2. Linguistic perspective on text and lexis

2.1 Natural language text

2.1.1 Text organisation

The definition of text and the nature of its organisation has been a matter of considerable debates in linguistics. Linguists disagree not only about the factors that determine internal organisation of text, but also what kind of language activity should be considered text and which - non-text. Text analysis has become a focus of some areas of linguistic study in the recent decades. As a linguistic discipline it is still at a stage of active research where controversies about key objects of research are not uncommon. It is not the aim of this chapter to give a complete coverage of all research activities in the area of text linguistics; instead the most influential approaches to the nature of text and its organisation will be introduced, some of which determined the view of the notion of text adopted in this work.

Lyons gave examples of two polar approaches to the nature of text [Lyons77]. The first is the view of the text as a "product of more or less conscious and controlled literary composition" [Lyons77, vol.2, p.631]. The second approach is shared by linguists who adopted a broader understanding that the text can be the product of everyday language-behaviour including phonologically transcribed conversations [Lyons77, vol.2]. However neither of these approaches to defining the notion of text clearly indicate the criteria for distinguishing text from non-text. As Halliday and Hasan pointed out, native speakers intuitively know whether any product of language activity is a text or not. This means there should be some objective characteristics of text that determine our perception of it as a text and not as a sequence of unrelated sentences.

One possible way of defining a text is to attribute to it the idea of a *whole*. The view of a text as a whole gained substantial recognition among linguists working within different frameworks of linguistic study: text grammars, discourse analysis, semantics. Halliday and Hasan emphasised this idea, defining the text as "any passage spoken or written, of whatever length, that does form a unified whole" [Halliday76]. They recognise however that distinction between a text as a whole and a sequence of disconnected sentences is a matter of degree, which depends on the stylistic factors such as literary genre and the author's language ability and competence. Nevertheless they point out that if a piece of written or spoken language activity is to be considered a text, it must have some internal features that differentiate it from a collection of unrelated sentences.

Although many linguists agree that the pivotal characteristic of text is its being a whole unity, their opinions diverge considerably as to what exactly holds a text together. The differences are rooted in the overall approach to text: it is viewed either as a *grammatical entity*, or as a *semantic unit*.

Linguists who look at a text as a grammatical entity come from the structuralist background and study text within the framework of the so-called “text grammar”.

Most text grammarians use some aspects of the sentence grammar in the analysis of the text. They claim that text has a rigid structure comparable to that of the sentence. Influential work in this area of text study belongs to Van Dijk [Dijk77]. He claims that the sentences are interrelated structurally in text just as clauses are interrelated within the sentence, and that both sentences and texts can be similarly described at some level using the same methods. However the global structure of a text, according to Van Dijk, also requires another level of description, which must be formulated as a systematic set of rules to derive the semantic structure of a text from the semantic structures of sentences. He introduces the notion of *macro-structures* to describe the semantic representation of a text or its part. Macro-structures are a formal equivalent to the intuitive notion of a topic of discourse. Van Dijk assumes that a topic of discourse, which denotes what a text or a part of it is about, can be described as a complex logical proposition which is derived from a set of propositions contained in the sequences of sentences. The semantic structure of a text can be described using several levels of macro-structures, starting from the simple macro-structure of a single sentence to the macro-structures of paragraphs and the entire text. [Dijk77]

Such a formalist approach to a text has been criticised by some linguists [Halliday76, Brown83]. Brown and Yule argue that an attempt to formulate general rules for deriving the semantic structure of a text objectively is an illusion. What Van Dijk attempts is to produce a sentence representing the underlying topic of the text and transform it into a set of logical propositions which would have a relation of entailment with the set of propositions derived from the sentences in the text. Brown and Yule argue that this approach is a way of determining just one of the possible topics in the text and cannot claim objectivity [Brown83].

Some linguists do not recognise a text as an object of grammatical study, claiming that the sentence is the highest grammatical unit [Halliday76, Lyons77]. Halliday and Hasan argue that a text is not a structural unit. It cannot be said that it is similar in kind to a sentence, and it does not consist of sentences in the same way as sentences consist of clauses. While it is possible to specify a limited number of structural relations within a sentence, it is not possible to draw up a list of structural relations within a text, or to categorise sentences into classes which perform certain roles in text. They doubt that sentences can enter into generalised structural relationships to realise some function in the text as a unity [Halliday76].

Many opponents of the structural approach to text suggest that a text, rather than having a structure, has some organisation [Halliday76, Hoey91]. Instead of considering text as a grammatical unit that permits structural description, they view it as a semantic unit that has some patterns of organisation. An interesting distinction between a structural and organisational approaches to text was drawn by Hoey. He states that a structural description is "one that permits one to make predictive statements about the data under examination" [Hoey91, p.13]. Structural descriptions predict what is possible or impossible to occur, whereas patterns of organisation depict what has occurred. Therefore if a structural stance is adopted, the description must account for which combinations are possible and under what circumstances. Description as patterns of organisation, on the other hand, does not suggest the impossibility, but the probability of certain combinations. Hoey emphasises that " text analysts can describe what is common, the culturally popular patterns of organisation,

but they are not convincing when they seek to make predictive statements about what can and cannot occur" [Hoey91, p.204].

Halliday and Hasan describe text as a unity brought together not by a structure but by the organisation they term texture [Halliday76]. They introduce the notion of "texture" to emphasise the fact that a text is a whole unit, but the factors that contribute to its unity are not of structural nature. The property of the text that gives it texture is *cohesion*. Texture is formed by the relationships that bind together items in text, which they call cohesive ties. The concept of cohesion will be described in greater detail in section *2.1.3 Cohesion in text*.

The following section will give an overview of the elements of text as seen from two different areas of linguistic study: grammar and semantics.

2.1.2 Text elements

Like any unit of language in use, text exists in two indivisible planes: *the plane of contents* and *the plane of expression*. The linguistic disciplines that study means of language expression are phonology and grammar, while the discipline which concerns itself with the study of meanings is semantics.

Linguistic elements can be described as organised hierarchically: the smaller units are combined to form larger units of language, which are different in kind (figure 2.1).

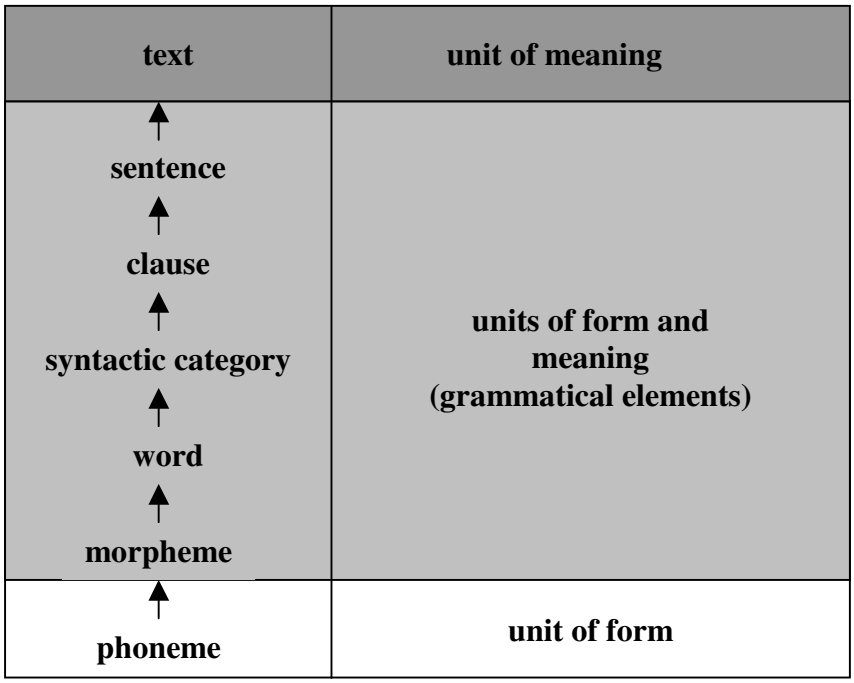


Figure 2.1. Stratal organisation of language elements

The smallest unit of the spoken language is the *allophone*, which is a sound variant of a structural element of *phoneme*. The linguistic discipline which studies the sound system of language is phonology. The phoneme, however, does not carry any meaning in itself, but serves only as a building block for higher ranked elements.

The next element in the hierarchy is a *morpheme*, which is a component of a word structure. The essential feature of a morpheme as contrasted to a phoneme is that the morpheme is a unit of meaning. It is defined as "the smallest unit of language that carries information about meaning or function" [O'Grady97, p.133]. Morpheme is the smallest grammatical element and is an object of the branch of grammatical study - morphology. Morphology is commonly defined as "the system of categories and rules involved in word formation and interpretation" [O'Grady97, p.132]. A *word* - the next element in the hierarchy of linguistic elements - can consist of one or more morphemes. Complex words normally comprise a root - a morpheme that carries the major part of the word's meaning and one or more affixes - morphemes carrying grammatical meaning. For example word *'walked'* consists of a root morpheme *'walk'* and of an affix *'ed'* which conveys the grammatical meaning of past. The word as a primary object of lexical study will be covered in a greater detail in the subsequent part *2.2 Lexis*.

The highest grammatical element in a text is a *sentence*. The branch of grammar that studies the sentence structure is syntax. Syntax can be formally defined as "the system of categories and rules that underlies sentence formation" [O'Grady97, p.181]. A sentence, however, does not simply consist of sequences of words. There are several levels of structural components that are combined together to form sentences. The smallest syntactic unit is called a *syntactic category*, which is defined as "a group of words or sequences of words in a given language that can replace one another in any sentence of the language whatsoever without affecting grammaticality" [Culicover76, p.13]. A syntactic category can either contain single words, in which case it is known as a lexical category, or it can contain sequences of words: a phrase category [Culicover76]. The most studied lexical categories are noun, verb, adjective and preposition. Phrase categories are based on lexical categories, and are structured as head + specifiers, where the head is the lexical category which the phrase is built on [O'Grady97]. Thus, if the head of a particular phrase is represented by a lexical category of noun, the category the phrase belongs to is a noun phrase.

Syntactic categories are combined into *clauses*. A clause is centred around a predicate - a verb or a verb phrase, and can also have other syntactic categories functionally related to the predicate, such as subject and complement. A sentence can consist either of only one clause - a *simple sentence*, or of more than one clause – *coordinate* and *complex sentences*. In coordinate sentences the clauses have a coordinate relationship, whereas in complex sentences one clause is grammatically subordinate to another.

The linguistic elements presented above function each at their own level to build the plane of expression of text, which serves in its entirety to realise the underlying contents of text. The discipline devoted to the study of meaning is semantics. Semantics as a study of meaning expressed by any sign system is not confined to linguistics, however language, being a major sign system is a dominant object of semantic study. Language semantics concerns itself with the meanings expressed by

language elements on different levels, such as word meanings and sentential meanings. Word meaning will be an object of a separate discussion in the later section of this work devoted to lexis (*2.2.1 Lexical unit and lexical meaning*).

Textual content is not a direct combination of the meanings conveyed by individual sentences. As described in the previous section, text is a more complex semantic unit than a simple combination of sentences. Therefore the meaning on the sentential level will only be touched upon briefly in this section, with the main focus on the larger semantic units existing in text.

Most discussions about the textual content centre around the intuitive notion of *topic* or 'aboutness' of a text or its parts. The notion of topic has proved difficult to define in linguistic study. Primarily, it is important to differentiate sentence topic and discourse topic. The topic, as it is understood in sentence grammar, is "its centre of attention" [Finegan94, p.200], in contrast to the notion of "comment", which is new information introduced in the sentence about the topic. The definition of the notion of discourse topic is a more difficult task than the definition of a sentential topic. Some linguists suggest that a topic is a proposition. Among such approaches is the rather formalist definition of a topic as a complex logical proposition suggested by Van Dijk [Dijk77], which was discussed in the previous section of this work. The proposition-based approach implies the idea that there exists a phrase or a sentence that is a single correct representation of the topic of the whole text or its part. Brown and Yule criticize this approach, arguing that for any text there can be several ways to represent the topic, each of which will reflect a particular judgement about the text contents [Brown83]. To determine the set of correct judgements about the discourse topic Brown and Yule suggest that extra-textual factors need to be taken into account. They describe the factors that are external to the text but necessary for the correct interpretation of its contents as the "activated topic framework" [Brown83].

The linguists whose approaches to topic have been discussed above try to define this notion from the point of view of its contents. Others, in contrast, try to derive its definition by formulating methods of locating boundaries between topics in text. Such approaches are based on the assumption that a shift from one topic to another is marked in some way in the text. The most straightforward way is to look for the boundary between topics in the orthographic boundaries between paragraphs. However, as some linguists point out, orthographic paragraphs often have stylistic rather than semantic functions [Longacre, Hinds cited in Brown83, p.95]. Therefore it is possible that the beginning of a new paragraph coincides with the start of a new topic, but this is not necessarily so [Brown83]. Another approach is to try to derive the information about topic boundaries from some other information which can be elicited from text. The information in question is various means of expression of cohesion in text, in particular its type - lexical cohesion, discussed in detail in section *2.3 Lexical cohesion*. Some implications of the analysis of lexical chains and bonds for the discovery of topic structure in text will be touched upon in sections *2.3.2* and *2.3.3*.

2.1.3 Cohesion in text

Following linguists who adopted an organisational approach to text [Halliday76, Hoey91], it is assumed in this thesis that text is a non-structural unity, whose entirety is dependent not on grammatical rules, but on non-structural relations existing between its elements. Halliday and Hasan introduced the concept of 'textual' or 'text-forming' component of the linguistic system, which they define as a "set of resources in a language whose semantic function is that of expressing relationship to the environment" [Halliday76, p.299]. They claim that it is the meaning realised through text-forming component of the language that creates text, and distinguishes it from the unconnected sequences of sentences. Cohesion constitutes the major unit of text-forming resources of the language, and its role is to express continuity throughout the parts of text. As Halliday and Hasan point out that the continuity created by cohesion consists in "expressing at each stage in the discourse the points of contact with what has gone before." [Halliday76, p.299]

It is important to emphasise that cohesion is a semantic concept which exists in a set of semantic relations between elements in text called cohesive relations or 'ties' in Halliday and Hasan's terms. These cohesive relations are realised through a set of devices, which will be covered later in this section, but the means of realisation of the cohesive relations are only of secondary importance. The meaning of the cohesive relation does not depend on the means of its realisation [Halliday76].

A cohesive relation occurs between two items in text when for the complete interpretation of one item the presence of the other is required. These two cohesively related items can occur, either within a sentence, or in different sentences throughout the text. In the former case, cohesion is realised in the syntactical structure of the sentence; in the latter it is realised through non-structural text-forming devices. A sentence, being subject to strict syntactic rules of structuring, imposes restrictions on how a cohesive relation can be expressed. For example there are rules of pronominalisation which require the use of a pronoun to name the entity mentioned for the second time in the same sentence [Halliday76]. Cohesion, however, is not divided into two distinct types: sentential and textual; it is first of all a semantic and not purely formal relation between two items, whose position in text is not governed by grammatical structure. "Cohesion is a general text-forming relation or a set of such relations, certain of which when incorporated within a sentence are subjected to certain restrictions." [Halliday76, p.9]

Cohesive relations are realised in the plane of expression of text through grammatical and lexical devices. Halliday and Hasan distinguished five categories of cohesive relations according to the means of their realisation in text [Halliday76]:

- Reference
- Substitution
- Ellipsis
- Conjunction
- Lexical cohesion

The first four categories are realised through grammatical structures, whereas the fifth category is realised through lexis.

Reference occurs when certain items do not refer to a concept directly, but refer to another item in text which in turn refers to that concept. In English such reference items are pronouns and determiners. Reference is a semantic relation, i.e. a relation between the meanings of the two items in the context and not between the linguistic units.

Substitution is a grammatical relation whereby certain words stand for other lexical units which took place earlier in text. The examples of such words are 'one' standing for a noun and 'do' standing for a verb which occurred earlier in text. The cohesive relation which takes place in substitution is the relation between two lexical items, as distinct from reference - relation between meanings.

Ellipsis can be defined as "substitution by zero" [Halliday76, p.143]. Substitution and ellipsis are the same process, only in case of ellipsis what stands for the item mentioned earlier in text is nothing.

Conjunction consists in using adjunct-like elements like 'although', 'however', 'nevertheless'. Conjunction is different from the above three types of cohesive relations, as conjunctive elements are not cohesive themselves in the sense that they are not directly related to any preceding item or its meaning. The role of conjunctive elements is to express certain meanings which presuppose other semantic components in text before them [Halliday76].

Lexical cohesion is achieved through semantic connectedness between lexical items in text. It is principally different from the above four types in the fact that cohesive relations falling into this category are in the first instance lexical relations and only secondarily textual. But as Hoey points the relation between lexical and textual is bi-directional, i.e. "the text provides the context for the creation and interpretation of lexical relations, just as the lexical relations help create the texture of the text" [Hoey91, p.8].

Lexical cohesion will be the subject of a more detailed discussion in section *2.3*. Since lexical cohesion is a relationship between words as lexical units and not as grammatical items in their functional roles in text, it is first necessary to illuminate another stratum of language organisation – *lexis*, which will be the focus of the discussion in the next section.

2.2 Lexis

2.2.1 Lexical unit and lexical meaning.

The question of what object in language should be defined as a lexical unit is controversial in linguistics. It is outside the scope of this thesis to give a comprehensive overview of all approaches to this problem. Therefore, first, the factors contributing to the complexity of the task of defining a lexical unit will be outlined and, secondly, the perception adopted in this thesis will be described.

Attributing the role of a lexical unit to a pre-theoretical notion of a 'word' as a sequence of letters delimited by spaces in text, will leave many questions unanswered.

Language in use is characterised by a complex interplay between grammatical and lexical factors. A word in text can incorporate both grammatical and lexical properties. Each word functioning as a grammatical element in a sentence has certain grammatical meaning or a set of meanings, but not every word has lexical meaning, i.e. referring to a concept. It is widely accepted to classify words into content words and function words. Function words like prepositions and articles have only grammatical meaning, which can be explained only in relation to other elements in the sentence. Content words, on the other hand, also have a lexical meaning, which is based on the reference to a concept. Function words are, therefore, the object of grammatical study, whereas content words are the object of both grammatical and lexical study. Each occurrence of a content word has a combination of a lexical and grammatical meaning, for example word *travelled* has a lexical meaning referring to the concept of 'travelling' and a grammatical meaning of past. *Travelled, travels, travelling* are not considered as different words, but different word forms of the same lexeme *to travel*. The lexeme of a word is a unit of lexis devoid of grammatical characteristics, representing all grammatical forms of this word in use. And it is a lexeme as a unit of a language structure of lexis that will be considered as a lexical unit in this thesis.

Another controversial question is whether to consider multiword entities like compound words or compound terms as lexical units. So far there is no reliable method for isolating such entities in text and distinguishing them from combinations of two or more lexical units. Compounding is recognised as a word forming mechanism in a language, whereby two or more lexemes are put together to form a new lexeme. While there is little doubt that words like *highranking* constitute a single lexical unit, items such as *search engine*, composed by the same word forming mechanism could be viewed as two lexical units or one. The essential characteristic of a compound word is that as a whole it denotes a new concept, different from the concepts denoted by any of its compounds. In contrast, a noun phrase like *powerful engine* does not refer to a new concept. Therefore, a compound word which as a whole denotes a new concept can be considered as a lexical unit, i.e. part of the lexical structure of language.

The definition of a lexical meaning is no less controversial than that of a lexical unit. The point of view that the lexical meaning of a word is a concept referred to by it is common in linguistics. Saussure's idea [Saussure16] that a word as a linguistic sign is a two-sided entity consisting of a sound-form and a concept referred to by it has received a wide acclaim in structural linguistics. In this work a related but somewhat different approach to defining a concept and a lexical meaning is adopted. A concept representing our knowledge about a real-world object is viewed as having an almost infinitely large number of attributes characterising this object from all facets: its characteristics, relations with other objects, etc. Lexical meaning is a subset of these attributes, including only the most characteristic attributes of the concept, necessary for differentiating it from other concepts [Apresjan74, Levkovskaja62, Novikov83]. This is evident in the existence of synonyms particularly in terminologies, whereby although all members of the synonym group are related to the same concept, they have different lexical meanings, because each of them highlights a particular aspect of this concept.

Another important point about the nature of lexical meaning, postulated in structural linguistics [Saussure16], is that lexical units do not exist in isolation in language. They are part of the systems of lexical units existing in two dimensions - paradigmatic and syntagmatic (see the next section). Therefore the complete meaning of the lexical unit can be derived only by considering the relationships it has with other lexical units in these two spheres of the language.

To summarise, the lexical meaning is established, on the one hand through its relatedness to extra-linguistic entities - concepts, and on the other hand through its relations with other lexical meanings in both the lexical structure of the language (the paradigmatic sphere) and in the language in use (the syntagmatic sphere).

Before moving on to the discussion of syntagmatic and paradigmatic relations it is necessary to clarify two concepts: *semantic relations and lexical relations*, which will be used further in this work. Semantic relations are relations between concepts and are extra-linguistic in their nature. Lexical relations exist between lexical meanings of the lexical units. [Evens88] They are however more complex than relations between concepts due to the fact that on the one hand they reflect the relations between the concepts underlying lexical meanings (e.g. *hyponymy*), and on the other hand, being part of the language, they include purely linguistic relations between lexical meanings (e.g. *synonymy*). Because lexical relations between words encompass not only the linguistic but also the conceptual sphere it is not uncommon to refer to them as *lexical-semantic relations* or even *semantic relations* to emphasise the relations between underlying concepts. In this work all three terms will be used to refer to relations between lexical meanings of words, the choice of the term depending on the desired emphasis: either on the purely linguistic relations or on the underlying conceptual relations. Where the distinction is unimportant the term *lexical-semantic relations* will be used.

2.2.2 Syntagmatic and paradigmatic relations

Saussure, the founder of modern structural linguistics, formulated the idea that any linguistic element exists and derives its value through relationships with other linguistic elements [Saussure16]. This became the central thesis of Saussurean structuralism. Linguistic elements at any level of analysis are interrelated in two dimensions - in the language as a system and the language in use (*'langue'* and *'parole'* in Saussurean terms or *'language-system'* and *'language-behaviour'* in Lyon's terms [Lyons77]). Relationships in the language-system are known as paradigmatic and in the language-behaviour as syntagmatic. Robins summarised the differences between the two types of relations as follows:

> "Syntagmatic relations are those holding between elements forming serial structures, or 'strings' (...) at a given level ... " [Robins89, p.47]
>
> "Paradigmatic relations are those holding between comparable elements at particular places in structures ..." [Robins89, p.47]

Syntagmatic relations between words as lexical units are relations between the lexical meanings of words co-occurring in the linearity of text. Saussure stressed that

syntagmatic relations exist between words consecutively occurring in text: "words acquire relations based on the linear nature of language because they are chained together" [Saussure16, p.123]. From this it follows that those relationships between words that realize lexical cohesion in text - repetition and collocation (see section *2.3*) - are of syntagmatic type.

Paradigmatic relations exist between lexical units as part of the lexical structure of the language-system. The lexical structure of the language can be viewed as an abstraction: the mental organisation of the lexicon, on the one hand mirroring the mental organisation of concepts, and on the other hand holding relations specific to the language itself.

Relations commonly recognised as holding within the lexical structure of the language are:

- Antonymy;
- Synonymy;
- Hyponymy;
- Meronymy (part-whole relationship).

Antonymy is a name for a wide range of relations, all of which are broadly characterised by the idea of 'oppositeness' of meaning between two lexical units.

Synonymy is a relationship between two or more lexical units which can be characterised as having 'sameness of meaning'. It is doubtful, however, that language keeps true synonyms, i.e. words with exactly the same lexical meanings. Usually synonymy in language is not absolute but relative, i.e. words have certain degree of similarity in meaning, but they have some differences - stylistic, emotive, collocational or dialectal [Palmer81].

Hyponymy is a hierarchical relation between more specific lexical units - hyponyms and more general units - hypernyms. There is a widely though not universally accepted point of view that lexicon of the language is organised hierarchically and can be represented as tree-diagrams [Lyons77]. The lexical meaning of a word in many cases can be specified by relating this word to its hypernym and contrasting it to its co-hyponyms.

Meronymy or part-whole relation is the second type of hierarchical relation. Though the part-whole relation does not organise the lexical structure like the hyponymy relation, it is important in the sense that lexical meanings of some words are specified through their relation to other words by the part-whole relation.

2.3 Lexical cohesion

Each lexical item used in a text acquires its own textual history, i.e. the lexical environment or context in which it is interpreted. By the time a word is encountered in text, the context for its instantial interpretation has already been built up. At the same time the fact that the previous lexical items which constitute the context of this word are necessary for the realisation of its 'instantial meaning' shows that it forms relationship of lexical cohesion with them [Halliday76]. It can be said that there is a

bi-directional relationship between the word in question and preceding lexical items which constitute its context: on the one hand the preceding lexical items provide a background for the correct interpretation of this word in the current stretch of text, and on the other hand the word itself through its cohesion with the earlier items functions as an element which contributes to the continuity of text.

The main feature distinguishing lexical cohesion from grammatical types of cohesion is the fact that lexical items that cohere in text are independent of each other in the realisation of their lexical meaning, though they rely on their common environment for the realisation of their *'meaning potential'* [Hoey91, p.71], or *'instantial meaning'* [Halliday76, p.289]. Grammatical items in grammatical cohesion, in contrast, not having a lexical meaning of their own, rely entirely on their lexical counterparts for the realisation of their contextual meaning.

It is notable that, in Halliday and Hasan's analysis of various types of cohesion in texts of various genres, lexical cohesion accounts for more than 40% of all cohesive relations [Halliday76].

The third stage of our experiments (chapter 8) explores the concept of lexical cohesion between query terms, realised through the similarity of their collocates. In particular, it studies the relation between documents' lexical cohesion and their property of relevance. It was, therefore, considered important to give in the following subsections a more detailed discussion of lexical cohesion and linguistic mechanisms for its realisation in text.

2.3.1 Types of lexical cohesion

Halliday and Hasan in their earlier work [Halliday76] distinguished two broad categories of lexical cohesion: *reiteration* and *collocation*.

Reiteration, as Halliday and Hasan understand it, refers to a broad range of relations between a lexical item and another word occurring before it in text, where the second lexical item can be an exact repetition of the first, a general word, its synonym or near-synonym or its superordinate. Lexical reiteration is different from the grammatical cohesion type - reference by several major factors. First of all the second word in this relation is a member of an open set, i.e. it is a lexical and not a grammatical unit. Secondly, it can not only refer to the previous item, but be related to it through a lexical-semantic relation, for example synonymy. And thirdly, it does not need to refer to the same referent as the first word. As Halliday and Hasan point out the idea of reference is irrelevant to lexical cohesion, since it is not a referential relation. The second word can be identical, inclusive, exclusive or unrelated to the referent expressed by the first word [Halliday76, p.283].

Another broad category that Halliday and Hasan distinguish in lexical cohesion - collocation - is defined rather vaguely in their earlier work [Halliday76]. Collocation will be discussed in section *2.4* of this thesis, however at this point it is necessary to introduce the notion by giving the earliest established understanding of collocation in linguistics formulated by Firth [Firth58]: collocation is a relationship between items which co-occur in text with more than random probability. Although Halliday and

Hasan admit that collocation is a relationship between lexical items that occur in the same environment, they, nevertheless, define this category rather generally, namely as lexical cohesion which is not reiteration [Halliday76, p.287].

Thus, the suggested categorisation of lexical cohesion in Halliday and Hasan's earlier work is as follows (taken from [Halliday76, p.288]):

Type of lexical cohesion:

1. Reiteration
    (a) same word
    (b) synonym (or near synonym)
    (c) superordinate
    (d) general word
2. Collocation

In her later work Hasan [Hasan84] admitted the intuitive character of their earlier category of collocation and left it out from her later categorisation of lexical cohesion altogether, introducing in its place a set of narrowly-defined lexical relations. The following categorisation is taken from [Hasan84, p.202]:

A	*General*	i. repetition	leave, leaving, left
		ii. synonymy	leave, depart
		iii. antonymy	leave arrive
		iv. hyponymy	travel, leave (including co-hyponyms, leave, arrive)
		v. meronymy	hand, finger (including co-meronyms, finger, thumb)
B	*Instantial*	i. equivalence	the *sailor* was their *daddy*; *you* be the *patient*, *i*'ll be the *doctor*
		ii. naming	the dog was called *Toto;* they named the *dog Fluffy*
		iii. semblance	the *deck* was like a *pool*; all my pleasures are like *yesterdays*

Hoey [Hoey91] in his categorisation of lexical cohesion omitted the category of collocation. Instead he categorised a rather broad range of relationships under the heading of *repetition*. He distinguished the following sub-categories:

1. Simple lexical repetition
2. Complex lexical repetition
3. Simple partial paraphrase
4. Simple mutual paraphrase
5. Complex paraphrase
6. Superordinate, hyponymic and co-reference repetition

By simple lexical repetition Hoey understands a relationship between two words in text which are the same lexical units, belonging to the same lexical category, but possibly having different grammatical characteristics, such as tense or number. Hoey points out that the seemingly uncomplicated nature of this relationship is deceptive. The major problem in identifying this relation lies in the possible polysemy of a word.

If only form is taken into account, then any two words with identical forms will be treated as an instance of repetition relationship. First of all Hoey points out that repetition does not presuppose identical meaning of its components. Their meanings can be related to different degrees. The major requirement, however, is that two words must have a similar collocational environment. If two occurrences of a word in a text have similar meanings, their collocational environment will be similar as well [Hoey91]. For more detailed discussion of collocation see section *2.4*.

Complex repetition occurs when two words have the same root morpheme, but their other word-forming morphemes are different (e.g., write, writing; visible, invisible), or when they are identical formally, but belong to different lexical categories, e.g., animal *(noun)*, animal *(adjective)*.

Both simple and complex repetitions cover the cases when two words have identical root morphemes. Words which enter paraphrase relations are different lexical units. The main criteria for defining a relation as a paraphrase is that one word could be used as a substitute for another word and its meaning must have the same level of generality. It could be compared to Hasan's category of synonymy, but, as Hoey points out, he relies on the context in identifying the relation as a simple paraphrase, whereas Hasan's synonymy is a lexical relation, i.e. that of the lexical structure of the language, rather than a textual relation which is interpreted in the context [Hoey91]. Hoey also includes into this category antonyms which have different morphological form (e.g., wet, dry), again stressing the point that this is primarily a textual relation, realised in context, and not lexical in the first place.

Hoey further distinguishes the cases of simple paraphrase as partial and mutual. A paraphrase is partial if only one of the participants of the relation can be substituted by the other, it is mutual when either of them can be used in place of each other [Hoey91].

Complex paraphrase is a relationship requiring three lexical items. It occurs when one item is a simple paraphrase of another item and a complex repetition of the third. According to Hoey the relationship between the second and the third items is a complex paraphrase [Hoey91, p.64]. For example, *publication* and *article* are in a relation of simple paraphrase and *publication* and *publish* are in a relation of complex repetition, hence there is a relation of complex paraphrase between *publish* and *article*.

Recognising superordinate, hyponymic and co-reference relationships as types of repetition, Hoey narrowed down their definition to only those cases where the identity of reference is retained. This is obvious in the case of co-reference, where any two words which point to the same referent are considered to be co-referential. In the cases of superordinate and hyponymical relation, Hoey argued that if identity of reference is not taken into account then, any words denoting a physical inanimate object, for example, must be considered repetition cases of an earlier occurrence of such general word as *thing* [Hoey91, p.69].

2.3.2 Lexical links and lexical chains

A single instance of a cohesive relation between two items was termed a *'tie'* by Halliday and Hasan. Since Halliday and Hasan were studying all types of cohesion, the term 'tie' refers to both grammatical and lexical cohesive relations. Hoey used the term *'link'* to denote an instance of repetition in the broad sense he defined in [Hoey91]. The difference between Hoey's definition of a 'link' and Halliday and Hasan's definition of a 'tie' is twofold: first, 'tie' is a more general notion than 'link', but more importantly, 'tie' implies directionality in a relation. The term 'link' has been used more extensively than 'tie' in other research to denote a lexical cohesive relation [Morris91, Hirst97, Ellman2000]. Since the study of grammatical types of relations is beyond the scope of this project, the term 'link' will be used throughout this work to refer to instances of lexical cohesion.

Lexical cohesion in text is normally realised through sequences of linked words - lexical chains. The term *'chain'* was first introduced by Halliday and Hasan to denote a relation where an element refers to an earlier element, which in turn refers to an even earlier element and so on. The notion of 'chain' in Halliday and Hasan's sense, again like 'tie', includes all types of cohesive relations. To distinguish the notion of a chain composed from lexical cohesive relations the term *'lexical chain'* will be used in this thesis.

Morris and Hirst [Morris91] define lexical chains as sequences of related words, which have distance relations between them. One of the prerequisites for the linked words to be considered units of a chain is that they should co-occur within a certain span.

While Hoey suggested using only information derivable from text to locate links in text, Morris and Hirst used Roget's thesaurus in identifying lexical chains. According to their algorithm two words are connected in a lexical chain if they have one of the following [Morris91]:

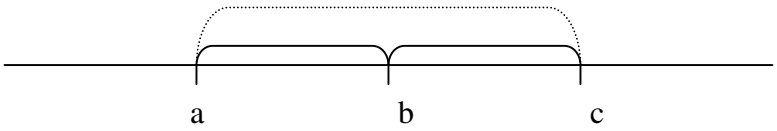
1. Their index entries point to the same thesaurus category or to adjacent categories;
2. The index entry of one points to a thesaurus category that contains the other;
3. The index entry of one contains the other;
4. The index entry of one points to a thesaurus category that in turn contains a pointer to a category pointed to by the index entry of the other;
5. The index entries of each point to thesaurus categories that in turn contain a pointer to the same category.

Morris and Hirst discovered that the first two are the most common types of relationships found in lexical chains, accounting for 90 percent of all relationships between elements of lexical chains built in their experiments.

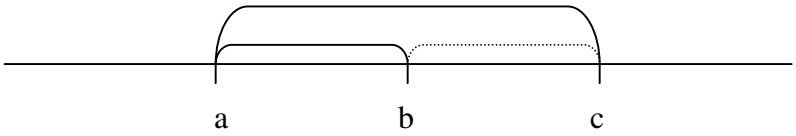
The use of the above criteria for the inclusion of a lexical item into a lexical chain results in the chain elements being connected with a diverse range of semantic relations. This prompted a question whether transitive connections can be considered as chain-forming relations. Morris and Hirst's analysis of sample texts showed that more than one transitive link weakens the relationship between words sharply. Therefore they considered only transitivity of one link for building lexical chains

[Morris91]. Two types of one-level transitive relations were counted when building lexical chains:

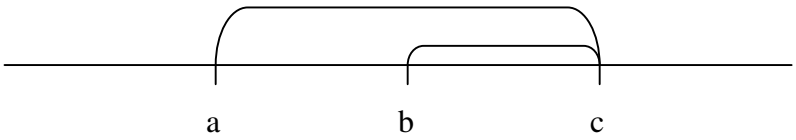
1. If word *a* is related to word *b*, and word *b* is related to word *c*, then word *a* is related to *c*.



2. If word *a* is related to word *b*, and word *a* is related to word *c*, then word *b* is related to *c*.



However, if two words are transitively linked through the third word which occurs later in the text, their relationship is not included into the chain as at the time of interpretation they were not related [Morris91].



In contrast Hoey, identifying only the cases of repetition where the range of semantic relations between words in a chain is less diverse, assumed that if a lexical item is related to one of the preceding elements in the chain, it is treated as being related to all other elements as well [Hoey91].

An important parameter characterising lexical chains, pointed out by Morris and Hirst, is the distance between chain elements. A text segment focusing on a particular subtopic is expected to have high density of related words. Morris and Hirst [Morris91] experimentally discovered that there can be up to two or three sentences between the chain elements. Subtopic development in the text reflects the flow of author's reasoning or the description of a certain pragmatic situation, therefore it is reasonable to expect that the author returns to the discussion of the same thing after digressing from it. Morris and Hirst called the resumption of the previous lexical chain after certain gap as *chain return*. Chain returns are indicators that the author resumes the subtopic entity he/she unfolded before. If the distance between chain elements is four or more sentences (no more than 19), than it is a signal of a chain return [Morris91].

Other parameters that characterise lexical chains and contribute to their strength are:

- Reiteration - the more words are repeated throughout the lexical chain, the stronger it is;

- Density - the more elements the lexical chain links within a certain text unit, the stronger is the chain;
- Length - the larger is the expanse of text that the lexical chain stretches through, the stronger it is [Morris91].

The motivation of Morris and Hirst's [Morris91] research of lexical chains was to see if lexical chains correspond to the intentional structure of discourse (proposed by Grosz and Sidner) and if they can be used to detect it. Experimentally they determined separately lexical chains and the intentional discourse structure. The comparison suggested strong correspondence between the two.

Morris and Hirst however did not automate their algorithm due to the unavailability of the electronic version of Roget's thesaurus. Therefore no large-scale testing of the algorithm was performed.

Later Hirst and St-Onge [Hirst97] attempted to adapt Morris and Hirst's algorithm for determining lexical chains, using WordNet relationships instead of Roget's thesaurus relationships.

Their method consists in attempting to link all the synsets associated with the two words being compared and ruling out those synsets which remain unconnected. They define three kinds of relations between words in the chain which are weighted according to their strength [Hirst97]:

- Extra-strong relation - between a word and its literal repetition;
- Strong relations
    (a) both words occur in the same synset;
    (b) there is a horizontal link between the two synsets associated with two different words;
    (c) there is any kind of link between a synset associated with each word if one word is a compound term that includes the other.
- Medium-strong relation occurs when there is an *allowable path* between two synsets associated with each word. The path must contain 2-5 links between synsets. The weight of such relation depends on the length of the path and the number of path direction changes.

The number of allowable paths between two words in Wordnet is restricted not only by the number of links that connect them, but also by the type of each lexical relation in the path. Since the semantics of the multilinked relationship is composed from the semantics of each of the links, their type and order is important. Hirst and St-Onge specified three rules according to which the multilink paths between words in Wordnet are acceptable for lexical chaining:

- No other relation may precede an upward link;
- No more than one change of direction is allowed;
- There can be a horizontal link between an upward and a downward relationships.

Hirst and St-Onge recognise the limitations of Wordnet, which caused problems such as including the wrong word into the chains or failing to include the word into the chain where it belongs. The major problem is that of a limited set of relations in Wordnet, which links words only by means of the most generic conceptual relations and does not provide for nonsystematic associative relations. Another problem is inconsistency in the semantic distance between words in Wordnet. This is evident when some words are semantically close but are separated by a large number of links in Wordnet and vice versa [Hirst97].

Hearst [Hearst94] in her research of the role lexical chains can play in text segmentation, discovered that locating individual chains in text does not give enough evidence of the topic boundaries. This is true especially for long texts with dense discussions, where it is not uncommon for multiple chains to overlap in the same stretch of text. Therefore it is important to analyse bunches of chains, eliciting information about topic boundaries from the points in text where one bunch of chain ends and another begins.

Interesting evidence about the distribution of link types in different types of texts was gathered by Ellman and Tait [Ellman2000]. Their algorithm for computing lexical chains is based on the algorithms by Morris and Hirst [Morris91] and Hirst and St-Onge [Hirst97]. The link types they compared were the following [Ellman2000]:

1. Two identical words;
2. Two words are members of the same Roget's thesaurus category;
3. Two words are members of the same group of categories in Roget's thesaurus;
4. Two words are linked through one level of thesaurus pointers.

The analysis of the link distribution showed that the most common is the link between two identical words, followed closely by the link between words belonging to the same category. The fourth type of link proved to be rather rare [Ellman2000]. This finding provides support for the methods which rely only on word repetition in the detection of lexical cohesion (cf. [Hoey91]). However the fact that the distribution of the link type 2 follows closely the distribution of the link type 1 means that similarity comparison between words belonging to the same thesaurus category can improve lexical chaining.

In our experimental study of lexical cohesion between query terms (chapter 8) we identify lexical links by detecting only the cases of simple lexical repetition, i.e. the identical stems (see section *8.2.1*). Ellman's findings suggest that our method identifies a large share of all lexical links.

2.3.3 Bonds between sentences and nets of bonds

Hoey pointed out that text cohesion is built not only of links between words, but also of semantic relationships between sentences. If sentences are not related as whole units, even though there are some lexically linked words found in them, they are no more than a disintegrated sequence of sentences sharing lexical context [Hoey91]. Following Winter [in Hoey91], he emphasised that it is important to interpret cohesion by taking into account the sentences where it is realised. For example, two

sentences in text can enter the relation, where the second one exemplifies the statement expressed in the previous sentence. Sentences do not have to be adjacent to be related, and relation can connect several sentences.

A cohesive relation between sentences is termed as *bond* [Hoey91]. Hoey defines bond between sentences as a sufficient number of lexical links between them. The number of lexical links the sentences must have to be bonded is a relative parameter, depending indirectly on the relative length and the lexical density of the sentences [Hoey91, p.92]. An empirical method for estimating a minimum number of links the sentences must have to form a bond is to rely on the proportion of sentence pairs that form bonds in text. If the proportion of sentences linked by any given number of links is too high, then it is important to increase the cut-off point, until the degree of connection is not above average [Hoey91].

It is notable that in Hoey's experiments, only 20% of bonded sentences were adjacent pairs. Analysing non-adjacent sentences, Hoey made and proved two claims about the meaning of bonds. The first claim is that the significance of bonds is greater than simply statistical evidence of the connection by repetition. Bonds between sentences are indicators of semantic relatedness, which is more than a sum of relations between linked words. The second claim is that a large number of bonded sentences are intelligible without recourse to the rest of the text. It is true that being semantically related, some bonded sentences are also coherent and can be interpreted on their own [Hoey91].

Our method of estimating the level of lexical cohesion between query terms was inspired by Hoey's method of detecting bonds between sentences. However, the two approaches have a number of fundamental differences in their aims and methodologies (for more detail see section *8.2.1*).

As already mentioned above, the relation between sentences can be multiple. All bonded sentences in a text or their subsets form what Hoey termed *nets* [Hoey91]. The analysis of nets of bonded sentences can throw light on the role each bonded sentence plays in text. Each sentence can be characterised by the number of preceding and following sentences in text it forms bonds with. Depending on these two sentence coordinates, Hoey distinguished four categories of sentences:

- Marginal sentences;
- Central sentences;
- Topic opening sentences;
- Topic closing sentences.

Sentences which form no bonds with other sentences are termed marginal, for their contents "neither builds lexically upon what has gone before nor provides the lexis for subsequent statements" [Hoey91, p. 105]. It is important, however, to note that although they do not directly contribute to the topic development, they can provide additional information, necessary for its understanding.

Central sentences are those that have exceptionally high number of bonds with other sentences. The minimum number of bonds required to define a sentence as central is relative, what is similar to the relative definition of the minimum number of links for bonding.

Sentences which are related to more following than preceding sentences are topic-opening, and those related to more preceding than following sentences are topic-closing.

An interesting study of nets was made earlier by Skorochod'ko [Skorochod'ko72], who formed a classification of nets, or "semantic networks" in his terminology. Identifying semantic relatedness between sentences through occurrence of related or repeated words, he built the semantic network of the text by representing it as a graph, where the sentences and the relations between them are shown as nodes and arcs respectively. He distinguished four types of nets that represent semantic organisation of text (figure 2.2).

The relatedness between sentences in Skorochod'ko's approach is relative. One of the factors determining strength of the relationship is the number of lexical links which sentences have between them.

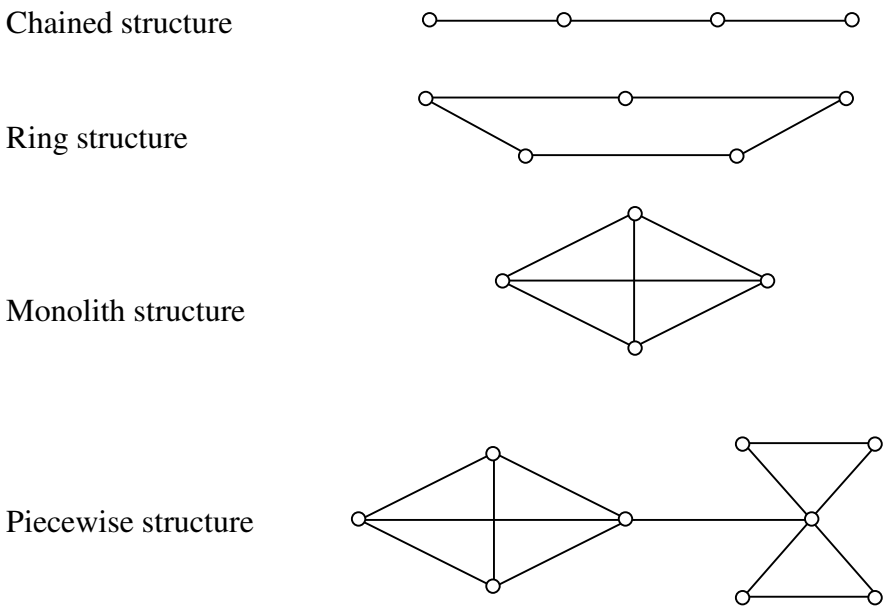


Figure 2.2. Semantic structure types by Skorochod'ko [Skorochod'ko72, p.1180]

In texts with nets of chained and ring type predominantly adjacent sentences are bonded, distance relations being uncommon. In monolith type of nets, each sentence is connected to the majority of others. Piecewise nets consist of groups with densely interconnected sentences, all groups being connected in turn.

The types of semantic networks presented above determine the semantic organisation of the text, in other words provide some information on the topic development in text, whether it is presented through sequential narration like in chain or monolith type, or

through dense discussion with frequent recourse to what was said before or what will follow, like in monolith or piecewise structures.

As can be seen from the research findings described in the above three subsections, cohesion is a text-organising component. By identifying cohesive relations it is possible to uncover (at least with some degree of accuracy) topic boundaries in text and ultimately the overall semantic organisation of text. However, lexical cohesion can also throw light on smaller scale relations existing between text elements. In our third group of study (chapter 8) we explored lexical cohesion between query terms and its association with relevance.

2.4 Collocation

2.4.1 Definitions of collocation

The earliest understanding of collocation, as mentioned in section *2.3.1*, was formed by Firth [Firth58]. Collocation in the Firthian sense is a name given to lexical items which co-occur with more than random probability in text. This is a rather broad understanding of the phenomenon of co-occurring words. Later many linguists attempted to narrow down the meaning of the term collocation or build a completely new vision of what collocation is. Many proposed definitions often reflect the background the linguists come from and are formulated in the light of the subject of their research.

Both Halliday and Hasan [Halliday76] and Hoey [Hoey91] who focused on the study of cohesion emphasised that collocation is a realisation of lexical cohesion in text. As mentioned earlier (section *2.3.1*) Halliday and Hasan gave a rather vague definition of collocation:

> "Here we shall simply group together (...) all lexical cohesion that is not covered by what we have called 'reiteration' - and treat it under the general heading of collocation, or collocational cohesion ..." [Halliday76, p.287]

They emphasise that there must be a lexical-semantic relation between words to be considered collocates. Although they recognise that cohesion between two collocates in text is not so much due to the existence of a lexical-semantic relation, as to the fact that they have a tendency to occur in the same environment, they claim that words co-occur because they are in some kind of lexical-semantic relation. Because they were unable to give a systematic classification of lexical relationships holding in collocations, Hasan later suggested to take out collocation from linguistic terminology until such classification is formed [Hasan84]. It would be more reasonable, perhaps, to admit the variety of lexical-semantic relationships existing between collocates, but to center the definition around the idea of typicality or regularity of their co-occurrence.

Hoey also interpreted the notion of collocation in the light of cohesion studies. He asserts that "collocation is the direct result of bonding (and of similar relations

between sentences in and across texts)." [Hoey91, p.219]. Unlike Halliday and Hasan he avoided including the undefined range of lexical-semantic relations into his understanding of collocation, focusing solely on the recurrence of words' appearance together in text: "Words only have collocations because they appear in recurring combinations in text ..." [Hoey91, p.219].

The recognition of collocation as the mechanism for creating text cohesion by [Hoey91] and [Halliday76] means that collocation in their sense is a topic-forming relationship, whose elements - collocates - can be separated by some stretches of text. Their co-occurrence is due not to grammatical patterns and restrictions, but to belonging to the same topic. They indeed exhibit certain semantic relations which are reflected in their usage in the same thematic environment, in the same topic. Below another widely accepted approach to collocation, based on a rather different stance, is given.

Some linguists understand by collocation a more than random use of words next to each other either due to referential association of their meanings, or due to habitual or customary patterns which evolved in the language in use [Robins89, Palmer81]. Robins distinguished collocations which are related to the situational and referential meaning of the collocates, e.g. *bright day*, from collocations less tightly connected with the reference, but whose use is more habitual, e.g. *white coffee*, where the word 'white' does not refer to the colour of the referent [Robins89, p.65]. Palmer drew a distinction between these types of collocates based on the idea of collocational restriction. Collocations can have different degrees of restriction. He distinguished three types of collocational restriction [Palmer81, p.79]:

1. Collocational restriction based entirely on the meanings of the words (e.g., *bright day*, *deep sea*);
2. Collocational restriction based on a range of words. A word can collocate with a range of words that have some common semantic feature (e.g., *pack of wolves*, *pack of hounds*, *pride of lions* but not *pack of lions* or *pride of wolves*);
3. Collocational restriction wholly due to habitual and customary use of the words (e.g., *rancid butter, rancid bacon*, *sour milk* but not *rancid milk* or *sour butter*).

It is possible to explain collocations of the first type entirely through the meaning of their elements. Collocations of the second and third types can or cannot be explained semantically through the meanings of their elements. For example, Palmer suggested that restricted collocations like 'white wine' and 'white coffee' could possibly be semantically explained by saying that 'white' means "with the lightest of the normal colours associated with the entity" [Palmer81, p.77]. But he admitted that it is impossible to draw a clear line between those collocations which can be semantically explained and those which cannot.

Palmer [Palmer81] and Robins [Robins89] recognised idioms as a special kind of collocations, whose meaning is not deducible from the meanings of their components and often has nothing to do with them semantically (e.g., *red herring*). Palmer also argued that the third type of restricted collocations can be referred to as partial idioms, where one word is used in its usual meaning (*'coffee'* in *white coffee*), but the other acquires a meaning specific to the collocation (*'white'*) [Palmer81].

Some scientists narrowed down the meaning of collocation to refer only to restricted type of collocations, whose meaning cannot be completely derived from the meaning of their elements. Manning and Schuetze [Manning99] defined collocation as grammatically bound elements occurring in a certain order which are characterised by limited compositionality, by which they mean the impossibility of deriving the meaning of the total from the meanings of its parts. They admit the existence of syntagmatic word associations across larger expanses of text, bound topically, not grammatically, but they suggest calling such associations 'co-occurrences' and to reserve the term 'collocation' only for grammatically bound combinations.

Manning and Schuetze highlighted the following categories among the collocations in the narrow sense as grammatically bound constructions [Manning99, p. 174]:

1. Light verbs (e.g., *make a decision*, *do a favour*);
2. Phrasal verbs (e.g., *to check in*, *to cut down*);
3. Proper names;
4. Terminological expressions (e.g., *fission reactor fuel*)

Defining collocation through the notion of limited compositionality, they note that collocations can have different degrees of invariability. Some collocations allow certain substitutions of its elements (this co-relates with Palmer's second and third type of collocational restriction), but some not - proper names, idioms, some terminological expressions.

In the light of the approaches described above in this thesis it is recognised that there are two major types of collocation:

1. Collocation due to lexical-grammatical or habitual restrictions. These restrictions limit the choice of words that can be used in the same grammatical structures with the word in question.
2. Collocation due to a typical occurrence of a word in a certain thematic environment. Two words hold a certain lexical-semantic relation, i.e. their meanings are close semantically, therefore they tend to occur in the same topics in texts.

Lexical-grammatical/habitual restrictions and lexical-semantic relations are the linguistic factors that cause the phenomenon of word collocation. The nature of these linguistic factors is quite complex and the development of a general method to identify collocations in text through the analysis of these factors is not an easy task, as evident from the failed attempt to impose strict classification on the phenomenon of collocation by [Halliday76] (see earlier section *2.3.1*). Instead it is possible to isolate this phenomenon in text empirically, not by looking for what causes it, but by identifying what characterises it most typically - namely, more than random probability of its occurrence.

The empirical approaches to identification of collocates from text were developed within the branch of linguistics based entirely on the analysis of empirical data - corpus linguistics. Corpus linguistics which developed in the last 30 years, emerging from the work of such linguists as Kucera and Francis [Kucera67] and Sinclair [Sinclair74], concerns itself with extrapolating general conclusions about various linguistic phenomena through the empirical analysis of corpora. Corpora are large

collections of texts, sampled to be maximally representative of a general/specific type of language in use [Wilson96]. Empirical identification of collocations is based on using various statistical parameters of lexical units elicited from the corpora. The statistical methods of finding collocations in text used in this project will be described in chapter 5.

Sinclair and Jones [Sinclair74] were the first to attempt corpus-based analysis of collocations. Although the size of the corpus they analysed was relatively small - 135000 words in spoken corpus and 12000 words in written corpus - compared to modern corpora typically several million words long, their work is of significant value as it defined the framework for corpus-based collocation analysis and yielded some information about patterns of collocation behaviour in text.

The major notions of collocation analysis introduced in [Sinclair74] and systemised further in [Sinclair91] are those of 'node', 'collocate', 'significant collocate' and 'span'. A 'node' is defined as "an item whose total pattern of co-occurrence with other words is under examination" [Sinclair74, p. 22]. While a 'collocate' is "any item which appears with the node within a specified environment." [ibid.].

Sinclair and Jones use the term 'collocation' to refer to any co-occurrence of two words within certain environment in text. To refer to those collocations which occur with more than random probability in text they use the term 'significant collocation'. "'Significant' collocation is regular collocation between items, such that they co-occur more often than their respective frequencies and the length of text in which they appear would predict." [Sinclair74, p. 25].

The term 'span' is used to refer to the stretch of text around the node within which words are considered to be its collocates [Sinclair74, p.27]

The above terms have been adopted in later corpus-based collocation research and will be also used in these senses throughout this thesis.

2.4.2 Patterns of collocation in text

This section will cover some patterns discovered by corpus linguists using statistical analysis of collocation and word distribution in text. The probability of occurrence of each word in text is affected by the presence of other words around it. More specifically, it is possible to say that due to the effect of collocation the appearance of each word in text influences to a varying degree the statistics of word occurrences around it [Beeferman97]. The major questions in this context are the following:

1. How far the influence of each word extends in text?
2. Is the influence of a word consistent throughout the word's environment?

The first question, concerning the size of the environment throughout which a word exerts its influence, remains arguable. Most researchers so far have made empirical decisions about the span of a word's environment, based mainly on the practical considerations of the particular research task. The span is measured either in syntactic units, such as phrases, sentences, paragraphs or even entire texts [Harper78], or by the number of words to the left and right of the node, for example, 4 words [Sinclair74], 5

words [Church90], 400 words [Beeferman97], 4,10, 50 words [Edmonds97]. Such lack of uniformity can be partly attributed to different interpretations of the concept *the environment of a word*. As already mentioned in the previous section of the thesis we distinguish two types of collocations: (1) those subject to lexical-grammatical restrictions and (2) collocations influenced by lexical-semantic or topical relations. The first type of collocations occur within a *short-span* environment, whereas the second type - within a *long-span* environment. Our experimental study of statistical parameters of the environment of a word indicated a marked difference between word distributions in the immediate span of the node (4-5 words either sides) and the bigger span, which correlates with the results of a similar experiment but on a smaller span of 10 words by [Sinclair74]. Our experiment consisted in counting the number of observed token types for each slot within 50 words both sides of the node. We analysed windows around all instances of 40 randomly selected indexing terms in the stemmed[1] TREC AP collection. The results showed that the variability of collocates within the immediate span was much lower than further away from the node. Immediate collocates maintained stable joint co-occurrence with the node, indicative of lexical-grammatical factors (e.g. phrasal verbs, compound terms). Term variability further away from the node was distinctly higher[2] (figure 2.3).

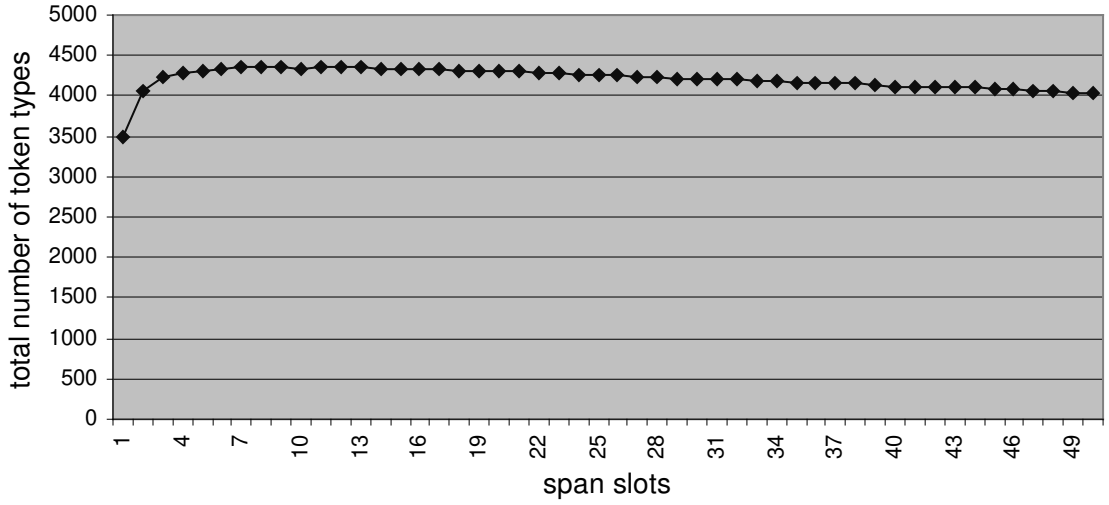


Figure 2.3. Distance distribution of token types in the windows of 50 words either sides of the nodes.

This experiment allowed us to identify only the boundary between the short-span and the long-span influences of the word. Beeferman et al. [Beeferman97] in their study of the effect distance has on triggering and prediction capabilities of collocates (or "trigger-words" in their terminology) obtained empirical evidence about the long-distance influence patterns of words. They analysed distance distributions in two groups of collocates: self-collocates - repetition of the same lexeme, e.g. (*gene, gene*), and non-self collocates, where collocate 1 and collocate 2 are different lexemes, e.g. (*gene, chromosome*). The technique they used consisted in calculating for each

[1] Okapi parsing algorithm was used to stem words and filter out stop words

[2] The gradual fall of the curve is due to the fact that windows are frequently truncated by hitting the document boundary.

distance k the probability that two trigger words are separated by exactly $k \pm 2$ words. The results showed that a word's influence on statistical distribution of words around it stretches as far as several hundred words, levelling off by about 400 words.

The main objective of Beeferman's research, however, was to prove the hypothesis that statistical parameters within a word's environment are not static, but are directly dependent on the distance away from the node. This leads us to the second question raised earlier about the consistency of the word's prediction capabilities. Beeferman's major discovery was that the influence of the word decays exponentially. Another observation was characteristic of the group of self-collocates and demonstrated that the word-triggering effect increases from the distance of 1 word to 25 words, but then falls gradually. They termed this phenomenon *the lexical exclusion principle*, due to the fact that lexical-syntactic restrictions prevent self-collocates or repetitions to appear within short distances in text [Beeferman97].

Mason [Mason97] discovered interesting patterns of the word's influence on its environment which he called *lexical gravity*. His corpus-based study of collocation proved among others the claims that (1) each word has an individual pattern of influence and (2) lexical and grammatical items have different patterns of influence on their environment. To test these claims he measured the variability of words in each slot for arbitrary distances using type-token ratio (TTR). Experiments showed that lexical gravity is not symmetrical, i.e. a word can have different degree of influence on its preceding and following neighbours and, most importantly, that each word has different patterns of influence on its immediate environment. These findings concern only short-span collocations, and can be theoretically explained by the fact that each word imposes different degrees of lexical-grammatical restrictions on its short-span environment, where, especially in restricted type of collocations, some words tend to occupy fixed slots relative to the node. Whereas, in the long-span environment of the node, words related to it semantically can occupy any slot.

While lexical units have a generally similar type of pattern, grammatical items have strikingly different patterns. Another feature characteristic of some grammatical items is negative gravity, i.e. higher than average variability. This is due to the fact that grammatical items form patterns with grammatical classes and not individual words. Articles, for example, are followed by any noun or noun phrase, i.e. they restrict the selection of the following word only grammatically, but not lexically or semantically.

Not all words in the environment of the node are influenced or predicted by its occurrence. The environment of a word consists of several types of units: high-frequency words (e.g. prepositions, articles, auxiliary verbs), which co-occur frequently with any word; one-off collocates, with low joint frequency and significant collocates with more than random probability of co-occurrence with the node. In the long-span environment, significant collocates are those that are related to the node semantically, their non-random co-occurrence being due to the fact that both belong to the same or related semantic fields and consequently tend to be used in the same topics. The number of such topic-discriminating words is relatively small in comparison to the total number of words in a stretch of text describing a topic [Renouf93]. Renouf argued that the majority of words in text are topic-independent or related to more than one topic. She distinguished eight types of such words: very common words, discourse organising words, homonyms, semi-technical words, words

with several technical senses, metaphors, typographically ambiguous words [Renouf93]. Genuine topically-dependent words, as opposed to high-frequency or context-independent words, are expected to have significant amount of association with the node. Significant collocates which can predict each other's occurrence are distinguished from chance word pairs by using various statistical measures of association. The statistical methods used in this project will be detailed in chapter 5. Section *3.3 Approaches to collocation in IR* will also cover other research of collocations in the IR context.