Chapter 9
Lexicography, terminology and ontologies

Serge Sharoff, Anthony Hartley

Published as
Sharoff, S., & Hartley, A. (2012). Lexicography, terminology and ontologies. *Handbook of Technical Communication (HAL 8)*, 317-346.

1. Words and concepts

A text – for example, this chapter – consists of sequences of characters usually separated by punctuation marks and white spaces. Humans reading this text interpret such sequences as words having particular meanings. The relationship between words and objects, both physical and abstract, can be illustrated by the semiotic triangle (Ogden and Richards, 1923), which was first introduced independently by Charles Sanders Pierce and Gottlob Frege at the end of the 19th century and later popularised by Ogden and Richards (see Figure 1). The semiotic triangle has three vertices corresponding to the word, its interpretation (concept or meaning) and the physical or abstract object it refers to (referent). The line between word and object is dotted, since a linguistic expression does not point to an object directly, but only via its interpretation. The word may be a single-word or a multi-word expression. The denotational form of concepts can change over time (e.g., *on line, on-line, online*) and is subject to the rules of word formation of a particular language (e.g., *watershed, ligne de partage des eaux*) without affecting its relationship to the referent. Reference, then, is a function that identifies objects in a domain. This model is clear and useful for many applications within the scope of this chapter. In particular, it helps us distinguish two perspectives for studying the relationship between words and meanings: the *semasiological* perspective starts from the linguistic form of the sign to ask how it is understood, while the *onomasiological* perspective starts from the content of the sign to ask how it is delimited and named.

The realm of ontology construction is the system of concepts itself and its relationship to objects; the emphasis is onomasiological. Terminologists are concerned with what Sager calls ‘special languages’: “semi-autonomous,

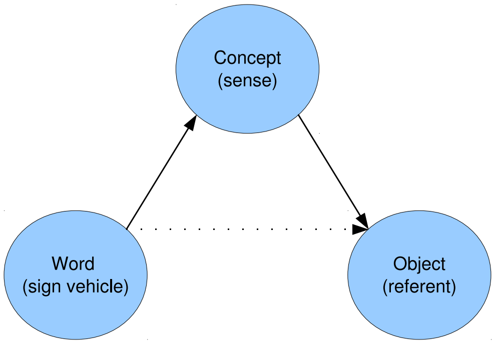


Figure 1. Semiotic triangle

complex semiotic systems [. . .] derived from general language [whose] use [. . .] is restricted to specialists for conceptualisation, classification and communication in the same or closely related fields" (Sager, 1993, p. 44). Thus the *terminology* of a discipline is an instrument of 'special reference'. While the traditional approach to terminological activity adopts the onomasiologial perspective, this view that concepts exist objectively prior to being named is challenged by advocates of a sociocognitive approach, e.g., (Kageura, 1995; Temmermann, 2000), who argue for equal emphasis on the semasiological perspective to uncover real usage in the 'parole' of a subject community. Lexicographers are concerned not with specialised areas of knowledge but with general knowledge, that is, with the *vocabulary* of 'general reference'; they take the semasiological perspective. In spite of these differences between the goals and methods of ontologists, terminologists and lexicographers, their activities share a common underlying task: to extract meanings from texts and represent them in a form appropriate to their end use.

It is tempting to view the set of interpretations of signs as a finite list, so that we could imagine enumerating the complete set of concepts existing in a given domain and the set of words or terms referring to them. However, when we apply the semiotic triangle to the use of words in the real world, many distinctions become blurred. Let us take an example from computer science of a number of 'mechanisms' with extensive abilities for processing text data. A webpage in HTML is not normally considered to be written in a *programming language*, yet a webpage can contain Javascript, which is normally referred to as a programming language. Again, while XSLT (eXtensible Stylesheet Language Transformations) is not commonly thought of as a programming language, a reference of this sort is possible in some contexts. On the other hand, CSS (Cascading Style Sheets) and LaTeX are not considered as programming languages at all. In other words, in many contexts it is difficult to determine the identity of a concept and the set of words that can be used to refer to it.

An aspect of representing the complex relationship between concepts and words concerns the need to accommodate the expectations and limitations of different audiences. For instance, the entries for *genome* in the Oxford Advanced Learners Dictionary (OALD) and the Concise Oxford English Dictionary (COED) define the concept using different words:

(1) **OALD** *the complete set of GENES in a cell or living thing: the human genome.*

(2) **COED** *1 the haploid set of chromosomes of an organism. 2 the complete set of genetic material of an organism.*

Although both definitions are intended to identify the same referent, the wordings and therefore contents are different, because the OALD is designed for non-native speakers of English while the COED aims to provide more comprehensive definitions.

Different language communities or communities of practice may operate with different systems of concepts, which makes even more difficult the task of determining the reference of a sign and its relationship to signs in another language. For example, Gottlob Frege described the relation between signs and meanings as:

(3) Es liegt nun nahe, mit einem Zeichen (Namen, Wortverbindung, Schriftzeichen) außer dem Bezeichneten, was die *Bedeutung* des Zeichens heißen möge, noch das verbunden zu denken, was ich den *Sinn* des Zeichens nennen möchte, worin die Art des Gegebenseins enthalten ist. (G.Frege, Sinn und Bedeutung)

(4) It is natural, now, to think of there being connected with a sign (name, combination of words, letters), besides that to which the sign refers, which may be called the *reference* of the sign, also what I should

> like to call the *sense* of the sign, wherein the mode of presentation is contained. (G.Frege, On Sense and Reference, translated by M.Black)

In their general reference occurrences, the two words italicised in the original German text (*Sinn* and *Bedeutung*) are both normally translated into English as *meaning*. However, the distinction drawn by Frege facilitated the development of a special reference in English which had not existed before Frege's text became influential. This illustrates how systems of terminology evolve over time, not only because of the invention of new artefacts and new scientific discoveries, but also because of the need to make distinctions important for a particular theory. Given that such development happens through the medium of language, language has a direct impact on the system of concepts, as well as on the context of their use. (Temmermann, 2000, p. 33) stresses the importance for terminologists of the semasiological perspective: "In a scientific community meaning is negotiated. A concept is not really recognised as such nor taken seriously unless it is named by a term." In linguistic research this is known as logogenesis, i.e., meaning making in the process of development of a discipline, a research area, or an individual text. Halliday and Matthiessen metaphorically refer to this fact as "there are no 'naked ideas' lurking in the background waiting to be clothed; it is language that *creates* meaning" (Halliday and Matthiessen, 1999, p. 602).

One important aspect of logogenesis concerns ways of naming new artefacts and constructs of human cognition, including:

- metaphoric extension of existing terms (*file* on a computer, *rib* of a tyre, *sense* and *reference* in semantics);
- borrowing from other languages (*ketchup, poltergeist, sputnik*);
- neoclassical creation (*vulcanisation, oscillograph, carcinoma, bronchitis, biotechnology*);
- prefixing (*pre-production, supraorbital, thermocouple*);
- compounding (*flywheel, sprocket wheel, fluid power transmission*).

The link between the world and words is further complicated by lexical expression of some grammatical categories, for example, definiteness, which in English or German is expressed by articles (*a, the*), which do not exist in

Chinese or Russian. Even if articles are used in a language, their use varies from one language to another. For instance, 'abstract' nouns in French and German take a definite article (e.g., *la paix est ...; der Frieden*), in English they do not (*peace is* ...).

Each domain has its own terminology (concepts and lexicalisations) and ways for using it. For example, the language of weather forecasts is considerably different from the language of atmospheric physics, even if their subject matter is, broadly speaking, the same. There is a long tradition of terminological studies of specific disciplines, reflected in, e.g., (Wüster, 1955). With respect to computational applications, especially in machine translation, this awareness gave rise to the concept of 'sublanguages', which highlights the importance of circumscribing the domain of terminographic or lexicographic research (Kittredge and Lehrberger, 1982; Grishman and Kittredge, 1986).

In the following sections we will present the three perspectives on the semiotic triangle with respect to ontologies (Section 2.1), terminology (Section 2.2) and lexicography (Section 2.4), followed by sub-sections devoted to methods for populating ontologies, termbanks and dictionaries (Section 3). After that we will consider their applications (Section 4), emphasising the use of ontologies and lexical resources in the creation and translation of documents.

2. Mapping between words, concepts and reality

2.1. Ontologies

The term 'ontology' originates from classic philosophy, where it refers to theory for describing objects of the universe and their properties (*onto-* comes from the Greek word for *being*). Aristotle was one of the first philosophers to develop a system of universal categories for describing the world in terms of entities and attributes predicated to them. The debate about the nature of the system of categories was one of the main subjects of philosophical debate throughout centuries, which contributed to knowledge representation techniques (Dreyfus, 1982; Sharoff, 2005b), also for information on the history of ontological debates in philosophy check `http://www.ontology.co`.

The need to input information into the computer and to perform reasoning on its basis gave ontological studies a more technological dimension, resulting in various approaches to knowledge representation. The bulk of commu-

Unstructured	**c1390** (?c1350) *Joseph of Arimathie* (1871) l. 452, A-non tholomers men woxen e biggore; sone beeren hem a-bac and brouhten hem to grounde.
Hierarchy	$\begin{array}{c} Example \\ \left[\begin{array}{ll} Created: & Date \\ & \left[\begin{array}{l} Year: 1390 \\ Range: c \end{array}\right] \\ Created?: & Date \\ & \left[\begin{array}{l} Year: 1350 \\ Range: c \end{array}\right] \\ Biblio: & PubStatement \\ & \left[\begin{array}{l} Source: \text{Joseph of Arimathie} \\ Publication: Date[Year: 1871] \\ Line: 452 \end{array}\right] \\ Eg: & Text \\ & \left[\begin{array}{l} Q: \text{A-non tholomers men woxen} \ldots. \\ Language: \text{Middle English} \end{array}\right] \end{array}\right] \end{array}$
Triples	$IS-A(Example,\text{Arimathie1871-big})$ $IS-A(Date,\text{Year-1390c})$ $Year(\text{Year-1390c}, 1390)$ $Range(\text{Year-1390c,'circa-date'})$ $Created(\text{Arimathie1871-big}, \text{Year-1390c}) \ldots$

Figure 2. Ontological analysis of an example from OED

nication between humans is not explicit enough for knowledge representation purposes. Figure 2 gives an example of the difference in the level of granularity needed for representing knowledge for the human and for the computer. The Oxford English Dictionary (OED) records one of the examples of uses of the word *big* from the 14^{th} century. The example quoted might be useful for the human reader, but for the computer this representation does not render the properties of the quote explicitly. Formally, the quote can be represented as a member of the class *Example*, which has such attributes as *Created, Biblio* or *Eg* filled with values, which, in turn, can be represented by the classes *Date, PubStatement* and *Text*. This more explicit ontological specification can be extended even further, e.g., by representing the confidence in knowledge about the two provisional dates for the quote, using greater granularity of information about the language variety (by adding explicit information that Middle English is a subtype of English), as well as representing the actual range of years for each provisional date (in Figure 2 *c* means *circa*; the actual meaning of *circa* has to be defined formally).

Even though the hierarchical representation of concepts is intuitively appealing, it is not sufficient in many cases. For instance, a book can be represented as consisting of pages, paragraphs and chunks of argumentation, each of which can overlap, e.g., a paragraph can be shared between two pages, or the same argument can be mentioned in two separate paragraphs. A more general (but more verbose) way of representing concepts is by using triples containing a relationship and two objects. The same information from the OED example can be represented by triples, specifying the class of each object or concept (an object has to be identified by a unique label, e.g., 'Arimathie1871-big' for this quotation), and relations, which join the attributes of labelled concepts with their values. Relations can also connect concepts. Formally, the representation by triples is equivalent to a graph, i.e., a network of labelled nodes connected by directed arcs (relations). Terms like 'semantic networks' or 'conceptual graphs' are used to refer to representations of this sort (Sowa, 2000).

There are two relations universally used in ontology engineering:

IS-A defining the class-instance relationship (*Bucephalus is a horse*) or the subclass-class relationship (*horses are equestrians; equestrians are mammals*). In some implementations, a separate relation Instance-Of is used in the first case. The subclass-class subtype of this relation is often referred to as hyponymy (the class *horse* is a hyponym of the class *equestrian*, while *mammal* is a hypernym of *equestrian*). Some implementations of the IS-A relationship allow multiple inheritance, when an object can be an instance of several classes defined with different sets of properties, e.g., this book is an information container, which consists of separate chapters, while at the same time, it can be considered as a physical object, which has its size, weight, etc.

HAS-A defining the relationship of containment of one object in another one, e.g., a book can contain chapters (the converse relationship to HAS-A is PART-OF). This relation is often referred to as meronymy (*chapter* is a meronym of *book*).

Any developed ontology defines concepts and relations specific to its domain (e.g., *Created* or *PubStatement* in the example above). More general (Upper) ontologies are designed to provide basic concepts and relations covering a wide range of possible domains, e.g., Cyc (Lenat and Gupta, 1990), Penman Upper Model (Bateman, 1990), or SUMO, Suggested Upper Merged Ontology (Niles and Pease, 2001). Cyc was a very ambitious project designed to represent the whole body of common-sense knowledge in a machine-tractable form, e.g., "Every tree is a plant" and "Plants die eventually". The Penman Upper Model was designed as a tool supporting various projects in multilingual generation, so it included such concepts as IdeaReporting (*Many Americans expressed concerns that . . .*) or Purpose (*Press Ctrl-S to save your changes*) for connecting respective facts. SUMO, on the other hand, was designed with knowledge engineering projects in mind, having sub-ontologies for various domains, e.g., for finances, military devices or viruses.

Often an ontology specification mechanism contains tools for providing information about default values of some attributes (e.g., a horse has 205 bones), facts (an animal consumes food), inference rules (if an animal is deprived of food for a long time, it can die), and events updating the set of facts (e.g., Bucephalus died in 326 BC, thus it was not true that Bucephalus was alive in the model of 327 BC).

The formal language for representing ontologies is sometimes based on the idea of a hierarchy, a network or (more typically) a combination of the two. For instance, RDF (Resource Description Framework), a popular representation framework for Web resources defined by the W3C consortium (Beckett, 2004), uses triples to make statements like those presented in the triples part of Figure 2. RDF is an abstract model whose actual representation is based on *serialisation*, i.e., ways for encoding graphs as a sequence of statements in a file. For example, information about a publication, like the RDF specification (Beckett, 2004), can be encoded in an XML form using the ontology based on the Dublin Core (Hillmann, 2005):

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
       xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description
       rdf:about="http://www.w3.org/TR/rdf-syntax-grammar/">
     <dc:title>RDF specification</dc:title>
     <dc:publisher>W3C</dc:publisher>
  </rdf:Description>
</rdf:RDF>
```

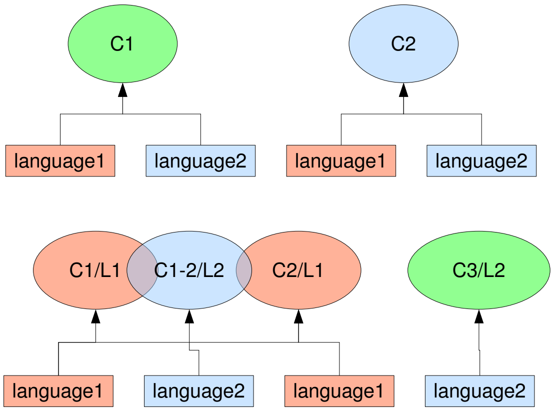


Figure 3. Mapping terms to concepts

RDF is a generic framework which does not specify many parameters needed for elaborate ontologies, but it led to development of Web Ontology Language (OWL, note the order of characters in the acronym), specifically designed for representing ontologies (Smith et al., 2004). Objects and relations from Figure 2 can be defined in OWL in the following way:

```
<Date rdf:ID="Year-1390c">
  <yearValue rdf:datatype="&xsd;positiveInteger">1390</yearValue>
  <rangeValue rdf:resource="circa-date"/>
</Date>
<Example rdf:ID="Arimathie1871-big"
  <Created rdf:resource="#Year-1390c"/>
  <Biblio  ... />
</Example>
```

In this notation the date of citation is defined as an identifiable resource consisting of a positive integer value with a range approximation, while the example makes an explicit reference to this resource.

2.2. Terminology

A termbank provides a link between concepts in an ontology and their names (*terms*), typically in several different languages. Thinking within ISO/TC 37 “Terminology and other language and content resources” on the nature of this link has followed the principles of the Vienna school of terminology founded by Eugen Wüster (Wüster, 1979). According to these principles, ideally concepts are defined by being assigned a place in the concept system before they are designated (the onomasiological perspective). Moreover, each concept should be designated by only one term (mononomy) and each term should refer to only one concept (monosemy) – the principle of *univocity*. This ideal case based on a one-to-one mapping is illustrated in the top part of Figure 3.

However, it is quite common that one designation can refer to several different objects even in the same domain, e.g., *terminology* taken as a term of its own can mean a collection of terms (= termbank), a set of practices for creating termbanks (= terminography, akin to lexicography), and also to a theory of the relationship between concepts and terms (akin to lexicology). Alternatively, several terms can be used to refer to the same object in slightly different contexts, e.g., *catalog, directory* and *folder* can all refer to a group of files in a file system.

Temmerman argues very persuasively that polysemy and synonymy of this sort are not only pervasive but also functional, since they enable specialised communities of practice to negotiate and refine their understanding of their shared knowledge space over time (Temmermann, 2000). This supposes a model of knowledge as a multidimensional space in which “the value of a concept with respect to a given axis is generally defined as a range and only exceptionally [. . .] as a point” (Sager, 1990, pp. 15-16). Temmerman proposes that, since many concepts are not clear-cut, they are better considered as categories showing different degrees of prototypicality and possibly overlapping with others.

In the multilingual context it is possible that one language (L2) has a word that refers to a object (C3) that is not lexicalised in another language (L1), e.g., *deux-roues* in French refers to the set of cycles, scooters and motorbikes in English (the bottom part of Figure 3). It is also possible that a word in L1 can refer to two separate objects, e.g., *Wortverbindung* in the quote from Frege (3) refers to single-word compounds as well as multiword expressions, while its English translation (4) only to *combination of words*. Similarly *Gegebensein* in (3) is a very precise description to a philosophical concept in German, whereas the word *presentation* in (4) does not function as a philosophical term in English. It is used as a term in special reference in religion and obstetrics, but in (4) only in general reference.

In some cases a term in one language functions as a lexicalised hypernym for a subset of concepts but does not have an equivalent in another language at the same point in the hierarchy. For instance, the French terms *bicyclette, cyclomoteur, véhicule* all have English equivalents, *bicycle, moped* and *vehicle*, respectively. However, while *deux-roues* covers all types of two-wheeled vehicles (bicycles, mopeds, scooters, etc) the English term *two-wheeler* is generally taken as referring to bicycles rather than to *motorised/powered two-wheelers* also.

A considerable part of the activity of a terminologist consists of organising the system of concepts, so in this respect it is similar to the work of an ontologist, except that it also requires assigning names to the concepts in a variety of languages (Sager, 1990; Wright and Budin, 2001). This organisation can often involve standardisation of the delimitation of particular concepts and their designation; this standardisation may be imposed locally, as part of the house style of a particular company, or nationally, regionally or internationally by standardisation bodies such as ANSI, CEN and ISO. One common facet of standardisation is the planned creation of neologisms, most often in response to massive borrowing from dominant languages, such as English; L'Office québécois de la langue française has undertaken probably the most concerted effort ever in this field. Such (legitimate) prescriptive activities distinguish terminography from contemporary lexicography, which aims to be descriptive of usage, although historically the dictionaries of language academies also set out to fix use.

However, in addition to terms themselves and their links to concepts, a termbank can contain:

- typical collocations, e.g., *open, save, delete a file* vs. *ouvrir, sauvegarder, supprimer un fichier*, to facilitate their coherent use and translation;

- boilerplate, i.e., fixed expressions, which are expected in a specific domain or specific document type, e.g., the copyright statement attached to Wikipedia articles *All text is available under the terms of the GNU Free Documentation License*;

- common abbreviations of multiword terms, e.g., *CALL* from *computer-assisted language learning*, *ABS* from *antilock brake system*;
- proper names, especially if the tradition of their spelling differs across languages, e.g., *München, Munich* or *Monaco* in German, English and Italian respectively;
- deprecated and preferred terms, e.g., *flame resistant* and *flame retardant*, respectively, in the "ISO Glossary of Fire Terms and Definitions".

2.3. Formats for storing terminology

Individuals and organisations commonly make use of a wide range of more or less sophisticated formats for storing and presenting terminology, from termbanks with rich and highly-structured data to simple flat bilingual lists of equivalent expressions managed in a spreadsheet. If we consider the sophisticated end of the spectrum, instantiated by the EU inter-institutional terminology database IATE [1] and other termbanks accessible via the International Telecommunications Union website [2], we can expect to see represented most if not all the following items:

- links to concepts in an ontology, primarily by means of intensional definitions specifying the characteristics of the concept and the nearest genus, additionally with notes on the scope of usage;
- links to hypernyms, hyponyms and co-hyponyms;
- domain(s) in which a term is used;
- terms themselves and their variant names, abbreviations;
- indicators that the term is standardised, or preferred or deprecated;
- reliability ratings;
- morphological information (gender, number, declension);
- examples of their use;
- links to other terms related by their form, e.g., *poverty, income poverty, Inter-agency Working Group on Poverty Elimination*;

- housekeeping information.

In the case of multilingual termbanks some information is shared between all terms referring to the same concept, while most information is language-specific.

In designing a termbank it is important to distinguish between the onomasiological and semasiological perspectives (Gibbon et al., 1997). With the first, concept-oriented perspective a given entry, with equivalent terms in multiple languages, corresponds to a single, language-independent concept; the second is able to represent polysemy. The two perspectives are complementary and can be accommodated in a single representation format, but the design of the termbase will differ depending on the preference for either perspective.

The late 1990's and early 2000's saw a flurry of activity, much of it associated with ISO/TC 37, aimed at designing standards for exchanging structured terminological data for the benefits of both human translation and the gamut of machine applications ranging from term extraction and ontology construction to IR and MT. They can be traced back to the Text Encoding Initiative (TEI) (Sperberg-McQueen and Burnard, 2002) and include the SGML-based MARTIF (Machine-Readable Terminology Interchange Format) and OLIF (Open Lexicon Interchange Format) developed for six European languages by a consortium of major translation and publishing companies. Currently, the leading contender is TBX (Term Base eXchange) [3], an XML application whose core structure is based on MARTIF. It is concept-oriented. TBX is sponsored by LISA (Localization Industry Standards Association) and published as ISO 30042.

2.4. Lexicography

Lexicography is one of the oldest human activities since the invention of writing. The need to catalogue a language, analyse obscure sources or communicate with other cultures led to production of word lists accompanied with their definitions, commentaries or translations. One of the earliest dictionaries are available in the form of Sumerian cuneiform tablets listing words of a foreign language (Akkadian) or a glossary of herbs for medicinal use (Hartmann and James, 2001).

In English language lexicography, the rise in the importance of dictionaries coincides with the spread of literacy and the appearance of a large number

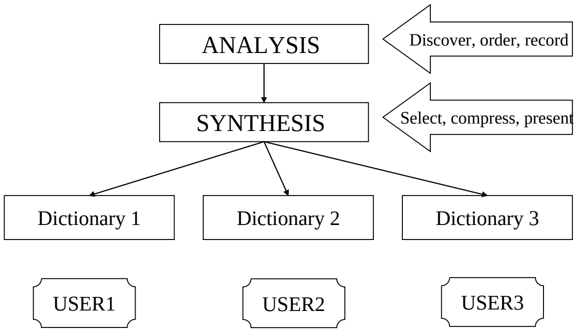


Figure 4. Two stages of dictionary development

of printed sources in the seventeenth century: newly educated readers needed an authoritative source to guide them in understanding rare words and resolving their disputes about word meanings. Dictionaries, such as Johnson's 1755 *Dictionary of the English Language*, appeared to satisfy this demand (Kilgarriff, 1997).

A more recent rise in dictionary making in the twentieth century coincides with economic globalisation, which has resulted in the need to use a variety of bilingual dictionaries for the tasks of translation and international communication. Moreover, globalisation has increased the importance of dictionaries for language learners, primarily learners of English, as English had established itself as the world's global language by the second half of the twentieth century.

Traditionally dictionaries were developed using a large collection of slip cards for individual words containing citations, which illustrated their use. At a later stage, these groups of examples were used to define senses. Often the decisions were made on the basis of introspection. For instance, a native English speaker might suggest that the noun *landing* can mean either an aircraft coming to earth or an intermediate platform in a staircase, and can invent examples of their use, e.g., *it was a bumpy landing*.

Development of corpora in the last quarter of the twentieth century has changed the situation considerably: lexicographic evidence in modern dictionaries is mostly based on analysis of concordance lines taken from representative corpora (McEnery and Wilson, 1999). The first dictionary created from scratch on the basis of corpus evidence was COBUILD, developed in a joint project between Collins and Birmingham University (Sinclair, 1987).

The frequency of words in a language follows a ‘long-tail’ distribution, also known as Zipf’s law (Zipf, 1935): very few words are extremely frequent (*the, of, have*), while the frequency of a very large number of words is fairly modest. For instance, in one million words of the Brown Corpus (Kučera and Francis, 1967) there are 23 examples of *landing* and only two occurrences of the verb *to stab*, none of which refers to the metaphorical sense (*to stab somebody in the back*, ‘to betray someone’), while in 100 million words of the BNC (Aston and Burnard, 1998) there are 1,049 occurrences of *stab*, with at least 31 examples of its metaphoric use.

This suggests that corpora used for producing dictionaries for general rather than special reference purposes have to be big enough to cover the range of expressions possible with each individual word (the BNC, Bank of English or Oxford English Corpus used for this task measure in hundreds of millions of words). However, going through thousands of examples of *landing* or *stab* in a corpus like the BNC is a very time consuming process, which can be facilitated by tools designed for extracting and condensing lexicographic information, such as SketchEngine (Kilgarriff, 2003). Such tools can identify words more commonly modifying the target, e.g., *forced, Normandy, first-floor, emergency, downwind, safe* as modifiers for *landing*, and *gear, strip, craft, stage* as words modified by it. These two lists can suggests adding other senses which might have been missed in introspection, e.g., *landing* in the sense of a military operation.

It is quite common that dictionaries are developed through two stages: through analysis of language data and then its synthesis to present facts to the users (Figure 4). In the first stage lexicographers analyse the behaviour of a word, create sense distinctions, determine its contexts of use and select a range of examples for each context. This results in a systematic overview of facts about the word, which are stored in a database. In the second stage this information is used for producing a number of actual dictionaries. Each dictionary in this case can contain a subset of words recorded in the database. Some senses can be discarded (as less frequent or less relevant to the anticipated users) or collapsed, joining several senses together. Alternatively extra senses can be distinguishes depending on the needs of the users. Definitions can be adapted to the level expected by the user, and translations into a specific language can be defined.

The issue of presentation is important in termbanks, but for dictionaries the way information is delivered to their users is crucial. The same word in the same language can be described in different ways depending on the purpose of a dictionary. A dictionary can be aimed at being a guide to understanding a word in as many contexts of its use as possible. Alternatively, a dictionary can be used as a guide for active production in a foreign language, which implies offering greater guidance to the user about more typical constructions of its use (Rundell, 1999). A dictionary aimed as being a guide for translation can have more senses than reasonable for the same word in a monolingual dictionary, since the target language may make more sense distinctions, so justifying several translations from the source language.

In Figure 5 we list examples of entries for the same word from four dictionaries: LDOCE (LDOCE, 1995), a dictionary for non-native speakers, OED (OED, 1989), a comprehensive dictionary recording historical information about word for researchers and native speakers (only two senses out of 19 are shown), WordNet (Miller, 1990), a concept-oriented thesaurus, and Oxford-Hachette French Dictionary (OFD, 1994), which is designed as a guide for translators. The bilingual dictionary makes distinctions between two senses of landing from a boat (for people and for cargo), which are not made in the two monolingual dictionaries. The definitions used in LDOCE are considerably simpler and it contains more information to guide the users.

A dictionary entry can contain various types of information:

- the headword and its variations (*favour, favor*; *ping pong, table tennis*);
- hyphenation and pronunciation (using different notations to indicate them);
- inflection (especially when it is irregular, e.g., *take, took, taken*);
- sense distinctions (they can be presented as a flat list or as a structured list with several levels of subsenses, like in the OED example; in other dictionaries distinctions between subsenses can be made implicitly, e.g., *crash landing* in LDOCE);
- basic morphological categories (this can affect the structure of the entry, as some dictionaries list all morphological classes as senses in the same entry, while others treat them as different headwords; some grammatical information can be confined to a specific sense, e.g., [C] in the

LDOCE example indicates that *landing* used in Senses 2 and 3 is a countable noun);

- definitions or translations of senses (a bilingual dictionary can also indicate the nature of senses by conditions on their use, e.g., *at turn of stairs, by plane*, as well as properties of translations, e.g., their grammatical gender or translation of collocation patterns *landing on→atterissage sur*);
- cross-references to other entries (references to synonyms or antonyms, as well as generic 'see' links);
- more or less fixed constructions (in some dictionaries they are presented as separate headwords, but most of the time they are incorporated into respective entries, e.g., Sense 8 in the full entry for *landing* in the OED lists 33 compounds);
- examples (they can also contain full bibliographic information, like in the OED example; they are normally associated with senses, but can also act as indications of subsenses, e.g., *Apollo moon landings* in the LDOCE example)
- various constraints on usage, such as the domain, register, region or period of use, e.g., *Aeronaut., slang, British, obsolete*;

The concept of a dictionary is based on a list of headwords, each of which is taken as a separate entity. On the other hand, the goal of a dictionary is to help its users use words or understand them in context. This often leads to introduction of information about word uses into its entry. One way of doing this is based on definitions. For instance, one sense of *leave* in COBUILD is defined as *If you leave someone to do something, you go away from them, so that they do it on their own*, which intends to show the context of its use (e.g., you=a human actor) and possible intentions of the speaker. However, this information is not encoded in any explicit way. One possibility of formalising the context of use is by using systemic networks, which can link meaning-making intentions to lexical realisation options (Sharoff, 2005a). Some researchers have proposed a very elaborate mechanism for representing lexical co-occurrence links between words. For instance, Mel'čuk proposed the notion of lexical functions as generalisations of lexical co-occurrences and designed their typology (Mel'čuk, 1996). Thus, expressions like *heavy* *rain, strong desire, strict control, sleep firmly* imply the high degree of a quality, so such collocations can be captured by a lexical function *Magn*, such that $Magn(smoker) = heavy; Magn(Raucher) = stark; Magn(fumeur) = grand$

3. Populating ontologies, termbanks and dictionaries

3.1. Corpora

As we stated earlier, ontologists deal with concepts, which are manifested in texts. The link between texts and the work of terminologists and lexicographers is even more direct. Hence, they need corpora, large collections of texts representative of a chosen domain and/or genre. Raw texts in corpora are usually accompanied with some kind of annotation, such as text metadata (provenance, text type, intended audiences, etc), linguistic annotation (parts of speech, lemmas, syntactic structures, etc) or alignment with corresponding translated texts (McEnery and Wilson, 1999).

The traditional way of creating corpora (still in use in some situations, especially for lexicography and linguistic research) is based on collecting texts manually on a case-by-case basis. A classic example is the development of the Bank of English (Sinclair, 1987) in the 1980s from scanned and OCR'ed printed texts. The growth of the amount of texts available in electronic form, often over the Internet, has led to a proliferation of corpora in a variety of domains and languages, often resulting in the possibility of using 'disposable corpora' (Varantola, 2003), i.e., corpora created for a particular task and discarded once the task has been fulfilled.

There are several methods for creating disposable corpora from the web. One is based on "focused crawling", which, in its simplest form, involves selecting several websites containing a large number of texts in the target domain and retrieving the entire set of these texts. Given that Wikipedia dumps in a variety of languages are available for download, a version of focused crawling can be based on retrieval of Wikipedias for several language and clustering them into subdomains on the basis of explicit metadata available in Wikipedia (categories, templates and links between pages). One problem with this simple approach to crawling is that only a small subset of relevant data can be retrieved, which might bias the corpus and the results of the ontological, terminological or lexicographic study.

More advanced methods of focused crawling involve starting with a seed set of links and then collecting links to other relevant websites, with the relevance assessed by keywords and/or hypertext links between pages, as more relevant pages tend to have more inter-connections with each other (Chakrabarti et al., 1999).

Yet another method for corpus collection relies on making automated queries to a major search engine, using words defining a domain. Since the initial set of seed queries is likely to be incomplete, corpus collection can include a bootstrapping phase: an initial corpus is created from seed queries, terms are automatically extracted from it and then fed to the search engine to collect a bigger corpus (Baroni and Bernardini, 2004). For instance, the seed query set can include words like *dissociative, epilepsy, pseudo-seizures*. This leads to a corpus yielding new keywords like *amnesia, convulsions, paroxysmal*, which can be used in the second iteration. It is also possible to collect a large (100-200 million words) corpus of general language using this method, if the queries consist of words from the general lexicon (Sharoff, 2006). Finally, keyword-based corpus development can be combined with focused crawling to get a wide variety of seed websites (Stamatakis et al., 2003).

For multilingual applications, parallel corpora containing original texts and their translations can provide useful data for developing bilingual dictionaries and termbanks. Traditionally, again, parallel corpora were created by painstaking collection and alignment of published translations. Nowadays large amount of texts are available on the Web, either in institutional repositories, e.g., United Nations (Eisele and Chen, 2010), European Parliament (Koehn and Knight, 2002), or in multilingual webpages, e.g., newswire stories and their translations in different languages. Resnik and his colleagues developed a method for harvesting webpages that are structurally similar, cross-linked and written in different languages (Resnik and Smith, 2003), while lexical-based methods can be used to detect the degree of similarity between potentially parallel pages (Patry and Langlais, 2011).

Once a corpus has been collected, it has to be ‘cleaned’ in order to become suitable for further processing (Baroni et al., 2008). Usually, the cleaning steps include:

- removal of duplicates and near-duplicates, e.g., normal and ‘print-friendly’ versions of the same page;

- unification of encodings used on webpages, e.g., KOI-8, CP1251 and UTF-8 are frequently used for Russian;

- identification of boilerplate, navigation frames and foreign language fragments (ovals in Figure 6 indicate areas to be cleaned, otherwise words like *Wikipedia* or *Main page* can become disproportionately frequent);

- identification of paragraph breaks (and, possibly, of sentence boundaries).

The next step in corpus processing is tokenisation, i.e., identification of breaks between word-like units. In languages without orthographically marked word boundaries, e.g., Chinese, the procedure is complicated enough to warrant a special competition on automatic text segmentation (Ng and Kwong, 2006). In European languages the procedure is relatively straightforward, but even in this case there can be problems with making tokenisation decisions, e.g., on the apostrophes in *the Zero-X's course* vs. *Cote d'Azur* in English. A separate problem concerns tokenisation of compounds, which lack explicit word boundaries, e.g., *Agrarstukturverbesserungsmaßnahmen* in German ('measures to improve the agrarian structure').

Finally, texts are usually processed by part-of-speech (POS) taggers and lemmatisers. POS tagging helps in identification of basic text chunks, such as noun phrases, while lemmatisation helps in reducing data sparsity. The use of word forms or simple stemming (removal of common endings) is feasible in English, but a word in languages with agglutinative or inflective morphology (e.g., Slavonic or Turkic) can have a large number of different forms. Lack of lemmatisation in such cases considerably reduces the coverage (since many morphological variations in the corpus correspond to one and the same lemma). Some lemmatisers can perform disambiguation, e.g., deciding the lemma of *left* in *on the left* and *he has left*, while others list all possible options for each form.

Finally, for open-ended retrieval tasks corpora have to be encoded to keep and retrieve information concerning metatextual descriptors of documents (at least, their provenance, possibly also their domain and language), results of tokenisation, POS tagging and lemmatisation. Encoding of modern corpora is generally based on XML using two approaches to annotation. One entails the use of tags to "enrich" the text, so that each document has its metatextual tags (e.g., describing the author, title, date of its creation), while each word has information on its POS tag, lemma and other properties. This makes it possible to use standard XML indexing tools (e.g., Berkeley DB XML) or tools designed with corpus queries in mind (e.g., IMS Corpus Workbench or XAIRA). Another approach is to use a “stand-off” annotation, e.g., (Ide et al., 2000), in which each token has an identifier, while its properties are described in separate XML files referring to the identifier. The advantage of the stand-off scheme is that new layers of annotation can be added or removed without disturbing the original corpus. Its disadvantage is related to the assumption that a corpus is a fixed collection of texts with eternal identifiers assigned to each word. However, many modern corpora are frequently updated, e.g., by downloading new versions of the Wikipedia or crawling updated websites, thus putting strains on the need to maintain the link between identifiers in the corpus and its stand-off annotations.

3.2. Term extraction

Large domain-specific corpora offer tremendous advantages for creating and maintaining termbanks that reflect the current state of the art in their domain. However, getting terms out of corpora is not a reliable procedure, which is complicated by the statistical nature of term extraction and differences between languages. The terminology used by researchers in this field is also confusing, as different researchers use different names to refer to slightly different aspects of the procedure, e.g., term extraction, term recognition, keyword identification, keyword recognition, glossary extraction. There are also important differences between methods for extracting single-word and multiword terms. Some term extraction methods can extract terms of both types, while others are restricted to single-word terms only.

The earliest method of keyword identification is the TF*IDF method (Term Frequency × Inverse Document Frequency) first proposed by Karen Spärk Jones in the beginning of 1970s (Spärck Jones, 1972) and modified in many different ways since. The definition used below assumes a collection of $|D|$ documents, $f_{i,j}$ is the number of occurrences of term t_i in document d_j, and $|D_i|$ is the number of documents term t_i appears in. Then TF*IDF is computed for each term in each document (Manning et al., 2008):

$$TF*IDF_{i,j} = TF_{i,j} \times IDF_i; TF_{i,j} = \frac{f_{i,j}}{\sum_k f_{k,j}}; IDF_i = \log \frac{|D|}{|D_i|}$$

Words with the highest TF*IDF score across the whole collection are treated as terms (the cut-off threshold for selecting the terms depends on the document collection and application, and is usually selected in a trial-and-error manner). The method is fast and does not require any additional data apart from a document collection, thus making it language-independent. As this method can miss terms common to the entire collection (if a term occurs in every document, its IDF will be zero), more successful term extraction methods compare the frequency of words in a document collection against a reference corpus by measuring the importance of seeing t_i in D rather than in the reference corpus. Various methods for computing the statistical significance of this fact are mutual information (MI), χ^2, log-likelihood (LL), etc (Kageura and Umino, 1996), which are all based on the following table for each word:

	D collection	Reference corpus	Total
Frequency of term	a	b	a+b
Frequency of other terms	c-a	d-b	c+d-a-b
Corpus size	c	d	c+d

Then the expected values E1 and E2 and the log-likelihood value G2 are calculated by taking into account the relative frequency of terms in the two corpora as well as the absolute number of their occurrences as evidence of its statistical significance (Rayson and Garside, 2000):

$$G2 = 2(a\ln(\frac{a}{E1}) + b\ln(\frac{b}{E2})); E1 = c\frac{a+b}{c+d}; E2 = d\frac{a+b}{c+d}$$

These methods are successful in extracting single words, but they miss multiword terms, which take a very large proportion of the termset in any domain. Methods like C-value (Frantzi and Ananiadou, 1999), Glossex (Kozakov et al., 2004), Termex (Sclano and Velardi, 2007) can also identify multiword units by selecting the most frequent n-grams (sequences of n words) and filtering their list, either using linguistic patterns (e.g., ADJ+NOUN, NOUN+of+NOUN, NOUN+NOUN in English) or the heuristics of containment, i.e., if the frequency of a sequence of words is comparable to the frequency of a longer sequence that contains it, the former n-gram is discarded. For instance, even though *graphical user* is frequent in computer science texts and it follows the pattern of ADJ+NOUN, it is not selected as a potential term, since it is nearly always embedded in a longer expression *graphical user interface*.

Judge	TF-IDF	Weirdness	C-value	Glossex	Termex	Voted
1	0.67	0.80	0.59	0.81	0.93	0.97
2	0.79	0.85	0.69	0.83	0.95	0.97
3	0.77	0.77	0.68	0.83	0.95	0.97

Table 1. Qualitative evaluation of term extraction methods, from (Zhang et al., 2008)

Termex utilises a slightly more complex strategy by considering for each term its domain relevance (as the ratio of its frequency within the domain in comparison to a reference corpus), domain consensus (the term is distributed over a large number of documents in the domain), lexical cohesion (similar to the containment filter) and structural relevance (terms that are used in the title or keywords or are italicised/underlined in a web document are more likely to be terms).

The accuracy of an algorithm extracting a list of terms (T_e) from a domain-specific corpus can be measured by precision (the proportion of terms in the list that are relevant terms in the domain), recall (the proportion of relevant terms in the list in comparison all relevant terms contained in this corpus, T_r) and F-measure, which is a harmonic mean between precision and recall:

$$P = \frac{|T_r| \cap |T_e|}{|T_e|}; R = \frac{|T_r| \cap |T_e|}{|T_r|}$$

$$F1 = \frac{2 \cdot (P \cdot R)}{(P + R)}$$

Usually, there is a trade-off between recall and precision: greater precision (less noise in the list of extracted terms) means fewer terms overall, thus negatively affecting recall, and vice versa. Also computing recall is more difficult as this requires manual extraction of all terms form a potentially large corpus, whereas it is easier to judge precision by assessing the top N terms in the list returned by a term extraction algorithm. A recent study of term extraction methods in (Zhang et al., 2008) shows that precision for the top 100 terms extracted from a corpus of 1.3 million words from the English Wikipedia reaches 97%, see Table 1. The voting algorithm listed in the last column of Table 1 is based on the weighted voting strategy: the rank of term *t* is based on the ranks R(t_i) of this term produced in all other algorithms weighted by their precision.

3.3. Development of ontologies

The traditional way of developing ontologies is by organising the body of knowledge in a domain into a network of concepts, manual selection of relevant facts and representing the concepts and facts using the ontology language. The Cyc project is a prime example of this method. Upper ontologies are also normally created in this way. However, population of ontologies with facts is much easier to automatise. The CIA Worldfact book, which contains information about countries, their capitals, borders and economic indicators in a somewhat structured form, is relatively easy to convert into a formal ontology. Wikipedia 'infoboxes' can be also used for this purpose.

A diverse corpus with less-structured facts can be also used for semi-automatic population of large ontologies. First of all, a considerable part of the effort involved in constructing ontologies is devoted to extracting the list of names for concepts. The names themselves can be detected by using one of the term extraction mechanisms described above. Some links between names and concepts can be extracted by detecting patterns and applying machine learning techniques. For instance, the contexts in which the names of countries are used share many words, e.g., *to visit/to return to/GDP of Germany/Italy/Spain/Switzerland*; they also occur in lists of two-three names (*in France, Italy and Spain; between Germany and Switzerland*). This information can be used for creating a cluster of similar words, such as names of countries and big cities (Lin, 1998; Sharoff et al., 2006). Once the lists of countries and cities are known, the set of facts about which city is the capital of which country can be retrieved from the same corpus using patterns like *Tirana, capital of Albania; the Hungarian capital of Budapest*, etc. Similar patterns can be used to retrieve IS-A or HAS-A relations from free text definitions like (1) or (2) above. Such methods invariably produce noise, e.g., an example like *Berlin took over from Bonn as the capital of Germany* can produce a false assertion that Bonn *is* the capital of Germany.

More advanced methods for populating ontologies involve also automatic detection of patterns like *is a* or *the capital of* from the seed set of concepts and terms. For instance, if we start with seed facts that *apples, peaches, pears* are kinds of fruit, we can discover automatically frequent ways for express-ing the relation between a class Y and its members X in English, including *X, such as Y; Y, including X; X, RB called Y* (*RB* in the last case means that an adverb can be frequently used before the word *called*, e.g., *often, sometimes, popularly, officially*). Then, we can apply the newly learned patterns for the IS-A relation to retrieve more facts. Pantel and *et al*, who presented this method, also evaluated the precision of fact detection using automatically extracted patterns and found it to be in the range of 50-70% (Pantel et al., 2004). For more information on techniques in ontology extraction, see an overview in (Staab and Studer, 2004).

4. Applications in document authoring and translation

In considering multilingual applications of these activities, it is useful to bear in mind the three major functions of translation, whether by human or machine: dissemination (of outgoing information), assimilation (of incoming information) and communication (dialogue within a group).

Dissemination starts with the authoring of the source text. To be effective and safe, technical documentation must be unambiguous and easy to understand, using words and terms believed to be accessible to the reader (and to the translator). A significant factor in promoting comprehension is consistency – if the same part of a machine is variously referred to as the *cover* and the *flap*, the user is likely to be confused. The provision of consistent and transparent terminology is an asset which offer significant advantages. In larger organisations, efforts to this end may go as far as establishing 'structured' or 'controlled' authoring.

A controlled language (CL) is a version of a human language that embodies explicit restrictions on vocabulary, grammar and style for the purpose of authoring technical documentation (Kittredge, 2003; Nyberg et al., 2003). With roots in the Simplified English of the 1930s, the initial objective was to minimize ambiguity and maximize clarity for human readers, including non-native speakers of English, and so to avoid the need for translation altogether. Probably the best-known CL is AECMA Simplified English [4], which is a *de facto* standard in the aerospace industry; the concept has also been widely adopted in the automotive and IT sectors. Even within the same sector CLs vary from one company to another while respecting the same general principles.

Accordingly, at the lexical level (which constitutes the major part of a CL specification), a CL will define the approved technical terms and often explicitly list any deprecated terms which authors tend to use in error. Moreover, the prescriptions extend to non-technical expressions. For example, AECMA restricts the use of *about* to ‘concerned with’, specifying that *approximately* should be used for the other frequent sense; *support* can only be used as a count noun (*Put a support under the item* but not *Offer support*), and when a verb is required to express this idea it must be *hold* (*Hold the item* but not *Support the item*). Thus, the principle of univocity applies even beyond terminology to what would usually be considered ‘vocabulary’.

Tools exist to automate compliance checking at all levels of a given CL [5] and even to propose corrections for recurrent error patterns, some of which will involve misuse of terms. The rules and lexicons of these tools are often user-definable.

Clarity and consistency at source is a major factor in reducing translation costs, and the benefits multiply in direct proportion to the number of target languages. In the dissemination scenario, it is imperative that the terminology be shared across all the applications that contribute to the creation and translation processes – mono- and multi-lingual online glossaries, authoring tools, translation memory tools, machine translation engines – in order to maximise lexical coverage. Hence the importance of exchange formats like TBX, designed for interoperability.

Increasingly, the component parts of technical documents are managed in Content Management Systems (CMS). Modularization on a large scale entails significant design and implementation costs, but DITA (Darwin Information Typing Architecture [6] provides an increasingly widely adopted infrastructure. Designed for developing technical product documentation, it specifies three basic topic types: *concept* (for background information), *reference* and *task*. A task topic, for example, is intended for instructional procedures and is itself modularized into sub-elements containing content for pre-requisites (e.g., preparation of ingredients before cooking begins), steps, options, results and post-requisites (e.g., re-setting or cleaning equipment after a process), among others, each instance potentially re-usable in many places. This is analagous to ontology building, except that what is represented is not objects but states (e.g., *The stack is empty.*) or events (e.g., *The dialog box appears.*). This has many implications for the work of ontologists, terminologists and ‘information architects’.

The assimilation of documents from unrestricted third-party sources is well known to be a much more complex task and the inconsistencies and constant evolution of individual and group usage will continue to trouble the core task of reliably extracting meanings from the words of texts which we identified at the beginning of this chapter.

Much more attention needs to be paid to the shifting nature of terminology within subject communities. The corpora used for constructing ontologies and extracting terminology for both statistical and rule-based MT systems must be constituted to reflect not only a common subject domain and genre but also the strength of networks of communication.

LDOCE: *landing* **land·ing** /’lændin/

1. [C and U] the action of bringing an aircraft down to the ground after being in the air [≠ take-off]: *take-off and landing procedures*; crash/emergency landing (=a sudden landing caused by a problem with the engine etc); *the Apollo moon landings*→**soft landing**
2. [C] the floor at the top of a set of stairs or between two sets of stairs: *the first-floor landing*
3. [C] the action of bringing soldiers onto land that is controlled by the enemy: *the first landings of American Marines at Da Nang*

OED: **I.** The action of the verb **LAND**.

1.a. The action of coming to land or putting ashore; disembarkation.
c1440 *Promp. Parv.* 312/1 Londynge fro schyppe and watur, applicacio. **1577-87** *HOLINSHED Chron.* I. 9/2 They take landing within the dominion of king Goffarus. ...

1.d. The (or an) action of approaching and alighting on the ground or some other surface after a flight. **happy landings!**: see HAPPY a. 3.
1784 [see LAND v. 8b]. **1909** *Flight* 13 Feb. 93/1 (*heading*) Flight ‘landings’. ...

8. *attrib.* and *Comb.*, as (sense 1) *landing area, fee, ... tower, vehicle;* (sense 3) *landing-gaff, -hook, -ring*; **landing beam** *Aeronaut.*, a radio beam to guide aircraft when landing; **landing card** ...

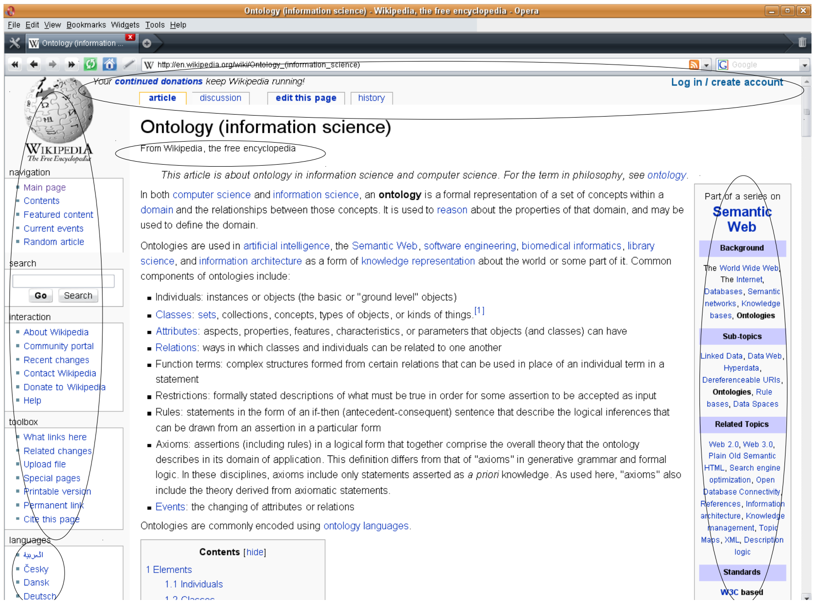
WordNet: *landing* Noun

1. landing (an intermediate platform in a staircase)
2. landing, landing place (structure providing a place where boats can land people or goods)
   - Direct hyponym: dock, dockage, docking facility
   - Part meronym: landing stage
   - Direct hypernym: structure, construction
   - Part holonym: seaport, haven, harbor, harbour
3. landing (the act of coming down to the earth (or other surface)) *the plane made a smooth landing*; *his landing on his feet was catlike ...*

Pocket Oxford-Hachette French Dictionary: *landing* noun

1. (*at turn of stairs*) palier m; (*storey*) étage m;
2. (*from boat*) (*of people*) débarquement m; (*of cargo*) déchargement m;
3. (*by plane*) atterrissage m (**on** sur).

Figure 5. Examples of entries for *landing* from four dictionaries



Ontology (information science)

From Wikipedia, the free encyclopedia

This article is about ontology in information science and computer science. For the term in philosophy, see ontology.

In both computer science and information science, an **ontology** is a formal representation of a set of concepts within a domain and the relationships between those concepts. It is used to reason about the properties of that domain, and may be used to define the domain.

Ontologies are used in artificial intelligence, the Semantic Web, software engineering, biomedical informatics, library science, and information architecture as a form of knowledge representation about the world or some part of it. Common components of ontologies include:

- Individuals: instances or objects (the basic or "ground level" objects)
- Classes: sets, collections, concepts, types of objects, or kinds of things.[1]
- Attributes: aspects, properties, features, characteristics, or parameters that objects (and classes) can have
- Relations: ways in which classes and individuals can be related to one another
- Function terms: complex structures formed from certain relations that can be used in place of an individual term in a statement
- Restrictions: formally stated descriptions of what must be true in order for some assertion to be accepted as input
- Rules: statements in the form of an if-then (antecedent-consequent) sentence that describe the logical inferences that can be drawn from an assertion in a particular form
- Axioms: assertions (including rules) in a logical form that together comprise the overall theory that the ontology describes in its domain of application. This definition differs from that of "axioms" in generative grammar and formal logic. In these disciplines, axioms include only statements asserted as *a priori* knowledge. As used here, "axioms" also include the theory derived from axiomatic statements.
- Events: the changing of attributes or relations

Ontologies are commonly encoded using ontology languages.

Contents [hide]

1 Elements
1.1 Individuals
1.2 Classes

Part of a series on
Semantic Web

Background

The World Wide Web, The Internet, Databases, Semantic networks, Knowledge bases, **Ontologies**

Sub-topics

Linked Data, Data Web, Hyperdata, Dereferenceable URIs, **Ontologies**, Rule bases, Data Spaces

Related Topics

Web 2.0, Web 3.0, Plain Old Semantic HTML, Search engine optimization, Open Database Connectivity, References, Information architecture, Knowledge management, Topic Maps, XML, Description logic

Standards

W3C based

Figure 6. Text cleaning example from Wikipedia

Bibliography

Aston, G. and Burnard, L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press, Edinburgh.

Baroni, M. and Bernardini, S. (2004). Bootcat: Bootstrapping corpora and terms from the web. In *Proc Conference on Language Resources and Evaluation (LREC'04)*, Lisbon.

Baroni, M., Chantree, F., Kilgarriff, A., and Sharoff, S. (2008). Cleaneval: a competition for cleaning web pages. In *Proc Sixth Language Resources and Evaluation Conference, LREC 2008*, pages 638–643, Marrakech.

Bateman, J. A. (1990). Upper modeling: Organizing knowledge for natural language processing. In *Proc. of the Fifth International Workshop on Natural Language Generation*, pages 54–61, Dawson, PA.

Beckett, D., editor (2004). *RDF/XML Syntax Specification (Revised)*. W3C Recommendation. http://www.w3.org/TR/rdf-syntax-grammar/.

Chakrabarti, S., van den Berg, M., and Dom, B. (1999). Focused crawling: a new approach to topic-specific web resource discovery. In *Proc. 8^{th} International World Wide Web Conference*, Toronto.

Dreyfus, H. L., editor (1982). *Husserl, Intentionality, and Cognitive Science*. MIT Press, Cambridge, MA.

Eisele, A. and Chen, Y. (2010). MultiUN: A multilingual corpus from United Nations documents. In *Proc Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Frantzi, K. T. and Ananiadou, S. (1999). The C-value / NC-value domain independent method for multi-word term extraction. *Journal of Natural Language Processing*, 6(3):145–179.

Gibbon, D., Nord, H. U., and Trippel, T. (1997). *EAGLET: EAGLES Termbank for Spoken Language Systems*. EAGLES SLWG. http://coral.lili.uni-bielefeld.de/EAGLES/WP5/termdeliv97/.

Grishman, R. and Kittredge, R., editors (1986). *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*. Lawrence Erlbaum.

Halliday, M. A. K. and Matthiessen, C. M. I. M. (1999). *Construing experience through meaning: a language-based approach to cognition*. Cassell, London.

Hartmann, R. R. K. and James, G., editors (2001). *Dictionary of Lexicography*. Routledge.

Hillmann, D. (2005). *Using Dublin Core*. DCMI. http://dublincore.org/documents/usageguide/.

Ide, N., Bonhomme, P., and Romary, L. (2000). XCES: An XML-based encoding standard for linguistic corpora. In *Proceedings of the Second International Language Resources and Evaluation Conference*, Athens.

Kageura, K. (1995). Towards the theoretical study of terms — a sketch from the linguistic viewpoint. *Terminology*, 1(1):103–119.

Kageura, K. and Umino, B. (1996). Methods of automatic term recognition: A review. *Terminology*, 3(2):259–289.

Kilgarriff, A. (1997). “I don’t believe in word senses”. *Computers and the Humanities*, 31(2):91–113.

Kilgarriff, A. (2003). Linguistic Search Engine. In *Shallow Processing of Large Corpora: workshop held in association with Corpus Linguistics 2003*, Lancaster.

Kittredge, R. (2003). Sublanguages and controlled language. In Mitkov, R., editor, *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Kittredge, R. and Lehrberger, J., editors (1982). *Sublanguage: Studies of Language in Restricted Semantic Domains*. Walter de Gruyter.

Koehn, P. and Knight, K. (2002). Learning a translation lexicon from monolingual corpora. In *Proc. of the ACL-02 workshop on Unsupervised lexical acquisition*, pages 9–16.

Kozakov, L., Park, Y., Fin, T., Drissi, Y., Doganata, Y., and Cofino, T. (2004). Glossary extraction and utilization in the information search and delivery system for ibm technical support. *IBM System Journal*, 43(3).

Kučera, H. and Francis, W. N. (1967). *Computational analysis of present-day American English*. Brown University Press, Providence.

LDOCE (1995). *Longman Dictionary of Contemporary English*. Longman, Harlow, 3rd edition edition.

Lenat, D. B. and Gupta, R. V. (1990). *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Addison-Wesley.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proc. Joint COLING-ACL-98*, pages 768–774, Montreal.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

McEnery, T. and Wilson, A. (1999). *Corpus Linguistics*. Edinburgh University Press.

Mel'čuk, I. A. (1996). Lexical Functions: a tool for the description of lexical relations in a lexicon. In Wanner, L., editor, *Lexical Functions in Lexicography and Natural Language Processing*, pages 37–102. John Benjamins, Amsterdam.

Miller, G. (1990). WordNet: an online lexical database. *International Journal of Lexicography*, 3(4).

Ng, H. T. and Kwong, O. O. Y., editors (2006). *Proc. Fifth SIGHAN Workshop on Chinese Language Processing at joint COLING-ACL*, Sydney.

Niles, I. and Pease, A. (2001). Towards a standard upper ontology. In *FOIS'01: Proceedings of the international conference on Formal Ontology in Information Systems*, pages 2–9.

Nyberg, E., Mitamura, T., and Huijsen, W.-O. (2003). Controlled language for authoring and translation. In Somers, H., editor, *Computers and Translation. A translator's guide*, pages 245–281. John Benjamins.

OED (1989). *Oxford English Dictionary*. Clarendon Press, Oxford.

OFD (1994). *Oxford-Hachette French Dictionary*. Oxford University Press, Oxford.

Ogden, C. K. and Richards, I. A. (1923). *The Meaning of Meaning*. Routledge, London.

Pantel, P., Ravichandran, D., and Hovy, E. (2004). Towards terascale knowledge acquisition. In *COLING'04: Proceedings of the 20th international conference on Computational Linguistics*, pages 771–777.

Patry, A. and Langlais, P. (2011). Identifying parallel documents from a large bilingual collection of texts: Application to parallel article extraction in Wikipedia. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 87–95, Portland, Oregon.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In *Proc. of the Comparing Corpora Workshop at ACL 2000*, pages 1–6, Hong Kong.

Resnik, P. and Smith, N. (2003). The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Rundell, M. (1999). Dictionary use in production. *International Journal of Lexicography*, 12(1):35–53.

Sager, J. C. (1990). *A practical course in terminology processing*. Benjamins.

Sager, J. C. (1993). *Language Engineering and Translation: Consequences of automation*. Benjamins, Amsterdam/Philadelphia.

Sclano, F. and Velardi, P. (2007). TermExtractor: a web application to learn the common terminology of interest groups and research communities. In *Proc. 9th Conf. on Terminology and Artificial Intelligence, TIA 2007*, Sophia Antinopolis.

Sharoff, S. (2005a). The communicative potential of verbs of "away-from" motion in English, German and Russian. *Functions of language*, 12(2):205–240.

Sharoff, S. (2005b). Phenomenology and cognitive science. In Franchi, S. and Guzeldere, G., editors, *Mechanical Bodies, Computational Minds*, pages 471–487. MIT Press, Cambridge.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In Baroni, M. and Bernardini, S., editors, *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna.

Sharoff, S., Babych, B., and Hartley, A. (2006). Using comparable corpora to solve problems difficult for human translators. In *Proc International Confenrence on Computational Linguistics and Association of Computational Linguistics, COLING-ACL 2006*, pages 739–746, Sydney.

Sinclair, J., editor (1987). *Looking up: an account of the COBUILD Project in lexical computing*. Collins, London and Glasgow.

Smith, M. K., Welty, C., and McGuinness, D. L. (2004). *OWL Web Ontology Language*. W3C. http://www.w3.org/TR/owl-guide/.

Sowa, J. F. (2000). *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks/Cole Publishing Co., Pacific Grove, CA.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Sperberg-McQueen, C. M. and Burnard, L., editors (2002). *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Oxford, 4 edition.

Staab, S. and Studer, R., editors (2004). *Handbook on Ontologies*. Birkhäuser.

Stamatakis, K., Karkaletsis, V., Paliouras, G., Horlock, J., Grover, C., Curran, J. R., and Dingare, S. (2003). Domain-specific web site identification: The crossmarc focused web crawler. In *Proc. Second Workshop on Web Document Analysis (WDA)*, Edinburgh.

Temmermann, R. (2000). *Towards New Ways of Terminology Description*. John Benjamins.

Varantola, K. (2003). Translators and disposable corpora. In Zanettin, F., Bernardini, S., and Stewart, D., editors, *Corpora in Translator Education*, pages 55–70. St Jerome, Manchester.

Wright, S. E. and Budin, G., editors (2001). *Handbook of terminology management*. Benjamins.

Wüster, E. (1955). *Bibliography of Monolingual Scientific and Technical Glossaries*. Unesco.

Wüster, E. (1979). *Introduction to the General Theory of Terminology and Terminological Lexicography*. Springer.

Zhang, Z., Iria, J., Brewster, C., and Ciravegna, F. (2008). A comparative evaluation of term recognition algorithms. In *Proc Language Resources and Evaluation Conference (LREC'08)*, Marrakech.

Zipf, G. (1935). *The psycho-biology of language*. Houghton Mifflin.