2. Technology and grammar

2.1 Indexing and retrieval in textual databases

2.1.1 Indexing

The indexing of a textual database is based on a table which specifies for each letter all the positions (addresses) where it occurs in the storage medium of the database.

2.1.2 Advantages of an electronic index

- Power of search
- Flexibility
  - General specification of patterns
  - Combination of patterns
- Automatic creation of the index structure
- Ease, speed, and reliability
  - Query
  - Retrieval

2.1.3 Definition of recall and precision

Recall measures the percentage of relevant texts retrieved as compared to the total of relevant texts contained in the database.

For example: a database of several million pieces of text happens to contain 100 texts which are relevant to a given question. If the query returns 75 texts, 50 of which are relevant to the user and 25 are irrelevant, then the recall is 50 : 100 = 50%.

Precision measures the percentage of relevant texts contained in the result of a query.

For example: a query has resulted in 75 texts of which 50 turn out to be relevant to the user. Then the precision is 50 : 75 = 66.6%.

2.2 Using grammatical knowledge

2.2.1 Linguistic methods of optimization

A. Preprocessing the query

- Automatic query expansion
  (i) The search words in the query are automatically ‘exploded’ into their full inflectional paradigm and the inflectional forms are added to the query.
  (ii) Via a thesaurus the search words are related to all synonyms, hypernyms, and hyponyms. These are included in the query – possibly with all their inflectional variants.
  (iii) The syntactic structure of the query, e.g. A sold x to B, is transformed automatically into equivalent versions, e.g. B was sold x by A, x was sold to B by A, etc., to be used in the query.
- Interactive query improvement
  Prior to the search, the result of a query expansion is presented to the user to allow elimination of useless aspects of the automatic expansion and allow for an improved formulation of the query.

B. Improving the indexing

- Letter-based indexing
  This is the basic technology of search, allowing to retrieve the positions of each letter and each letter sequence in the database.
- Morphologically-based indexing
  A morphological analyzer is applied during the reading-in of texts, relating each word form to its base form. This information is coded into an index which for any given word (base form) allows to find all corresponding (inflected) forms.
- Syntactically-based indexing
  A syntactic parser is applied during the reading-in of texts, eliminating morphological ambiguities and categorizing phrases. The grammatical information is coded into an index which allows to find all occurrences of a given syntactic construction.
- Concept-based indexing
  The texts are analyzed semantically and pragmatically, eliminating syntactic and semantic ambiguities as well as inferring special uses characteristic of the domain. This information is coded into an index which allows to find all occurrences of a given concept.

C. Postquery processing

- The low precision resulting from a nonspecific formulation of the query may be countered by an automatic processing of the data retrieved. Because the raw data retrieved are small as compared to the database as a whole they may be parsed after the query and checked for their content. Then only those texts are given out which are relevant according to this post query analysis.

2.3 Smart versus solid solutions

2.3.1 Smart solutions

avoid difficult, costly, or theoretically unsolved aspects of the task at hand, such as

- Weizenbaum's Eliza program, which appears to understand natural language, but doesn't.
- Statistically-based tagging, which can guess the part of speech.
- Direct and transfer approaches in machine translation, which avoid understanding the source text.

2.3.2 Solid solutions

aim at a complete theoretical and practical understanding of the phenomena involved. Applications are based on ready-made off-the-shelf components such as

- Automatic word form analysis based on an online lexicon of the language and a rule-based morphological parser handling inflection, derivation and composition
- Syntactic parsing of free text based on the automatic word form analysis of the language
- Semantic interpretation of the syntactic analysis deriving the literal meaning
- Pragmatic interpretation relative to a context of use deriving the speaker meaning.

2.3.3 Comparison

- As long as the off-the-shelf components are not available, a smart solution seems initially cheaper and quicker. But smart solutions are costly to maintain and their accuracy cannot be substantially improved.
- The components of grammar are a long term investment that can be used again and again in a wide variety of different solid solutions. Improvements in the components of grammar lead directly to better applications.

2.3.4 Choice between smart or solid solution depends on application

- A smart solution providing a 70% recall in a large database is more than the user could hope to find by hand. Also, the user is not aware of what is missing in the query result.
- In contrast, the deficiencies of a smart solution providing 70% accuracy in machine translation are painfully obvious to the user. Furthermore, there are human translators available which are able to do a much better job.

2.4 Beginnings of machine translation

2.4.1 Language pairs

French $\rightarrow$ *English* and *French* $\leftarrow$ *English* are two different language pairs.

2.4.2 Formula to compute the number of language pairs

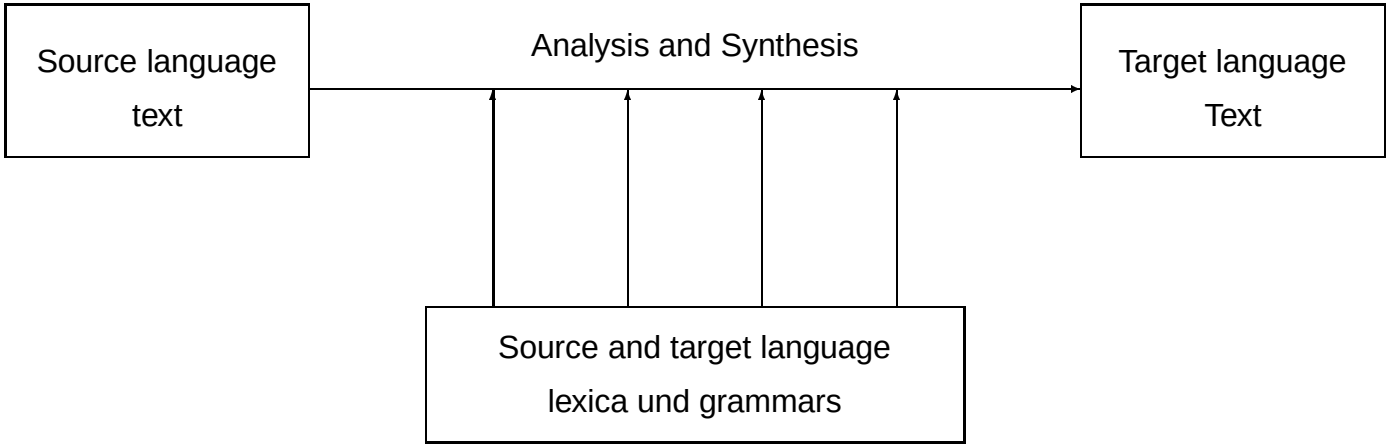
$n \cdot (n - 1)$, *where n = number of different languages*

For example, an EU with 11 different languages has to deal with a total of $11 \cdot 10 = 110$ language pairs.

2.4.3 Translating a French document in the EU

French $\rightarrow$ English	French $\rightarrow$ Spanish
French $\rightarrow$ German	French $\rightarrow$ Portugese
French $\rightarrow$ Italian	French $\rightarrow$ Greek
French $\rightarrow$ Dutch	French $\rightarrow$ Danish
French $\rightarrow$ Swedish	French $\rightarrow$ Finnish

2.4.4 Schema of direct translation



2.4.5 What is FAHQT?

FULLY AUTOMATIC HIGH QUALITY TRANSLATION

2.4.6 Examples of automatic mis-translations

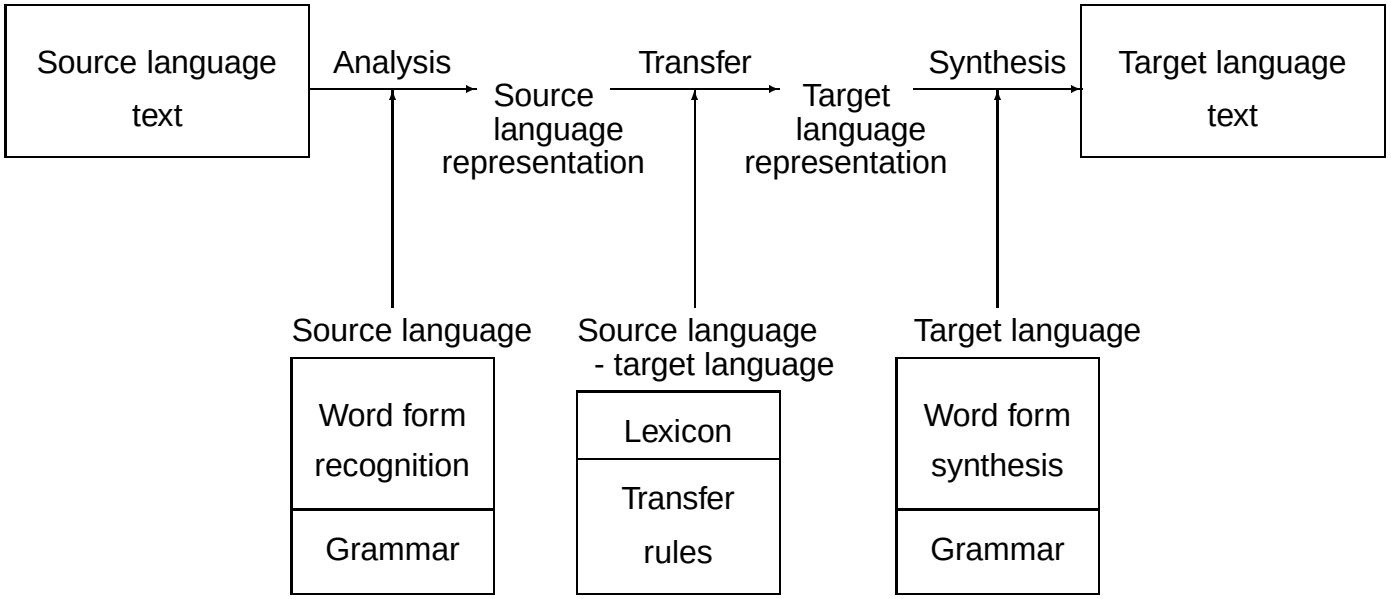
Out of sight, out of mind. $\Rightarrow$ *Invisible idiot.*
The spirit is willing, but the flesh is weak. $\Rightarrow$ *The whiskey is alright, but the meat is rotten.*
La Cour de Justice considère la création d'un sixième poste d'avocat général. $\Rightarrow$ *The Court of Justice is considering the creation of a sixth avocado station.*

2.4.7 The transfer approach

aims for a modular separation of the

- source language analysis and target language synthesis, of
- linguistic data and processing procedures, and of the
- lexica for source language analysis, target language transfer, and target language synthesis.

2.4.8 Schema of the transfer approach



2.4.9 Three phrases of a word form transfer *English-German*

1. Source language analysis:

   Unanalyzed surface: `knew`

   Morphological and lexical analysis: `(knew (N A V) know)`

   The source language analysis produces the syntactic category (N A V) of the inflectional form (categorization) and the base form know (lemmatization).

2. Source-target language transfer:

   Using the base form resulting from the source language analysis, a source-target language dictionary provides the corresponding base forms in the target language.

   `know` ⟹ `wissen`
   `kennen`

3. Target language synthesis

   Using the source language category (resulting from analysis) and the target language base forms (resulting from transfer), the desired target language word forms are generated based on target language morphology.

wußte	kannte
wußtest	kanntest
wußten	kannten
wußtet	kanntet

2.4.10 Shortcomings of the direct and the transfer approach

- Each language pair requires a special source-target component.
- Analysis and synthesis are limited to single sentences.
- Semantic and pragmatic analysis are avoided, attempting automatic translation without understanding the source language.

2.5 Machine translation today

2.5.1 Illustrating the importance of language understanding

- **Syntactic ambiguity in the source language**
  1. Julia flew and crashed the air plane.
     Julia (flew and crashed the air plane)
     (Julia flew) and (crashed the air plane)
  2. Susan observed the yacht with a telescope.
     Susan observed the man with a beard.
  3. The mixture gives off dangerous cyanide and chlorine fumes.
     (dangerous cyanide) and (chlorine fumes)
     dangerous (cyanide and chlorine) fumes

- **Lexical differences between source and target**
  1. The men killed the women. Three days later they were caught.
     The men killed the women. Three days later they were buried.
  2. know: wissen savoir
     kennen connaître
  3. The watch included two new recruits that night.

- **Syntactic differences between source and target**
  - German:
    Auf dem Hof sahen wir einen kleinen Jungen, der einem Ferkel nachlief.
    Dem Jungen folgte ein großer Hund.

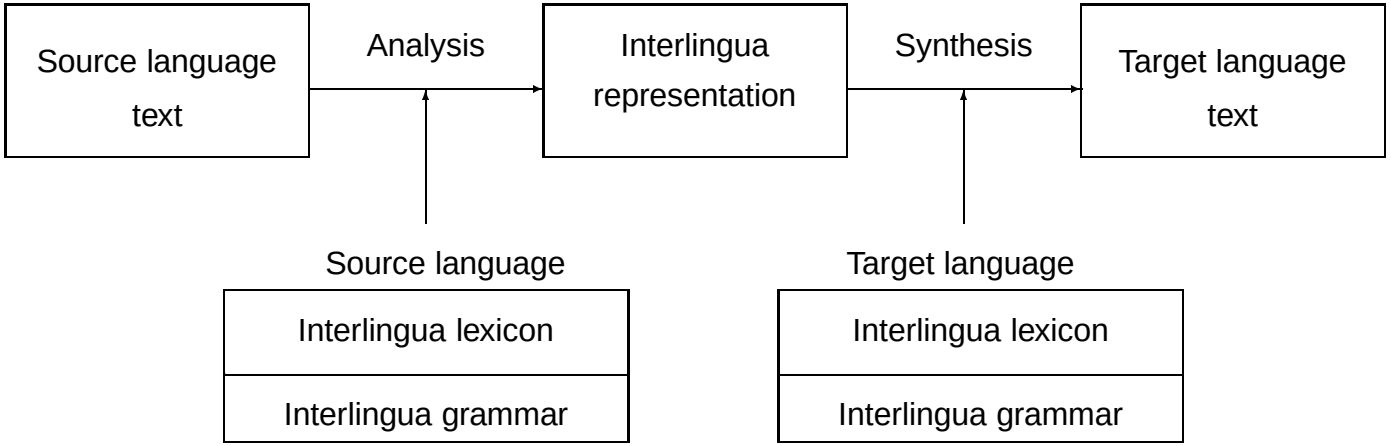
  - English:
    In the yard we saw a small boy running after a piglet.
    A large dog followed the boy.
    The boy was followed by a large dog.
- **Collocation and idiom**

  strong current | high voltage (but: *high current | *strong voltage)
  bite the dust | ins Gras beißen (but: *bite the grass | *in den Staub beißen)

2.5.2 Partial solutions for practical machine translation

1. *Machine aided translation* (MAT) supports human translators with comfortable tools such as on-line dictionaries, text processing, morphological analysis, etc.
2. *Rough translation* – as provided by an automatic transfer system – arguably reduces the translators' work to correcting the automatic output.
3. *Restricted language* provides a fully automatic translation, but only for texts which fulfill canonical restrictions on lexical items and syntactic structures.

2.5.3 Schema of the interlingua approach



2.5.4 Candidates proposed as interlingua

- an artificial logical language,
- a semi-natural language like Esperanto which is man-made, but functions like a natural language,
- a set of semantic primitives common to both, the source and the target language, serving as a kind of universal vocabulary.

3. Cognitive foundation of semantics

3.1 Prototype of communication

3.1.1 Variants of language communication

- two speakers are located face to face and talk about concrete objects in their immediate environment
- two speakers talk on the telephone about events they experienced together in the past
- a merchant writes to a company to order merchandise in a certain number, size, color, etc., and the company responds by filling the order
- a newspaper informs about a planned extension of public transportation
- a translator reconstructs an English short story in German
- a teacher of physics explains the law of gravitation
- a registrar issues a marriage license
- a judge announces a sentence
- a conductor says: Terminal station, everybody please get off.
- a sign reads: Do not step on the grass!
- a professor of literature interprets an expressionistic poem
- an author writes a science fiction story
- an actor speaks a role

3.1.2 Prototype of communication

> The basic prototype of natural communication is the direct face to face discourse of two partners talking about concrete objects in their immediate environment.

Possible alternatives: complete texts or signs of nature, such as smoke indicating fire.

3.1.3 Three components of the communication prototype

- Specification of the external *task environment*
- Structure of the *cognitive agent* including the internal *problem space*
- Specification of the *language*

3.1.4 Objects in the world of CURIOUS

- triangles (scalene, isoceles, etc.)
- quadrangles (square, rectilinear, etc.)
- circles and ellipses