as separate resource. Moreover, semantic parsing can be aided by identification of places where covert constituents are said to exist.

Next I will describe how these problems have been addressed in the current work, what are the goals of the thesis and what has been left out for the future work.

1.7 Goals and scope of the thesis

This thesis aims at a modular method for parsing unrestricted English text into a Systemic Functional constituency structure using fragments of Systemic Functional Grammar (SFG) and dependency parse trees.

The computational complexity, the lack of proper syntagmatic description in SFL and perhaps for other hidden reasons the results of parsing with SFGs so far are not

have to provide two things. First a description in terms of a formal structure of the sentence revealing the constituents plus their syntactic relations to each other. And second, a description in terms of a (complete) set of features, detailed to the extent that grammar permits, applicable to each constituent of structure.

One of the grammars successfully used in generation tasks is the Nigel grammar developed within Penman generation project (Mann 1983a). The efficiency in generation tasks is, in part, due to decomposition of language along the paradigmatic axis using functionally motivated sets of choices between functionally motivated alternatives (McDonald 1980). The Nigel grammar contains 767 grammatical systems defined over 1381 grammatical features which Bateman evaluates as "a very large computational grammar by current standards, although nowadays by no means the broadest when considered in terms of raw grammatical coverage" (Bateman 2008: 29).

The computational processes driving natural language generation relied heavily on the notion of *search*. A well defined search problem is defined in terms of a precise description of the search space which then helps a navigation process effectively to find solutions. The paradigmatic organization of the *lexicogrammar* as system networks assumed within SFL turns out to organise the search space for possible grammatical units appropriate for expressing communicative goals in generation in almost ideal manner (Bateman 2008: 28).

If, in the generation process, the abstract semantic specifications are increasingly materialised through choice making by traversing the system network towards finally generated text (see example in Section 1.5), then, in the parsing process, the reverse is the case. The process starts from a given sentence aiming to derive/search the feature choices in the system network afferent to each of the constituents. But if the paradigmatically organised lexicogrammatical resource is effective for generation it turns out, as we will see next, to be by far unsuitable for the analysis task because the size of the search space is too big to be computed in a reasonable time. Halliday himself mentions this problem when he asks *how big is a grammar?*

> Given any system network it should in principle be possible to count the number of alternatives shown to be available. In practice, it is quite difficult to calculate the number of different selection expressions that are generated by a network of any considerable complexity (Halliday 1996: 10).

The issue is that of handling a combinatorial space which emerges from the way connections and (cross-)classifications are organised in a system network. In addition to that, the orientation of systemic grammars towards choice means that a typical

grammar includes many disjunctions, which leads to the problem of search complexity. Also the abstract nature of systemic features leads to a structural richness that adds logical complexity to the task (O'Donnell 1993). So estimating the size of the grammar would in fact mean estimating the potential number of feature combinations.

For example, if we consider a hypothetical network of 40 systems then the "size of the grammar it generates lies somewhere between 41 and 2^{40} (which is somewhere around 10^{12})" (Bateman 2008: 28). Moreover, it is not easy to calculate where would the upper limit of a grammar fall even when the configuration of relations of a particular system network is known. To parse with Nigel grammar, mentioned above, would means exploring a search space of approximately 3×10^{18} feature combinations (Bateman 2008: 35). A more detailed break down the complexity by rank or primary class as provided in Table 1.1 below.

rank or primary class	*size*
adverbial-group	18
words	253
quantity-group	356
prepositional-phrase	744
adjectival-group	1045
nominal-group	$>2 \times 10^9$
clause	$>3 \times 10^{18}$

Table 1.1 Size of major components of the Nigel grammar expressed in terms of the number of selection expressions generated (Bateman 2008: 35)

For the generation task the size is not an issue because the number of choice points is actually rather small. The paradigmatic organisation is, in fact, a concise and efficient way to express the linguistic choices where the possible feature selections are relevant only when they are enabled by prior paradigmatic choices and it is only those alternatives that need to be considered (Halliday 1996: 12–13). This property of gradual exposure of choices characterises the traversal of the system networks which starts from the root and gradually advances towards more delicate features down to a leaf.

In the analysis task, the paradigmatic context of choice that helps navigation during the generation process is no longer available. It is not known any longer which features of a systemic network are relevant and which are not. This leads to a radical asymmetry between the two tasks. That is: in generation, the simple traversal of the network finds only the compatible choices because that is what the network leads to; whereas in analysis it is not evident in advance which path to follow therefore the task

constituent in Example 1 from Mood system network that is an adaptation of the Mood network proposed in Halliday & Matthiessen (2013: 162). These selections represent choices made by a natural language generation system when producing the utterance by a process similar to the one explained for the pronominal referent above. The traversal description is ommited for brevity. Organisation of the linguistic features is system networks is one of the main things that distinguishes SFL from other linguistic traditions. I will formally introduce system networks, how they are structured and how they function in Chapter 3 and 7 that follow below.

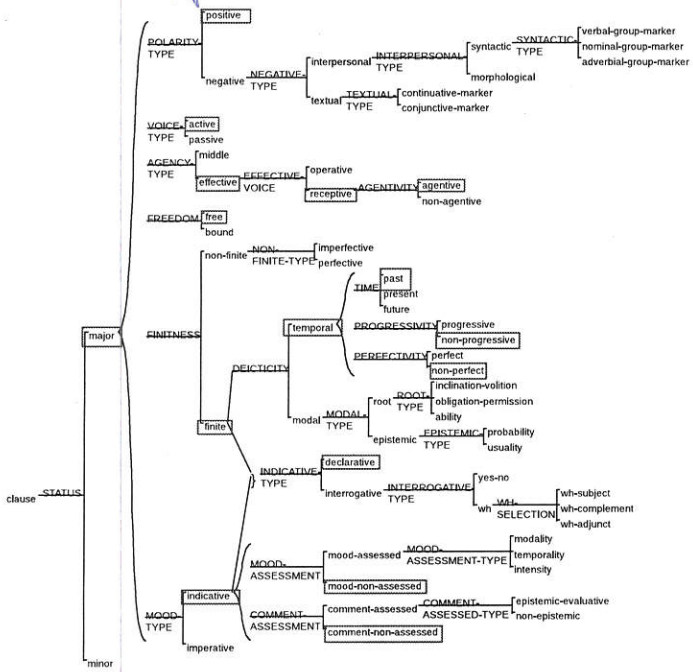


Fig. 1.5 The feature selections in the Mood system network for clause constituent in Example 1

So far we have seen constituents assigned syntactic functions such as Subject, Complement, Adjunct etc. In SFL, they are elements of the *interpersonal metafunction* which will be explained in Chapter 3. SFL provides more linguistic features and

functions depending on the kind of meaning it aims at describing. For example another view on the same clause can be provided from a perspective that in SFL is called *experiential* and corresponds to what in traditional linguistics is known as *semantics*. It is systematised, in SFL, as Transitivity which aims at providing domain independent *semantic frames* called *process configurations*. They describe semantic actions and relationships, along with *semantic roles* ascribed to their *participants*. These semantic frames generally are "governed" by verbs and more specifically each verb meaning has a dedicated semantic frame.

The clause in Example 1 corresponds to a Possessive semantic frame where "He" is the Agent and Carrier while "the cake" is the Affected and Possessed thing. Example 2 provides these annotations. These configurations and participant roles correspond to the Transitivity system network proposed by Neale (2002) which I will introduce in Chapter 4.

(2) [*Agent–Carrier* He] gave [*Affected–Possessed* the cake] away.

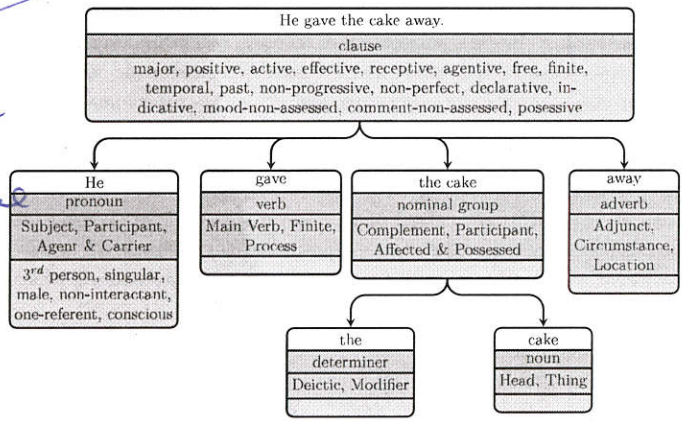


Fig. 1.6 Representation of Example 1 as feature rich constituency tree

There are more functions and features that can be assigned to the constituents in the Example 1 but this is sufficient for the current purposes of introduction. Figure 1.6 summarises everything discussed above into a partially filled constituency tree. The constituents that were not discussed are assigned only a few functions. The last (green) section of every node in the constituent tree is filled with a limited set of grammatical

the importance of the formal structural descriptions which from this perspective appear as realisation of (abstract) features.

A linguistic description is then provided at various levels of granularity, that in SFL are is called *delicacy*. Just as the resolution of a digital photo defines the clarity and the amount of detail in the picture, in the same way delicacy refers to the how fine- or coarse-grained distinctions are made in the description of the language.

There is no distinction, in SFL tradition, between lexicon and grammar. And to emphasize this fact, the term *lexico-grammar* is used which means the combination of grammar and lexis into a unitary body (see Section 3.1). A deeper description of the SFL theory of language is provided below in Chapter 3.

To present two major Systemic Functional Grammars (SFG) have been developed: the *Sydney Grammar* (Halliday & Matthiessen 2013) and the *Cardiff Grammar* (Fawcett 2008). The latter, as Fawcett himself regards it, is an extension and a simplification of the Sydney Grammar (Fawcett 2008: xviii). Each of the two grammars has advantages and shortcomings (presented in Chapter 3) which I will discuss from the perspective of theoretical soundness and suitability to the goals of the current project.

Both the Cardiff and Sydney grammars have been used as language models in natural language generation projects within the broader contexts of social interaction. Some researchers (Kasper 1988; O'Donoghue 1991; O'Donnell 1993; Souter 1996; Day 2007) consequently attempted to reuse the grammars for the purpose of syntactic parsing. I come back to these works in more detail in Section 2.

To sum up, in this thesis I adopt the Systemic Functional Linguistic (SFL) framework because of its versatility to account for the complexity and phenomenological diversity of human language providing descriptions along *multiple semiotic dimensions*, i.e. paradigmatic, syntagmatic, meta-functional, stratification and instantiation dimensions (Halliday 2003c) and at different *delicacy levels* of the *lexico-grammatical cline* (Halliday 2002; Hasan 2014). To what degree it is possible and what are the benefits of such descriptions still remains to be explored. Moreover it is still unexplored how much of the SFL descriptive potential needs to be employed in practice in order to achieve useful results or solve problems as those exemplified in Section 1.3. The concepts introduced above and other elements of the SFL theory will be addressed in Chapter 3 below. In order to provide a clearer picture on what the SFG analysis represents next section provides with an example.

1.5 A systemic functional analysis example

To provide with a better intuition on the current work, this section describes an analysis of a simple sentence in Example 1. It will guide us starting from a traditional "school grammar" concepts down to a detaile systemic functional description of the sentence. As stipulated in the previous section, SFL provides us with a variety of functions and features serving to express text meaning from several perspectives. Another source of the descriptive breadth is achieved through a practice of feature systematisation as mutually exclusive choices. The feature analysis provided here is partial and restricted to only two constituents (the clause and it's Subject) as this suffices to provide the reader with an intuition of what to expect from a full analysis.

(1) He gave the cake away.

School grammar teaches us how to perform a syntactic analysis of a sentence. So let's consider Example 1 in order to perform one. First we would assign a *part of speech* such as verb, noun, adjective etc. to each word; then we would focus on clustering words into constituents guided by the intuitive question "which words go together as a group". Following these actions we will arrive to a word clustering like the one in Figure 1.1.

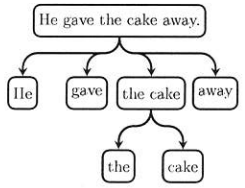


Fig. 1.1 Constituency diagram for Example 1

Figure 1.1 depicts a constituency division of Example 1.The nodes represent grammatical constituents and the edges stand for the *structure-substructure composition*. Next we can move on to assign constituent class and a grammatical function. Here the sentence is formed of a single clause which has four constituting functional parts: a subject designating who is the clause about, a predicate indicating the action performed by the subject, a complement denoting what was in scope of the action and an adjunct describing its manner. Each of these functional parts is filled correspondingly by a pronoun, a verb, a nominal group and an adverb. This analysis can be seen in Figure 1.2 as a constituency tree where the nodes carry classes and functions within parent

Such working components are already widely used to enable humans to explore and exploit large quantities of textual data for purposes that vary from the most theoretical, such as understanding how language works or the relation between form and meaning, to very pragmatic purposes such as developing systems with natural language interfaces, machine translation, document summarising, information extraction and question answering systems to name just a few. Nevertheless there is still a long way to go through before machines excel in these narrowly scoped tasks and even longer before machines start using language in the ways human do.

1.2 Living in a technologically ubiquitous world

The human language has become a versatile highly nuanced form of communication that carries a wealth of meaning which by far transcends the words alone. When it comes to *human-machine* interaction this highly articulated communication form is deemed impractical. So far humans had to learn to interact with computers and do it in a formal, strict and rigorous manner via graphical user interfaces, command line terminals and programming languages. Advancements in *Natural Language Processing* (NLP) are a game changer in this domain. NLP starts to unlocks the information locked in the human speech and make it available for processing to computers. NLP becomes an important technology in bridging the gap between natural data and digital structured data.

In a world such as ours, where technology is ubiquitous and pervasive in almost all aspects of life, NLP becomes of great value and importance regardless of whether it materializes as a spell-checker, an intuitive recommender system, spam filters, (not so) clever machine translators, voice controlled cars, or intelligent assistants such as Siri, Alexa or Google Now.

Every time an assistant such Siri or Alexa is asked for directions to the nearest Peruvian restaurant, how to cook Romanian beef stew or what is the dictionary definition for the word "germane", a complex chain of operations is activated that allows 'her' to understand the question, search for the information you are looking for and respond in a human understandable language. Such tasks are possible only in the past few years thanks to advances in NLP. Until now we have been interacting with computers in a language they understand rather than us. The next challenge is to develop a technology that enables computers to interact with us in a language we understand rather then they.

1.3 NLP for business

NLP opens new and quite dramatic horizons for businesses. Navigating with limited resources stormy markets of competitors, customers and regulators and finding an optimal answer/action to a business question is not a trivial task. In this section I present a few example application areas and use them to discuss tasks that need to be accomplished for NLP in such contexts. These examples underline the ever growing need for NLP putting into perspective the need of ever deeper and richer linguistic analysis across a broad range of domains and applications.

Markets are influenced by the information exchange and being able to process massive amounts of text and extract meaning can help asses the status of an industry and play an essential role in crafting a strategy or a tactical action. Relevant NLP tasks for gathering market intelligence are *named entity recognition* (NER), *event extraction* and *sentence classification*. With these tasks alone one can build a database about companies, people, governments, places, events together with positive or negative statements about them and run versatile analytics to audit the state of affairs.

Compliance with governmental, European or international regulations is a big issue for large corporations. One question for addressing this problem is whether a product is a liability or not and if yes then in which way. Pharma companies for example, once a drug has been released for clinical trials, need to process the unstructured clinical narratives or patient's reports about their health and gather information on the side effects. The NLP tasks needed for this applications are primarily *NER* to extract names of drugs, patients and pharma companies and *relation detection* used to identify the context in which the side effect is mentioned. NER task help transforming a sentence such as "Valium makes me sleepy" to "(drug) makes me (symptom)" and relation detection will apply patterns such as "I felt (symptom) after taking (drug)" to detect the presence of side effects.

Many customers, before buying a product, check online reviews about the company and the product regardless of whether it is pizza or a smartphone. Popular sources for such inquiry are blogs, forums, reviews, social media, reports, news, company websites, etc. All of these contain a plethora of precious information that stays trapped in unstructured human generated text. This information if unlocked can play a great deal in company's reputation management and decisions for necessary actions to improve it. The NLP tasks sufficient to address this business required are *sentiment analysis* to identify attitude, judgement, emotions and intent of the speaker, and *co-reference resolution*, which connects mentions of things to their pronominal reference in the following or preceding text. These tasks alone can extract the positive and negative