2

Semantic Domains

In this chapter we define the concept of Semantic Domain, recently introduced in Computational Linguistics [56] and successfully exploited in NLP [29]. This notion is inspired by the "Theory of Semantic Fields" [88], a structural model for lexical semantics proposed by Jost Trier at the beginning of the last century. The basic assumption is that the lexicon is structured into Semantic Fields: semantic relations among concepts belonging to the same field are very dense, while concepts belonging to different fields are typically unrelated. The theory of Semantic Fields constitutes the linguistic background of this work, and will be discussed in detail in Sect. 2.1. The main limitation of this theory is that it does not provide an objective criterion to distinguish among Semantic Fields. The concept of linguistic game allows us to formulate such a criterion, by observing that linguistic games are reflected by texts in corpora.

Even if Semantic Fields have been deeply investigated in structural linguistics, computational approaches for them have been proposed quite recently by introducing the concept of *Semantic Domain* [59]. Semantic Domains are clusters of terms and texts that exhibit a high level of lexical coherence, i.e. the property of domain-specific words to co-occur together in texts. In the present work, we will refer to these kinds of relations among terms, concepts and texts by means of the term *Domain Relations*, adopting the terminology introduced by [56].

The concept of Semantic Domain extends the concept of Semantic Field from a lexical level, in which it identifies a set of domain related lexical concepts, to a textual level, in which it identifies a class of similar documents. The founding idea is the lexical coherence assumption, that has to be presupposed to guarantee the existence of Semantic Domains in corpora.

This chapter is structured as follows. First of all we discuss the notion of Semantic Field from a linguistic point of view, reporting the basics of Trier's work and some alternative views proposed by structural linguists, then we illustrate some interesting connections with the concept of linguistic game (see Sect. 2.2), that justify our further corpus-based approach. In Sect. 2.3

A. Gliozzo and C. Strapparava, *Semantic Domains in Computational Linguistics*,
DOI: 10.1007/978-3-540-68158-8_2, © Springer-Verlag Berlin Heidelberg 2009

we introduce the notion of Semantic Domain. Then, in Sect. 2.4, we focus on the problem of defining a set of requirements that should be satisfied by any "ideal" domain set: *completeness*, *balancement* and *separability*. In Sect. 2.5 we present the lexical resource WordNet Domains, a large scale repository of domain information for lexical concepts. In Sect. 2.6 we analyze the relations between Semantic Domains at the lexical and at the textual levels, describing the property of *Lexical Coherence* in texts. We will provide empirical evidence for it, by showing that most of the lexicon in documents belongs to the principal domain of the text, giving support to the *One Domain per Discourse* hypothesis. The lexical coherence assumption holds for a wide class of words, namely *domain words*, whose senses can be mainly disambiguated by considering the domain in which they are located, regardless of any further syntactic information. Finally, we report a literature review describing all the computational approaches to represent and exploit Semantic Domains we have found in the literature.

2.1 The Theory of Semantic Fields

Semantic Domains are a matter of recent interest in Computational Linguistics [56, 59, 29], even though their basic assumptions are inspired from a long standing research direction in structural linguistics started in the beginning of the last century and widely known as "The Theory of Semantic Fields" [55]. The notion of *Semantic Field* has proved its worth in a great volume of studies, and has been mainly put forward by Jost Trier [87], whose work is credited with having "opened a new phase in the history of semantics" [89].

In that work, it has been claimed that the lexicon is structured in clusters of very closely related concepts, lexicalized by sets of words. Word senses are determined and delimitated only by the meanings of other words in the same field. Such clusters of semantically related terms have been called Semantic Fields,[1] and the theory explaining their properties is known as "The theory of Semantic Fields" [92].

This theory has been developed in the general framework of Saussure's structural semantics [20], whose basic claim is that a word meaning is determined by the "horizontal" paradigmatic and the "vertical" syntagmatic relations between that word and others in the whole language [55]. Structural semantics is the predominant epistemological paradigm in linguistics, and it is very much appreciated in Computational Linguistic. For example, many machine readable dictionaries describe the word senses by means of semantic networks representing relations among terms (e.g. WordNet [66]). The Semantic Fields Theory goes a step further in the structural approach to lexical

[1] There is no agreement on the terminology adopted by different authors. Trier uses the German term *wortfeld* (literally "word field" or "lexical field" in Lyons' terminology) to denote what we call here semantic field.

semantics by introducing an additional aggregation level and by delimiting to what extent paradigmatic relations hold.

Semantic Fields are conceptual regions shared out amongst a number of words. Each field is viewed as a partial region of the whole expanse of ideas that is covered by the vocabulary of a language. Such areas are referred to by groups of semantically related words, i.e. the Semantic Fields. Internally to each field, a word meaning is determined by the network of relations established with other words.

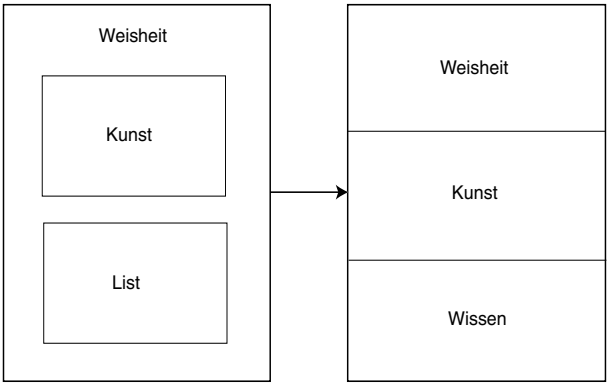


Fig. 2.1. The INTELLECTUAL field's structure in German at around 1200 AD (left) and at around 1300 AD (right)

Trier provided an example of its theory by studying the INTELLECTUAL field in German, illustrated in Fig. 2.1. Around 1200, the words composing the field were organized around three key terms: Weisheit, Kunst and List. Kunst meant knowledge of courtly and chivalric attainments, whereas List meant knowledge outside that sphere. Weisheit was their hypernym, including the meaning of both. One hundred years later a different picture emerged. The courtly world has disintegrated, so there was no longer a need for a distinction between courtly and non-courtly skills. List has moved towards its modern meaning (i.e. `cunning`) and has lost its intellectual connotations; then it is not yet included into the INTELLECTUAL field. Kunst has also moved towards its modern meaning indicating the result of artistic attainments. The term Weisheit now denotes religious or mystical experiences, and Wissen is a more general term denoting `knowledge`. This example clearly shows that word meaning is determined only by internal relations between the lexicon of the field, and that the conceptual area to which each word refers is delimitated in opposition with the meaning of other concepts in the lexicon.

A relevant limitation of Trier's work is that a clear distinction between lexical and conceptual fields is not explicitly done. The lexical field is the set of words belonging to the semantic field, while the conceptual field is the

set of concepts covered by terms of the field. Lexical fields and conceptual fields are radically different, because they are composed by different objects. From an analysis of their reciprocal connections, many interesting aspects of lexical semantics emerge, as for example ambiguity and variability. The different senses of ambiguous words should be necessarily located into different conceptual fields, because they are characterized by different relations with different words. It reflects the fact that ambiguous words are located into more than one lexical field. On the other hand, variability can be modeled by observing that synonymous terms refer to the same concepts, then they will be necessarily located in the same lexical field. The terms contained in the same lexical field recall each other. Thus, the distribution of words among different lexical fields is a relevant aspect to be taken into account to identify word senses. Understanding words in contexts is mainly the operation of locating them in the appropriate conceptual fields.

Regarding the connection between lexical and conceptual fields, we observe that most of the words characterizing a Semantic Field are domain-specific terms, then they are not ambiguous. Monosemic words are located only into one field, and correspond univocally to the denoted concepts. As an approximation, conceptual fields can be analyzed by studying the corresponding lexical fields. The correspondence between conceptual and lexical fields is of great interest for computational approaches to lexical semantics. In fact, the basic objects manipulated by most text processing systems are words. The connection between conceptual and lexical fields can then be exploited to shift from a lexical representation to a deeper conceptual analysis.

Trier also hypothesized that Semantic Fields are related between each other, so as to compose a higher level structure, that together with the low level structures internal to each field composes the structure of the whole lexicon. The structural relations among Semantic Fields are much more stable than the low level relations established among words. For example, the meaning of the words in the INTELLECTUAL field has changed largely in a limited period of time, but the INTELLECTUAL field itself has pretty much preserved the same conceptual area. This observation explains the fact that Semantic Fields are often consistent among languages, cultures and time.

As a consequence there exists a strong correspondence among Semantic Fields of different languages, while such a strong correspondence cannot be established among the terms themselves. For example, the lexical field of COLORS is structured differently in different languages, and sometimes it is very difficult, if not impossible, to translate names of colors, even whether the chromatic spectrum perceived by people in different countries (i.e. the conceptual field) is the same. Some languages adopt many words to denote the chromatic range to which the English term `white` refers, distinguishing among different degrees of "whiteness" that have no direct translation in English. Anyway, the chromatic range covered by the COLORS fields of different languages is evidently the same. The meaning of each term is defined by virtue of its opposition with other terms of the same field. Different languages have

different distinctions, but the field of COLORS itself is a constant among all the languages.

Another implication of the Semantic Fields Theory is that words belonging to different fields are basically unrelated. In fact, a word meaning is established only by the network of relations among the terms of its field. As far as paradigmatic relations are concerned, two words belonging to different fields are then unrelated. This observation is crucial form a methodological point of view. The practical advantage of adopting the Semantic Field Theory in linguistics is that it allows a large scale structural analysis of the whole lexicon of a language, which is otherwise infeasible. In fact, restricting the attention to a particular lexical field is a way to reduce the complexity of the overall task of finding relations among words in the whole lexicon, that is evidently quadratic in the number of words. The complexity of reiterating this operation for each Semantic Field is much lower than that of analyzing the lexicon as a whole. From a computational point of view, the memory allocation and the computation time required to represent an "all against each other" relation schema is quadratic on the number of words in the language (i.e. $O(|\mathcal{V}|^2)$). The number of operations required to compare only those words belonging to a single field is evidently much lower (i.e. $\mathrm{O}\left(\left(\frac{|\mathcal{V}|}{d}\right)^2\right)$, assuming that the vocabulary of the language is partitioned into d Semantic Fields of equal sizes). To cover the whole lexicon, this operation has to be iterated d times. The complexity of the task to analyze the structure of the whole lexicon is then $O\left(d\left(\frac{|\mathcal{V}|}{d}\right)^2\right) = O\left(\frac{|\mathcal{V}|^2}{d}\right)$. Introducing the additional constraint that the number of words in each field is bounded, where k is the maximum size, we obtain $d \geqslant \frac{|\mathcal{V}|}{k}$. It follows that $O\left(\frac{|\mathcal{V}|^2}{d}\right) \leqslant O(|\mathcal{V}|k)$. Assuming that k is an "a priori" constant, determined by the inherent optimization properties required by the domain-specific lexical systems to be coherent, the complexity of the task to analyze the structure of the whole lexicon decreases by one order (i.e. $O(|\mathcal{V}|)$), suggesting an effective methodology to acquire semantic relations among domain-specific concepts from texts

The main limitation of Trier's theory is that it does not provide any objective criterion to identify and delimitate Semantic Fields in the language. The author himself admits "what symptoms, what characteristic features entitle the linguist to assume that in some place or other of the whole vocabulary there is a field? What are the linguistic considerations that guide the grasp with which he selects certain elements as belonging to a field, in order then to examine them as a field?" [88]. The answer to this question is an issue opened by Trier's work, and it has been approached by many authors in the literature.

Trier's theory has been frequently associated to Weisgerber's "theory of contents" [93], claiming that word senses are supposed to be immediately given by virtue of the extra-lingual contexts in which they occur. The main

problem of this referential approach is that it is not clear how extra-lingual contexts are provided; then those processes are inexplicable and mysterious.

The referential solution, adopted to explain the field of colors, is straightforward as long as we confine ourselves to fields that are definable with reference to some obvious collection of external objects, but it is not applicable to abstract concepts. The solution proposed by Porzig was to adopt syntagmatic relations to identify word fields [74]. In his view, a Semantic Field is the range of words that are capable of meaningful connection with a given word. In other words, terms belonging to the same field are syntagmatically related to one or more common terms, as for example the set of all the possible subjects or objects for a certain verb, or the set of nouns to which an adjective can be applied. Words in the same field would be distinguished by the difference of their syntagmatic relations with other words.

A less interesting solution has been proposed by Coseriu [15], founded upon the assumption that there is a fundamental analogy between the phonological opposition of sounds and the "lexematic" opposition of meanings. We do not consider this position.

2.2 Semantic Fields and the *meaning-is-use* View

In the previous section we have pointed out that the main limitation of Trier's theory is the gap of an objective criterion to characterize Semantic Fields. The solutions we have found in the literature rely on very obscure notions, of scarce interest from a computational point of view. To overcome such a limitation, in this section we introduce the notion of Semantic Domain (see Sect. 2.3).

The notion of Semantic Domain improves that of Semantic Fields by connecting the structuralist approach in semantics to the *meaning-is-use* assumption introduced by Ludwig Wittgenstein in his celebrated Philosophical Investigations [94]. A word meaning is its use into the concrete "form of life" where it is adopted, i.e. the *linguistic game*, in Wittgenstein's terminology. Words are then meaningful only if they are expressed in concrete and situated linguistic games that provide the conditions for determining the meaning of natural language expressions. To illustrate this concept, Wittgenstein provided a clarifying example describing a very basic linguistic game: "... Let us imagine a language ... The language is meant to serve for communication between a builder A and an assistant B. A is building with building-stones; there are blocks, pillars, slabs and beams. B has to pass the stones, and that in the order in which A needs them. For this purpose they use a language consisting of the words block, pillar, slab, beam. A calls them out; – B brings the stone which he has learnt to bring at such-and-such a call. – Conceive of this as a complete primitive language." [94].

We observe that the notions of linguistic game and Semantic Field show many interesting connections. They approach the same problem from two different points of view, getting to a similar conclusion. According to Trier's

view, words are meaningful when they belong to a specific Semantic Field, and their meaning is determined by the structure of the lexicon in the field. According to Wittgenstein's view, words are meaningful when there exists a linguistic game in which they can be formulated, and their meaning is exactly their use. In both cases, meaning arises from the wider contexts in which words are located.

Words appearing frequently in the same linguistic game are likely to be located in the same lexical field. In the previous example the words block, pillar, slab and beam have been used in a common linguistic game, while they clearly belong to the Semantic Field of BUILDING INDUSTRY. This example suggests that the notion of linguistic game provides a criterion to identify and to delimit Semantic Fields. In particular, the recognition of the linguistic game in which words are typically formulated can be used as a criterion to identify classes of words composing lexical fields. The main problem of this assumption is that it is not clear how to distinguish linguistic games between each other. In fact, linguistic games are related by a complex network of similarities, but it is not possible to identify a set of discriminating features that allows us to univocally recognize them. "I can think of no better expression to characterize these similarities than 'family resemblances'; for the various resemblances between members of a family: build, features, color of eyes, gait, temperament, etc. etc. overlap and criss-cross in the same way. And I shall say: 'games' form a family" ([94], par. 67).

At first glance, the notion of linguistic game is no less obscure than those proposed by Weisgerber. The first relies on a fuzzy idea of family resemblance, the latter refer to some "external" relation with the real world. The main difference between those two visions is that the former can be investigated within the structuralist paradigm. In fact, we observe that linguistic games are naturally reflected in texts, allowing us to detect them from a word distribution analysis on a large scale corpus. In fact, according to Wittgenstein's view, the content of any text is located in a specific linguistic game, otherwise the text itself would be meaningless. Texts can be perceived as open windows through which we can observe the connections among concepts in the real world. Frequently co-occurring words in texts are then associated to the same linguistic game.

It follows that lexical fields can be identified from a corpus-based analysis of the lexicon, exploiting the connections between linguistic games and Semantic Fields already depicted. For example, the two words fork and glass are evidently in the same lexical field. A corpus-based analysis shows that they frequently co-occur in texts, then they are also related to the same linguistic game. On the other and, it is not clear what would be the relation among water and algorithm, if any. They are totally unrelated simply because the concrete situations (i.e. the linguistic games) in which they occur are in general distinct. It reflects the fact that they are often expressed in different texts, then they belong to different lexical fields.

Words in the same field can then be identified from a corpus-based analysis. In Sect. 2.6 we will describe in detail the lexical coherence assumption, that ensures the possibility of performing such a corpus-based acquisition process for lexical fields. Semantic Domains are basically Semantic Fields whose lexica show high lexical coherence.

Our proposal is then to merge the notion of linguistic game and that of Semantic Field, in order to provide an objective criterion to distinguish and delimit lexical fields from a corpus-based analysis of lexical co-occurences in texts. We refer to this particular view on Semantic Fields by using the name Semantic Domains. The concept of Semantic Domain is the main topic of this work, and it will be illustrated more formally in the following section.

2.3 Semantic Domains

In our usage, Semantic Domains are common areas of human discussion, such as ECONOMICS, POLITICS, LAW, SCIENCE, etc. (see Tab. 2.2), which demonstrate lexical coherence. Semantic Domains are Semantic Fields, characterized by sets of *domain words*, which often occur in texts about the corresponding domain. Semantic Domains can be automatically identified by exploiting a *lexical coherence* property manifested by texts in any natural language, and can be profitably used to structure a semantic network to define a computational lexicon.

As well as Semantic Fields, Semantic Domains correspond to both lexical fields and conceptual fields. In addition, the lexical coherence assumption allows us to represent Semantic Domains by sets of domain-specific text collections.[2] The symmetricalness of these three levels of representation, allows us to work at the preferred one. Throughout this book we will mainly adopt a lexical representation because it presents several advantages from a computational point of view.

Words belonging to lexical fields are called domain words. A substantial portion of the language terminology is characterized by domain words, whose meaning refers to lexical concepts belonging to the specific domains. Domain words are disambiguated when they are collocated into domain-specific texts by simply considering domain information [32].

Semantic Domains play a dual role in linguistic description. One role is characterizing word senses (i.e. *lexical concepts*), typically by assigning domain labels to word senses in a dictionary or lexicon (e.g. crane has senses in the domains of ZOOLOGY and CONSTRUCTION).[3] A second role is to characterize

[2] The textual interpretation motivates our usage of the term "Domain". In fact, this term is often used in Computational Linguistics either to refer to a collection of texts regarding a specific argument, as for example BIOMEDICINE, or to refer to ontologies describing a specific task.

[3] The WORDNET DOMAINS lexical resource is an extension of WORDNET which provides such domain labels for all synsets [56].

texts, typically as a generic level of Text Categorization (e.g. for classifying news and articles) [80].

At the lexical level Semantic Domains identify clusters of (domain) related lexical-concepts. i.e. sets of domain words. For example the concepts of `dog` and `mammal`, belonging to the domain ZOOLOGY, are related by the **is_a** relation. The same hold for many other concepts belonging to the same domain, as for example `soccer` and `sport`. On the other hand, it is quite infrequent to find semantic relations among concepts belonging to different domains, as for example `computer_graphics` and `mammifer`. In this sense Semantic Domains are shallow models for Semantic Fields: even if deeper semantic relations among lexical concepts are not explicitly identified, Semantic Domains provide a useful methodology to identify a class of strongly associated concepts. Domain relations are then crucial to identify ontological relations among terms from corpora (i.e. to induce automatically structured Semantic Fields, whose concepts are internally related).

At a text level domains are cluster of texts regarding similar topics/subjects. They can be perceived as collections of domain-specific texts, in which a generic corpus is organized. Examples of Semantic Domains at the text level are the subject taxonomies adopted to organize books in libraries, as for example the Dewey Decimal Classification [14] (see Sect. 2.5).

From a practical point of view, Semantic Domains have been considered as lists of related terms describing a particular subject or area of interest. In fact, term-based representations for Semantic Domains are quite easy to be obtained, e.g. by exploiting well consolidated and efficient shallow parsing techniques [36]. A disadvantage of term-based representations is lexical ambiguity: polysemous terms denote different lexical concepts in different domains, making it impossible to associate the term itself to one domain or the other. Anyway, term-based representations are effective, because most of the domain words are not ambiguous, allowing us to biunivocally associate terms and concepts in most of the relevant cases.

Domain words are typically highly correlated within texts, i.e. they tend to co-occur inside the same types of texts. The possibility of detecting such words from text collections is guaranteed by a *lexical coherence* property manifested by almost all the texts expressed in any natural language, i.e. the property of words belonging to the same domain to frequently co-occur in the same texts.[4]

Thus, Semantic Domains are a key concept in Computational Linguistics because they allow us to design a set of totally automatic corpus-based acquisition strategies, aiming to infer shallow Domain Models (see Chap. 3) to be exploited for further elaborations (e.g. ontology learning, text indexing, NLP systems). In addition, the possibility of automatically acquiring Semantic Domains from corpora is attractive both from an applicative and theoretical

[4] Note that the lexical coherence assumption is formulated here at a term level as an approximation of the strongest original claim that holds at the concept level.

point of view, because it allows us to design algorithms that can fit easily domain-specific problems while preserving their generality.

The next sections discuss two fundamental issues that arise when dealing with Semantic Domains in Computational Linguistics: (i) how to choose an appropriate partition for Semantic Domains; and (ii) how to define an adequate computational model to represent them. The first question is both an ontological and a practical issue, that requires us to take a (typically arbitrary and subjective) decision about the set of the relevant domain distinctions and their granularity. In order to answer the second question, it is necessary to define a computational model expressing domain relations among text, terms or concepts. In the following two sections we will address both problems.

2.4 The Domain Set

The problem of selecting an appropriate *domain set* is controversial. The particular choice of a domain set affects the way in which topic-proximity relations are set up, because it should be used to describe both semantic classes of texts and semantic classes of strongly related lexical concepts (i.e. domain concepts). An approximation of a lexical model for Semantic Domains can be easily obtained by clustering terms instead of concepts, assuming that most of the domain words are not ambiguous. At the text level Semantic Domains look like text archives, in which documents are categorized according to predefined taxonomies.

In this section, we discuss the problem of finding an adequate domain set, by proposing a set of "ideal" requirements to be satisfied by any domain set, aiming to reduce as much as possible the inherent level subjectivity required to perform this operation, while avoiding long-standing and useless ontological discussions. According to our experience, the following three criteria seem to be relevant to select an adequate set of domains:

Completeness: The domain set should be *complete*; i.e. all the possible texts/concepts that can be expressed in the language should be assigned to at least one domain.

Balancement: The domain set should be *balanced*; i.e. the number of text/concepts belonging to each domain should be uniformly distributed.

Separability: Semantic Domains should be *separable*, i.e. the same text/concept cannot be associated to more than one domain

The requirements stated below are formulated symmetrically at both the lexical and the text levels, imposing restrictions on the same domain set. This symmetrical view is intuitively reasonable. In fact, the larger the document collection, the larger its vocabulary. An unbalanced domain set at the text level will then reflect on an unbalanced domain set at the lexical level, and vice versa. The same holds for the separability requirement: if two domains

overlap at the textual level then their overlapping will be reflected at the lexical level. An analogous argument can be made regarding completeness.

Unfortunately the requirements stated below should be perceived as "ideal" conditions, that in practice cannot be fully satisfied. They are based on the assumption that the language can be analyzed and represented in its totality, while in practice, and probably even theoretically, it is not possible to accept such an assumption for several reasons. We try to list them below:

- It seems quite difficult to define a truly complete domain set (i.e. general enough to represent any possible aspect of human knowledge), because it is simply impossible to collect a corpus that contains a set of documents representing the whole of human activity.
- The balancement requirement cannot be formulated without any "a-priori" estimation of the relevance of each domain in the language. One possibility is to select the domain set in a way that the size of each domain-specific text collection is uniform. In this case the set of domains will be balanced with respect to the corpus, but what about the balancement of the corpus itself?
- A certain degree of domain overlapping seems to be inevitable, since many domains are very intimately related (e.g. texts belonging to Mathematics and Physics are often hard to distinguish for non-experts, even if most of them agree on separating the two domains).

The only way to escape from the problem of subjectivity in the selection of a domain set is to restrict our attention to both the lexicon and the texts contained in an available corpus, hoping that the distribution of the texts in it would reflect the "true" domain distribution we want to model. Even if from a theoretical point of view it is impossible to find a truly representative corpus, from an applicative point of view corpus-based approaches allows us to automatically infer the required domain distinctions, representing most of the relevant information required to perform the particular NLP task.

2.5 WordNet Domains

In this section we describe WordNet Domains,[5] an extension of WordNet [25], in which each synset is annotated with one or more domain labels.

The domain set of WordNet Domains is composed of about 200 domain labels, selected from a number of dictionaries and then structured in a taxonomy according to their position in the (much larger) Dewey Decimal Classification system (DDC), which is commonly used for classifying books in libraries. DDC was chosen because it ensures good coverage, is easily available and is commonly used to classify "text material" by librarians. Finally, it is

[5] Freely available for research from `http://wndomains.itc.it`.

officially documented and the interpretation of each domain is detailed in the reference manual [14].[6]

Table 2.1. WordNet Domains annotation for the senses of the noun "bank"

Sense	Synset and Gloss	Domains	Semcor
#1	depository financial institution, bank, banking concern, banking company (a financial institution...)	Economy	20
#2	bank (sloping land...)	Geography, Geology	14
#3	bank (a supply or stock held in reserve...)	Economy	
#4	bank, bank building (a building...)	Architecture, Economy	
#5	bank (an arrangement of similar objects...)	Factotum	1
#6	savings bank, coin bank, money box, bank (a container...)	Economy	
#7	bank (a long ridge or pile...)	Geography, Geology	2
#8	bank (the funds held by a gambling house...)	Economy, Play	
#9	bank, cant, camber (a slope in the turn of a road...)	Architecture	
#10	bank (a flight maneuver...)	Transport	

Domain labeling of synsets is complementary to the information already in WordNet. First, a domain may include synsets of different syntactic categories: for instance Medicine groups together senses of nouns, such as `doctor#1` and `hospital#1`, and from verbs, such as `operate#7`. Second, a domain may include senses from different WordNet sub-hierarchies (i.e derived from different "unique beginners" or from different "lexicographer files"[7]). For example, Sport contains senses such as `athlete#1`, derived from `life_form#1`, `game_equipment#1` from `physical_object#1`, `sport#1` from `act#2`, and `playing_field#1` from `location#1`.

The annotation methodology [56] was primarily manual and was based on lexicon-semantic criteria that take advantage of existing conceptual relations in WordNet. First, a small number of high level synsets were man-

[6] In a separate work [7] the requirements expressed in Sect. 2.4 were tested on the domain set provided by the first distribution of WordNet Domains, concluding that they have been partially respected. In the same paper a different taxonomy is proposed to alleviate some unbalancement problems that have been found in the previous version.

[7] The noun hierarchy is a tree forest, with several roots (*unique beginners*). The *lexicographer files* are the source files from which WordNet is "compiled". Each lexicographer file is usually related to a particular topic.

ually annotated with their pertinent domain. Then, an automatic procedure exploited some of the WordNet relations (i.e. hyponymy, troponymy, meronymy, antonymy and pertain-to) to extend the manual assignments to all the reachable synsets. For example, this procedure labeled the synset {beak, bill, neb, nib} with the code Zoology through inheritance from the synset {bird}, following a **part-of** relation. However, there are cases in which the inheritance procedure was blocked, by inserting "exceptions", to prevent incorrect propagation. For instance, barber_chair#1, being a **part-of** barbershop#1, which in turn is annotated with Commerce, would wrongly inherit the same domain. The entire process had cost approximately two person-years.

Domains may be used to group together senses of a particular word that have the same domain labels. Such grouping reduces the level of word ambiguity when disambiguating to a domain, as demonstrated in Table 2.1. The noun bank has ten different senses in WordNet 1.6: three of them (i.e. bank#1, bank#3 and bank#6) can be grouped under the Economy domain, while bank#2 and bank#7 belong to both Geography and Geology. Grouping related senses in order to achieve more "practical" coarse-grained senses is an emerging topic in WSD [71].

In our experiments, we adopted only the domain set reported in Table 2.2, relabeling each synset with the most specific ancestor in the WordNet Domains hierarchy included in this set. For example, Sport is used instead of Volley or Basketball, which are subsumed by Sport. This subset was selected empirically to allow a sensible level of abstraction without losing much relevant information, overcoming data sparseness for less frequent domains.

Some WordNet synsets do not belong to a specific domain but rather correspond to general language and may appear in any context. Such senses are tagged in WordNet Domains with a Factotum label, which may be considered as a "placeholder" for all other domains. Accordingly, Factotum is not one of the dimensions in our Domain Vectors (see Sect. 2.7), but is rather reflected as a property of those vectors which have a relatively uniform distribution across all domains.

2.6 Lexical Coherence: A Bridge from the Lexicon to the Texts

In this section we describe into detail the concept of *lexical coherence*, reporting a set of experiments we made to demonstrate this assumption. To perform our experiments we used the lexical resource WordNet Domains and a large scale sense tagged corpus of English texts: SemCor [51], the portion of the Brown corpus semantically annotated with WordNet senses.

The basic hypothesis of lexical coherence is that a great percentage of the concepts expressed in the same text belongs to the same domain. Lexical

Table 2.2. Domains distribution over WordNet synsets

Domain	#Syn	Domain	#Syn	Domain	#Syn
Factotum	36820	Biology	21281	Earth	4637
Psychology	3405	Architecture	3394	Medicine	3271
Economy	3039	Alimentation	2998	Administration	2975
Chemistry	2472	Transport	2443	Art	2365
Physics	2225	Sport	2105	Religion	2055
Linguistics	1771	Military	1491	Law	1340
History	1264	Industry	1103	Politics	1033
Play	1009	Anthropology	963	Fashion	937
Mathematics	861	Literature	822	Engineering	746
Sociology	679	Commerce	637	Pedagogy	612
Publishing	532	Tourism	511	Computer_Science	509
Telecommunication	493	Astronomy	477	Philosophy	381
Agriculture	334	Sexuality	272	Body_Care	185
Artisanship	149	Archaeology	141	Veterinary	92
Astrology	90				

coherence allows us to disambiguate ambiguous words, by associating domain-specific senses to them. Lexical coherence is then a basic property of most of the texts expressed in any natural language. Otherwise stated, words taken out of context show domain polysemy, but, when they occur into real texts, their polysemy is solved by the relations among their senses and the domain-specific concepts occurring in their contests.

Intuitively, texts may exhibit somewhat stronger or weaker orientation towards specific domains, but it seems less sensible to have a text that is not related to at least one domain. In other words, it is difficult to find a "generic" (Factotum) text. The same assumption is not valid for terms. In fact, the most frequent terms in the language, that constitute the greatest part of the tokens in texts, are generic terms, that are not associated to any domain.

This intuition is largely supported by our data: all the texts in SemCor exhibit concepts belonging to a small number of relevant domains, demonstrating the domain coherence of the lexical concepts expressed in the same text. In [59] a "one domain per discourse hypothesis" was proposed and verified on SemCor. This observation fits with the general lexical coherence assumption.

The availability of WordNet Domains makes it possible to analyze the content of a text in terms of domain information. Two related aspects will be addressed. Section 2.6 proposes a test to estimate the number of words in a text that brings relevant domain information. Section 2.6 reports on an experiment whose aim is to verify the "one domain per discourse" hypothesis. These experiments make use of the SemCor corpus.

We will show that the property of lexical coherence allows us to define corpus-based acquisition strategies for acquiring domain information, for example by detecting classes of related terms from classes of domain related

texts. On the other hand, lexical coherence allows us to identify classes of domain related texts starting from domain-specific terms. The consistency among the textual and the lexical representation of Semantic Domains allows us to define a "dual" *Domain Space*, in which terms, concepts and texts can be represented and compared.

Domain Words in Texts

The lexical coherence assumption claims that most of the concepts in texts belongs to the same domain. The experiment reported in this section aims to demonstrate that this assumption holds into real texts, by counting the percentage of words that actually share the same domain in them.

We observed that words in a text do not behave homogeneously as far as domain information is concerned. In particular, we have identified three classes of words:

- *Text Related Domain words (TRD)*: words that have at least one sense that contributes to determine the domain of the whole text; for instance, the word bank in a text concerning ECONOMY is likely to be a text related domain word.
- *Text Unrelated Domain words (TUD)*: words that have senses belonging to specific domains (i.e. they are non-generic words) but do not contribute to the domain of the text; for instance, the occurrence of church in a text about ECONOMY does not probably affect the whole topic of the text.
- *Text Unrelated Generic words (TUG)*: words that do not bring relevant domain information at all (i.e. the majority of their senses are annotated with FACTOTUM); for instance, a verb like to be is likely to fall in this class, whatever the domain of the whole text.

In order to provide a quantitative estimation of the distribution of the three word classes, an experiment has been carried out on the SemCor corpus using WORDNET DOMAINS as a repository for domain annotations. In the experiment we considered 42 domains labels (FACTOTUM was not included). For each text in SemCor, all the domains were scored according to their frequency among the senses of the words in the text. The three top scoring domains are considered as the prevalent domains in the text. These domains have been calculated for the whole text, without taking into account possible domain variations that can occur within portions of the text. Then each word of a text has been assigned to one of the three classes according to the fact that (i) at least one domain of the word is present in the three prevalent domains of the text (i.e. a TRD word); (ii) the majority of the senses of the word have a domain but none of them belongs to the top three of the text (i.e. a TUD word); (iii) the majority of the senses of the word are FACTOTUM and none of the other senses belongs to the top three domains of the text (i.e. a TUG word). Then each group of words has been further analyzed by part of speech and the average polysemy with respect of WORDNET has been calculated.

Table 2.3. Word distribution in SemCor according to the prevalent domains of the texts

Word class	Nouns	Verbs	Adjectives	Adverbs	All
TRD words	18,732 (34.5%)	2416 (8.7%)	1982 (9.6%)	436 (3.7%)	21%
Polysemy	3.90	9.55	4.17	1.62	4.46
TUD words	13,768 (25.3%)	2224 (8.1%)	815 (3.9%)	300 (2.5%)	15%
Polysemy	4.02	7.88	4.32	1.62	4.49
TUG words	21,902 (40.2%)	22,933 (83.2%)	17,987 (86.5%)	11,131 (93.8%)	64%
Polysemy	5.03	10.89	4.55	2.78	6.39

Results, reported in Table 2.3, show that a substantial quantity of words (21%) in texts actually carry domain information which is compatible with the prevalent domains of the whole text, with a significant (34.5%) contribution of nouns. TUG words (i.e. words whose senses are tagged with FACTOTUM) are, as expected, both the most frequent (i.e. 64%) and the most polysemous words in the text. This is especially true for verbs (83.2%), which often have generic meanings that do not contribute to determine the domain of the text. It is worthwhile to notice here that the percentage of TUD is lower than the percentage of TRD, even if it contain all the words belonging to the remaining 39 domains.

In summary, a great percentage of words inside texts tends to share the same domain, demonstrating lexical coherence. Coherence is higher for nouns, which constitute the largest part of the domain words in the lexicon.

One Domain per Discourse

The One Sense per Discourse (OSD) hypothesis puts forward the idea that there is a strong tendency for multiple uses of a word to share the same sense in a well-written discourse. Depending on the methodology used to calculate OSD, [26] claims that OSD is substantially verified (98%), while [49], using WORDNET as a sense repository, found that 33% of the words in SemCor have more than one sense within the same text, basically invalidating OSD.

Following the same line, a One Domain per Discourse (ODD) hypothesis would claim that multiple uses of a word in a coherent portion of text tend to share the same domain. If demonstrated, ODD would reinforce the main hypothesis of this work, i.e. that the prevalent domain of a text is an important feature for selecting the correct sense of the words in that text.

To support ODD an experiment has been carried out using WORDNET DOMAINS as a repository for domain information. We applied to domain labels the same methodology proposed by [49] to calculate sense variation: it is sufficient for just one occurrence of a word in the same text with different meanings to invalidate the OSD hypothesis. A set of 23,877 ambiguous words with multiple occurrences in the same document in SemCor was extracted and the number of words with multiple sense assignments was counted. Sem-

Table 2.4. One sense per discourse vs. one domain per discourse

Pos	Tokens	Exceptions to OSD	Exceptions to ODD
All	23,877	7469 (31%)	2466 (10%)
Nouns	10,291	2403 (23%)	1142 (11%)
Verbs	6658	3154 (47%)	916 (13%)
Adjectives	4495	1100 (24%)	391 (9%)
Adverbs	2336	790 (34%)	12 (1%)

cor senses for each word were mapped to their corresponding domains in WordNet Domains and for each occurrence of the word the intersection among domains was considered. To understand the difference between OSD and ODD, let us suppose that the word bank (see Table 2.1) occurs three times in the text with three different senses (e.g. `bank#1`, `bank#3`, `bank#8`). This case would invalidate OSD but would be consistent with ODD because the intersection among the corresponding domains is not empty (i.e. the domain Economy).

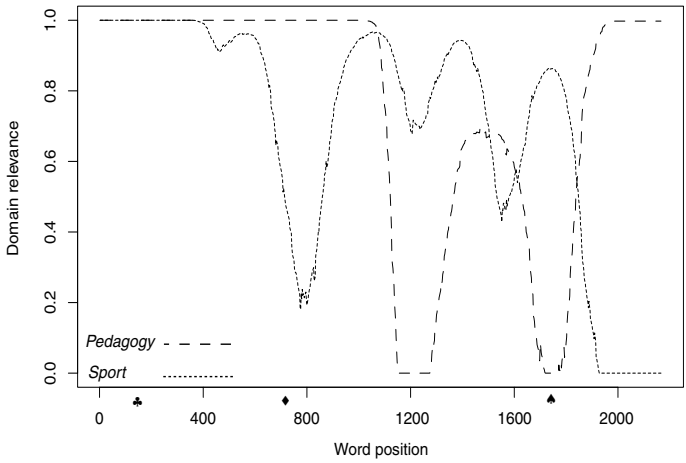
Results of the experiment, reported in Table 2.4, show that ODD is verified, corroborating the hypothesis that lexical coherence is an essential feature of texts (i.e. there are only a few relevant domains in a text). Exceptions to ODD (10% of word occurrences) might be due to *domain variations* within SemCor texts, which are quite long (about 2000 words). In these cases the same word can belong to different domains in different portions of the same text. Figure 2.2, generated after having disambiguated all the words in the text with respect to their possible domains, shows how the relevance of two domains, Pedagogy and Sport, varies through a single text. Domain relevance is defined in Sect. 3.1.

As a consequence, the idea of "relevant domain" actually makes sense within a portion of text (i.e. a context), rather than with respect to the whole text. This also affects WSD. Suppose, for instance, the word acrobatics (third sentence in Fig. 2.2) has to be disambiguated. It would seem reasonable to choose an appropriate sense considering the domain of a portion of text around the word, rather than relevant for the whole text. In the example the local relevant domain is Sport, which would correctly cause the selection of the first sense of acrobatics.

2.7 Computational Models for Semantic Domains

Any computational model for Semantic Domain is asked to represent the domain relations in at least one of the following (symmetric) levels.

Text level: Domains are represented by relations among texts.
Concept level: Domains are represented by relations among lexical concepts.
Term level: Domains are represented by relations among terms.



♣ ... The Russians are all trained as dancers before they start to study gymnastics. ...
♢ ... If we wait until children are in junior-high or high-school, we will never manage it. ...
♠ ... The backbend is of extreme importance to any form of free gymnastics, and, as with all acrobatics, the sooner begun the better the results. ...

Fig. 2.2. Domain variation in the text `br-e24` from the SemCor corpus

It is not necessary to explicitly define a domain model for all those levels, because they are symmetric. In fact it is possible to establish automatic procedures to transfer domain information from one to the other level, exploiting the lexical-coherence assumption. Below we report some attempts we found in the Computational Linguistics literature to represent Semantic Domains.

Concept Annotation

Semantic Domains can be described at a concept level by annotating lexical concepts into a lexical resource [56]. Many dictionaries, as for example LDOCE [76], indicate domain-specific usages by attaching Subject Field Codes to word senses. The domain annotation provides a natural way to group lexical concepts into semantic clusters, allowing us to reduce the granularity of the sense discrimination. In Sect. 2.5 we have described WordNet Domains, a large scale lexical resource in which lexical concepts are annotated by domain labels.

Text Annotation

Semantic Domains can be described at a text level by annotating texts according to a set of Semantic Domains or categories. This operation is implicit when annotated corpora are provided to train Text Categorization systems. Recently, a large scale corpus, annotated by adopting the domain set of WordNet Domains, is being created at ITC-irst, in the framework of the EU-funded MEANING project.[8] Its novelty consists in the fact that domain-representativeness has been chosen as the fundamental criterion for the selection of the texts to be included in the corpus. A core set of 42 basic domains,

[8] `http://www.lsi.upc.es/`*sim*`nlp/meaning/documentation/`.

broadly covering all the branches of knowledge, has been chosen to be represented in the corpus. Even if the corpus is not yet complete, it is the first lexical resource explicitly developed with the goal of studying the domain relations between the lexicon and texts.

Topic Signatures

The topic-specific context models (i.e. neighborhoods) as constructed by [35] can be viewed as signatures of the topic in question. They are sets of words that can be used to identify the topic (i.e. the domain, in our terminology) in which the described linguistic entity is typically located.

However, a topic signature can be constructed even without the use of subject codes by generating it (semi-)automatically from a lexical resource and then validating it on topic-specific corpora [38]. An extension of this idea is to construct "topics" around individual senses of a word by automatically retrieving a number of documents corresponding to this sense. The collected documents then represent a 'topic out of which a topic signature may be extracted, which in turn corresponds directly to the initial word sense under investigation. This approach has been adopted in [1].

Topic signatures for sense can be perceived as a computational model for Semantic Domains, because they relate senses co-occurring with a set of lexically coherent terms. Topic signatures allows us to detect domain relations among concepts, avoiding taking any a-priori decision about a set of relevant domains. In addition topic signatures provide a viable way to relate lexical concepts to texts, as required for any computational model for Semantic Domain.

Finally, topic signatures can be associated to texts and terms, adopting similar strategies, allowing us to compare those different objects, so to transfer domain information from one level to the other.

Domain Vectors

Semantic Domains can be used to define a vectorial space, namely the *Domain Space* (see Sect. 3.3), in which terms, texts and concepts can be represented together. Each domain is represented by a different dimension, and any linguistic entity is represented by means of Domain Vectors (DVs) defined in this space. The value of each component of a DV is the domain relevance (see Sect. 3.1) estimated between the object and the corresponding domain.

Typically, DVs related to generic senses (namely FACTOTUM concepts) have a flat distribution, while DVs for domain-specific senses are strongly oriented along one dimension. As is common for vector representations, DVs enable us to compute domain similarity between objects of either the same or different types using the same similarity metric, defined in a common vectorial space. This property suggests the potential of utilizing domain similarity

between various types of objects for different NLP tasks. For example, measuring the similarity between the DV of a word context and the DVs of its alternative senses is useful for WSD.