Glossary

adjective phrase (AP) a complete construction **headed** by an adjective. APs typically modify nouns and occur as complements to verbs such as *be, seem, become*. For example: *The man* **guilty of this heinous crime** *was imprisoned*. *John seems rather stupid*.

adjunct or modifier an optional or secondary element in a construction which can be removed without affecting the structural status of the rest of the construction. For example, *yesterday* in: *John kicked the ball yesterday* (Compare *John kicked the ball* where *the ball* is not an adjunct, because **John kicked yesterday* is ungrammatical).

affix morpheme placed at the beginning (prefix), middle (infix), or end (**suffix**) of the **root** or **stem** of a word, e.g. **re***legal***ize**.

agreement the process whereby the form of one word requires a corresponding form of another - for example, the plural form *boys* requires a plural form of the demonstrative determiner *these*/**this*: *these boys* vs **this boys*.

algorithm a prescribed set of well-defined rules or instructions for the solution of a problem.

analysis the phase in natural language processing systems (including MT systems) in which a structure or representation is assigned to source language (input) sentences or the representation itself or the name for the module of linguistic rules involved.

anaphor a word or phrase which refers back to some previously expressed word or phrase or meaning (typically, pronouns such as *herself, himself, he, she*).

antecedent the word or phrase to which a later word or phrase (e.g. an **anaphor**) refers.

Artificial Intelligence (AI) the branch of Computing Science concerned with simulating aspects of human intelligence such as language comprehension and production, vision, planning, etc.

ASCII American Standard Code for Information Interchange - a standard set of codes used for representing alphanumeric information in a computer.

aspect a property of verbs or sentences, which refers primarily to the duration or type of activity described, e.g. the distinction between *Sam sang* and *Sam was singing*.

attribute value pair Many contemporary linguistic analyses use collections of features or attribute value pairs to encode various properties of a linguistic entity. In the pair [number sing], *number* is the attribute and *sing* is the value.

auxiliary (AUX) in English, auxiliary verbs are those which carry distinctions of **tense**, **aspect**, etc, such as *do*, *be* and *have*. The **modal** auxiliaries include *can/could*, *may/might*, *shall/should*, *ought to*, *need* and *used to*. Auxiliary verbs are opposed to **main** verbs (*walk*, *play*, etc.)

batch (processing) as opposed to interactive processing. In batch processing, a computer does not perform tasks as soon as requested, but groups similar jobs together into batches and carries them out together at some later time (e.g. overnight). Interactive processing allows the user to issue an instruction and have it carried out more or less instantly.

bitext a bilingual text which is aligned so that within each bilingual chunk the texts are translations of each other. The use of the term does not necessarily commit one as to the level at which a text is chunked and aligned, e.g. into sentences or paragraphs, but the chunks are very often sentences.

case a property of words, primarily nouns, which varies according to their syntactic function. English distinguishes three cases of pronouns, one used for pronouns which are the subject of finite verbs (*he*, *I*) one for possessive pronouns (*his*,*my*) and one for pronouns elsewhere (*him*, *me*). The case system of many other languages is much more extensive.

CD-Rom a compact disc used for the storage of data in read-only (ROM) format.

collocation phrases composed of words that co-occur for lexical rather than semantic reasons, for example, a *heavy smoker* is one who smokes a great deal, but someone who writes a great deal is not a *heavy writer*. This seems to be a lexical fact, not related to the meanings of *smoker* or *writer*.

common sense reasoning reasoning on the basis of common knowledge, as opposed to purely logical reasoning, or reasoning that depends solely on the meanings of words. A purely logical inference might be from *If it is Tuesday, Sam is in London* and *It is Tuesday* to the conclusion *Sam is in London*. An example of common sense reasoning might be the inference that if someone asks for a phone book it is because they want to look up a number, and make a phone call.

complement a term for all constituents of the sentence required by a verb except for the subject (e.g. the object is a complement of the verb).

compound two or more words which function as one word (e.g. *fireplace*, *video-tape*, *door handle*). Most common in English and closely related languages are noun-noun

compounds functioning as nouns. Because such compounds have the external behaviour and distribution of a lexical item, they are often taken to be morphological structures.

constituent a linguistic unit which is a component of a larger construction. These units can, in turn, be analysed into further constituents (e.g. a **noun phrase** can be analysed into a determiner and a **noun**).

constituent structure the structure of an expression in terms of the **constituent** syntactic parts and their categories (as opposed to analysis in terms of grammatical or semantic relations).

context all the factors which systematically determine the form, meaning, appropriateness or translation of linguistic expressions. One can distinguish between linguistic context (provided by the preceding utterances or text) and non-linguistic context (including shared assumptions and information).

controlled language a specially simplified version of a language which is adopted (typically by a company or a documentation section of a company) as a partial solution to a perceived communication problem. Both the vocabulary and the syntactic structures may be restricted.

corpus collection of linguistic data, either written texts or a transcription of recorded speech. Typically, corpora have to be quite large to be of any linguistic use (upwards of 100,000 tokens).

critiquing system a computer program which analyses a text and indicates where it deviates from the norms of language use.

database generally, any collection of information that can be created, accessed, and processed automatically. Many sophisticated software packages exist for creating and accessing databases of information.

dependency grammar a type of grammar which operates essentially in terms of types of dependencies or grammatical relation between heads and dependent elements of a construction rather than in terms of constituent structure.

derivational a term used in **morphology** to refer to one of the two main processes of work-formation, the other being **inflectional**. Derivational processes result in words of a different class. In English, the major derivational process is suffixation, e.g. *derive - derivation, happy - happiness, nation - national*.

electronic dictionary dictionary which is stored on computer and can be accessed by programs, e.g. so that definitions can be looked up and displayed on screen.

feature see **attribute-value pair**

finite a form of a **verb** that can occur as the head of a sentence. In *Sam wants to leave*, *wants* is finite, *leave* is non-finite.

gender 2 types of gender are distinguished in linguistics — natural gender, where items refer to the sex of real world entities, and grammatical gender, which has nothing to do with sex, but which signals grammatical relationships between words in a sentence and which is shown e.g. by the form of the article or the noun.

generation (also synthesis) the phase in a natural language processing system (including MT systems) in which a strings or sentences are produced from some sort of underlying representation, typically a meaning representation of some sort or the name for the module of linguistic rules which causes this to happen.

grammar the term is generally used to include **syntax** and **morphology** but may also be used in a wider sense to include rules of **phonology** and **semantics**. A grammar is a collection of linguistic rules which define a language.

grammatical relations the relations which hold between a **head** (such as a verb) and its **dependents**. For example, subject and object are grammatical relations borne by constituents in a sentence.

head the central or most important element in a construction which determines the external distribution of the construction and places certain requirements on the words or constituents it occurs with. For example, the verb *saw* is head of the sentence *The big man saw Mary* and of the VP *saw Mary*. Nouns are heads of NPs, prepositions are heads of PPs, adjectives of APs, etc. In lexicography, head is another term for **headword**.

headword word forming the heading of an entry in a dictionary.

homographs words which have the same spelling but which differ in meaning, e.g. *bank* (financial institution) and *bank* (of a river).

idiom a sequence of words which functions semantically as a unit and with an unpredictable meaning (e.g. *kick the bucket*, meaning *die*). This is generally accompanied by a degree of syntactic restriction.

imperative verb forms or sentence types that are used to express commands (e.g. *Go away!*)

indexical a word which depends on the context of utterance for its meaning (e.g. *I, you, here*).

indirect object (IOBJ) the constituent of a sentence most typically associated with the goal or recipient role. In English indirect objects are often PPs with the preposition *to*, e.g. *Lee gave the book* **to his friend**.

inflectional term in **morphology** assigned to affixes which encode grammatical properties such as **number**, **tense** and do not change the **part of speech** of the **stems** to which they are attached.

interlingual language independent, a linguistic knowledge based approach to MT where translation proceeds in 2 stages - **analysis** (where input string is mapped onto a language independence representation) and **generation**, cf. transfer.

intransitive a verb that does not take a **direct object** (e.g. *die*).

lexicon used synonymously with dictionary.

light verbs (also **support verbs**) verbs that are semantically empty or relatively empty (e.g. *take* in *take a walk*).

markup codes in some (text formatting) description language which determine how text will look when printed.

metaphor in metaphorical usage, expressions are used in a way that appears literally false. For example, using the word *boiling* to describe water which is simply too hot for comfort.

mood a term applied to **sentences** and **verbs** to signal a wide range of meanings, especially speaker's attitude to the factual content of utterances, e.g. certainty, possibility (e.g. *Sam must/may be at home*). The distinction between active and passive sentences/verbs is also sometimes considered a mood.

morphology the branch of **grammar** which studies the structure or forms of words. The main branches are **inflectional morphology**, **derivational morphology**, and **compounding**.

natural language a term which denotes a (naturally occurring) human language as opposed to computer languages and other artificial languages.

NLP (Natural Language Processing) the field of inquiry concerned with the study and development of computer systems for processing natural (human) languages.

noun phrase (NP) a complete construction **headed** by a **noun**. It can be substituted by, or act as antecedent for, a pronoun of the appropriate sort:
$[_{NP}$ *The man who I saw yesterday* $]$ *has just knocked at the door. Can you let* **him** *in?*

number the number of a noun or noun phrase generally corresponds to the number of real world entities referred to (e.g. singular NPs denote single individuals (*a table*), plural NPs denote collections of individuals (*two tables*). However the relationship between real number and grammatical number is not always straightforward - *trousers* is plural in form yet denotes a singular entity (as in *the committee are considering that question this afternoon*) and some nouns do not have distinct singular and plural forms (*sheep*, *salmon*).

object (OBJ) also direct object - the constituent of a sentence generally associated with the entity which undergoes the action. In English, the direct object of a verb is a NP and normally follows the verb, e.g. *Peter saw* **Mary**.

OCR Optical Character Reader. A device which scans printed textual material and converts it into electronic form, storing it in a file on the computer or disc. OCR technology has improved dramatically in recent years and is now a reasonably accurate way of making text available in electronic form.

participle the term covers both a word derived from a **verb** and used as an **adjective**, as in *a* **singing** *woman*, and the -ing and -en non-finite forms of the verb, as in *was* **singing** (present participle), *has* **given** (past participle).

particle an element which occurs in a single form (like a preposition in English) and with a function that does not easily fit into standard parts of speech classifications. Particles very often occur in constructions with certain verbs in English with varying degrees of idiosyncratic interpretation: *John took* **off** *at great speed (i.e. left). May gave herself* **up** *(i.e. surrendered)*

part of speech (category) the class of units used in the description of a language, e.g. **noun**, **verb**, **noun phrase**, **verb phrase**.

phonology the branch of linguistics which studies the sound systems of languages. Phonological rules describe the patterns of sounds used distinctively in a language, and phonologists are interested in the question of what constitutes a possible sound system for a natural language.

post-editing program that performs some operations on the output of another program, typically formatting the output for some device or filtering out unwanted items.

predicate traditional and modern grammars often divide sentences so that constituents other than the **subject** are considered together to form the predicate (e.g. *John (***subject***) kicked the ball (***predicate**)).

prepositional phrase (PP) a phrase headed by a preposition, a word such as *on*, *in*, *between*. Prepositions combine with other constituents (usually noun phrases) to form **prepositional phrases**, as in *The man sat* **on the bench**.

probabilistic a term for approaches to natural language processing (including MT) which rely to some extent on statistical methods.

pronoun a word that can substitute for a **noun phrase** (e.g. *he* can substitute for *John*).

prosodic indicating stress or intonation.

reading a sense of a word that can be distinguished from other senses or meanings of the

same word.

relative clause a clause which qualifies or restricts the meaning of the noun in a **noun phrase**. It may be introduced by words such as *who*, *which* and *that* in English: *the man* **who I saw this morning**, *the woman* **(that) I sent the letter to**.

root that part of a word that is left when all **affixes** have been removed (*industry* is the root of *preindustrial*).

selectional restrictions selectional restrictions are essentially semantic restrictions on combinations of words. For example, verbs place such restrictions on their subjects and objects - the verb *frighten* generally requires as (active) subject something animate which can experience fear.

semantics the branch of linguistics which studies meaning in language. One can distinguish between the study of the meanings of words (lexical semantics) and the study of how the meanings of larger constituents come about (structural semantics).

semantic role also called **deep case**, **semantic relation** or **thematic role**. A semantic role is a description of the relationship that a constituent plays with respect to the verb in the sentence. The subject of an active sentence is often the **agent** or **experiencer**. Other roles include **instrumental**, **benefactive**, **patient**: *Peter* **(experiencer)** *died. The cat* **(agent)** *chased the dog* **(patient)**.

SGML Standard Generalized Markup Language. A generic language for marking various formatting and other textual relationships in a text.

source language when translating, the language one is translating out of; in French to English translation, French is the source language.

speech act a declarative sentence can be used to perform a number of different **speech acts**. In uttering *It's cold in here* a speaker may perform an act of requesting the hearer to close the window or turn up the heating.

stem that part of a word to which **inflectional affixes** are attached (it consists of the **root** plus any **derivational affixes**.

subcategorization the pattern of complements selected by head, e.g. the verb *put* subcategorizes for an NP and a PP. *We put the car in the garage*, but not **We put the car*.

subject the constituent of an active sentence most typically associated with the 'doer' or 'undergoer' of an action. The verb agrees with the subject in person and number in English.

sublanguage a language used to communicate in a specialized technical domain or for a specialized purpose, for example, the language of weather reports, expert scientific

polemic or other modes of scientific discourse, user or maintenance manuals, drug interaction reports, etc. Such language is characterised by the high frequency of specialized terminology and often also by a restricted set of grammatical patterns. The interest is that these properties make sublanguage texts easier to translate automatically.

suffix an **affix** that is added following a **root** or **stem**, for example the boldface parts of *legal***ize**, *nation***al**.

syntax the rules of a **grammar** which govern the way words are combined to form sentences and other phrases in a language.

tag to tag a text is to annotate it with grammatical information. Usually tagging takes the form of **part-of-speech** annotations but semantic tags or tags encoding other linguistic information can be used. Tagging is usually performed automatically or semi-automatically.

target language when translating, the language one is translating into; in French to English translation, English is the target language.

tense a property of verbs relating primarily to the time at which the action or event denoted by the verb takes place. For example, past tense verbs, as in *Sam left*, describe events in the past.

testsuite a collection of **sentences** or sentence fragments collated to test the capabilities of a translation system or other NLP application.

thesaurus a list of words arranged according to meaning, rather than alphabetically as in a standard dictionary.

transfer the phase in MT where a source language representation is mapped onto a target language representation, a linguistic knowledge based approach to MT where translation proceeds in three stages — analysis (where input string is mapped onto a source language representation) transfer and generation.