

*Language Engineering
for UCLAF*

*Ralf Steinberger
JRC - Ispra*

*ralf.steinberger@jrc.it
Tel: + 39 - 332 78 6271*

23 October 1998

Agenda

- *What is Language Engineering (LE)*
- *Some LE applications*
- *Types of information ‘contained’ in language*
- *Some tools*
- *Common Problems in LE*
- *Methods used in LE*
- *Current and future LE activities of our group*
- *all in 30 minutes!*

Warning: *Many explanations in this presentation are simplified; distinctions made are often gradual!*

What is Language Engineering

- *Computational Linguistics (CL) is the cross section between computers and natural language.*
- *Natural Language Processing (NLP),
Language Technology (LT),
Language Engineering (LE)
are newer terms referring to the application areas of CL.*

Applications

- *Machine Translation (MT)*
- *Automatic document indexing: assign indexing words automatically to a text to facilitate retrieval (e.g. for libraries)*
- *Automatic text summarisation / abridgement: present only the most relevant information of a text*
- *Document / text retrieval: find one or more documents matching certain search criteria (e.g. a certain letter or texts on a certain subject) in a large repository of documents*
- *Cross-language document / text retrieval: query multilingual documents in one language*

Applications - 2

- ***Document classification:*** assign documents to categories of a given subject classification, e.g. decide whether newswires relate to sports, finance, agriculture, etc.
- ***Document filtering / Personal information filtering:*** select (user-) relevant documents from a large flow of incoming messages such as email messages or newswires and present only these to the user
- ***Document navigation:*** in a large repository of documents, find relevant ones and move from one document to other related ones, e.g. by comparing the overlap of common indexing terms or using other similarity indicators
- ***Extraction of entities from texts:*** e.g. proper names, company names, place names, etc.
- ***Extraction of scenarios from texts:*** e.g. succession of company leaders, bank transfers, etc.

Applications - 3

- *Spell checkers:*
- *Grammar / syntax checkers*
- *Style checkers*
- *Interference checkers*
- *Computer-Assisted Language Learning (CALL)*
- *Monolingual or multilingual document generation*
- *Translator's workbench: a set of tools improving translators' efficiency including MT, translation memory, on-line dictionaries (bilingual, synonyms, ...), automatic terminology extraction, a terminology bank, etc.*
- *Speech recognition and synthesis: to make machines more user-friendly or for dictation machines, etc.*
- ...

Phonetics, Phonology, Morphology, Syntax, Semantics, Pragmatics

Phonetics: *sound production, physical difference between sounds*

Phonology: *sound system*

Morphology: *the form of words, inflection*
(*cas+a/e; vend+o/i/e/.../eró/...*)

Syntax: *the combination of words to bigger units*
(*NP → Det + Adv + Adj + N + ...*)

Semantics: *the meaning (signifiant and signifié) and the difference between meanings*

Pragmatics: *the intention*
(*Could you give me the butter? → ‘give me’, not: ‘are you able?’*)

Linguistic knowledge vs. contextual knowledge vs. world knowledge

Linguistic knowledge: knowledge contained in the language

She **wrote** a letter to her parents at the bar.

- *past tense (tense system of the grammar; morphology)*
- *SU + write + DOBJ + IOBJ (+ ADV) (sub-categorisation information of the verb write)*
- *SU ¹ OBJ (syntactic word order knowledge)*

Contextual knowledge: knowledge contained in the same text (or present due to common environment)

... . **She** wrote a letter to **her** parents sitting at the **bar**.

- *Anaphora (pronoun) resolution*
- *Word sense disambiguation (bar = 1. place for drinks, 2. straight piece of metal, 3. legal organisation, 4. stripe, ...)*

Linguistic knowledge vs. contextual knowledge vs. world knowledge - 2

World knowledge: *knowledge not contained in the language or in the text such as knowledge on the shape and character of things in the world, knowledge on inference mechanisms, etc.*

The man saw **the elephant** with the telescope.

(telescope is a tool to see)

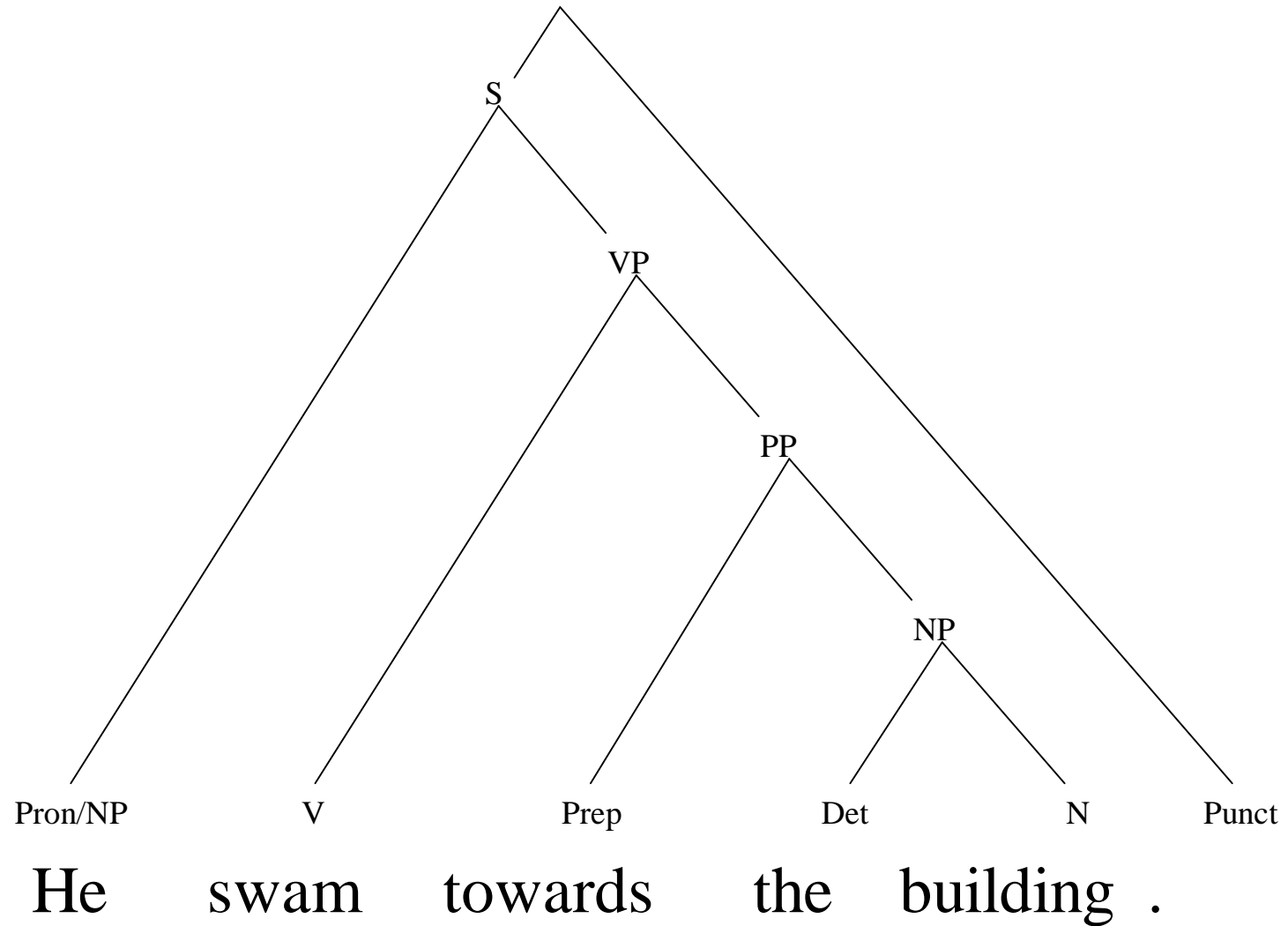
The elephant saw **the man** with the telescope.

(telescope is an attribute of the man)

A **small** elephant vs. A **large** fly

Tools & Resources

- ***Morphological analyser:***
takes word and yields morphological analysis
swims → *3rd person singular present tense indicative of the verb swim*
lu=swim, pos=verb, person=3, number=sg, tense=present, mode=indicative
- ***Lemmatiser:*** *yields the dictionary citation form of each word*
(infinitive of verb, singular nominative of noun)
swims → swim, swimming → swim, houses → house
- ***Part of speech (POS) tagger:*** *assigns a POS to every word in a text*
He-Pron swam-V towards-Prep the-Det building-N.
- ***Parser:***
assigns a syntactic structure (e.g. phrase structure) to a sentence
or a shorter or longer string



[[[[[He]_{Pron}]_{NP}[[swam]_v[[towards]_{Prep}[[the]_{Det}[building]_N]_{NP}]_{PP}]_{VP}]_S[.]_{Punct}]

Tools & Resources - 3

- ***Semantic tagger***: tries to identify the semantic reading of ambiguous words

bar: (a) place to have a drink (c) legal organisation
 (b) metal rod (d) stripe (bar code)

...

They met at the bar. → *place to have a drink (or metal rod?)*

They put him behind bars. → *metal rods (prison) (or ...)*

- ***Dictionaries***

- *bilingual*

- *monolingual with morphological, syntactic, semantic information*

- *thesaurus (expressing relations between words: part-of, associative, ...)*

- *names (places, people)*

- *tagger (possible POS of each word, statistical likelihood of each POS)*

- *synonyms, antonyms, related words, etc.*

- *false friends, easily confused words, sound alike, ...*

- ...

- ...

Common problems in Language Engineering

- ***Ambiguity of POS*** (e.g. building, saw (*Noun or Verb? see or saw?*))
→ need for POS disambiguation and sometimes semantic disambiguation
- ***Structural (syntactic) ambiguity***
[Time]_{SU} flies [like an arrow]_{ADV}. vs.
[Time flies]_{SU} like [an arrow]_{OBJ}.
- ***Semantic ambiguity*** (e.g. bar, suit)
- ***Ellipsis*** (*Es*: cantaría (1st, 3rd) = *En*: I/he/she would sing)
- ***Morphological ambiguity*** (*Es*: cantaría = *It*: canterei (1st)/canterebbe(3rd))
- ***Multi-word expressions*** (MWU) (colour liquid crystal display, door frame varnish colour, vs. ‘time flies’) (Flüssigkristallfarbbildschirm)

Common problems in LE - 2

- *Orthographic errors in texts (typos, bad OCR output)*
- *Discontinuous elements*
Er brachte das Buch 1993 heraus. (herausbringen = publish)
make good use of
- *Unclear sentence borders (...in the U.S.A. [Chancellor Kohl promised 3.2 million DM to Mr. Smith while Mr. Chirac avoided the issue.] ...)*
- *Structural difference between languages (complex MT transfer)*
He swam across the river. ≈
Il a traversé la rivière en nageant.
- *Use of metaphors, stylistic and lexical variation, ...*
- ...

Methods

- **Rule-based (symbolic, linguistic):**
 - *description of all the rules in a formalism, having large and completely coded dictionaries → complex applications are very time-consuming to develop.*
 - *typical applications are morphological analysers, tools to recognise names and multi-word expressions, MT and CALL systems*
- **Statistical:**
 - *using absolute or relative frequency of words, letters, co-occurrence of words, etc.*
 - *typical applications are language recognisers based on bigram frequency and summarising / abridging systems.*
 - *systems are often trained on manually encoded texts (e.g. POS taggers)*

Methods - 2

- *Hybrid methods combining rules and statistics:*
 - *e.g. POS tagging, lemmatisation, etc. for applications is done in a rule-based manner and word frequency statistics are used to identify indexing terms or similarities between documents, or*
 - *rule-based MT system identifies syntactic or lexical ambiguity and statistical data is used to choose the most likely reading*

Current activities of our group

- *Automatic **document indexing** for faster categorisation and retrieval of similar or typical fraud cases in IRENE95*
- *Automatic **subject domain recognition***
- *Automatic **language identification***
- ***Document clustering** to see whether it is possible to derive a fraud case classification; derivation of meta-knowledge from texts*
- *Also needed: POS-tagging, stop word lists, lemmatisation tool, tool to recognise multi-word terms, several dictionaries (synonyms, thesaurus, subject domain information, ...)*
- ***Building up a LE infrastructure** (tools developed now can be reused in other applications)*

Potential future activities of our group

- *Improve and refine current work (indexing, clustering, derivation of meta-knowledge from texts)*
- *Expand work to dealing with more EU languages*
- *Tackling the multilinguality problem:*
 - *multilingual document indexing*
 - *indexing using controlled vocabulary (e.g. Eurovoc or combined nomenclature):*
«PREP ALIM SAUCES, CONDIMENTS, FARINE» (CN 2103)
 - *cross-language document retrieval (problems: 1-to-n translations, how to search for information in other languages, ...)*
- *Help UCLAF's staff to find information using **query-expansion**, also multilingual (BSE ® cow, sheep, disease, bovine, UK, encephalopathy)*
- *Personal information filtering*
- *Summarisation / Abridging of texts*

En-Fr dictionary entry for 'bar'

Oxford SuperLex for Windows

File Edit View Search Book Help

bar

I noun

- (strip of metal, wood) barre *f*;
- (on cage, cell, window) barreau *m*; to put sb/be behind bars mettre qn/être derrière les barreaux;
- a bar of soap une savonnette; a bar of gold un lingot d'or; a bar of chocolate une tablette de chocolat;
- (block) (of soap, gold, chocolate) barre *f*;
- (obstacle) obstacle *m* (to pour; to doing pour faire); your age is not a bar votre âge ne constitue pas un obstacle;
- law (profession) the bar le barreau; to study for the bar se destiner au barreau; to be called to the bar entrer au barreau;
- law (in court) barre *f*, to come to the bar venir à la barre; the prisoner at the bar l'accusé/-e *m/f*;
- Sport (in gym, across goal) barre *f*; to practise on the bars s'exercer aux barres;
- music mesure *f*, two beats in a/to the bar deux temps dans une/par mesure;
- (in electric fire) résistance *f*;
- military GB (on medal) barrette *f*, US (on uniform) galon *m*;
- heraldry barre *f*.

II preposition sauf; all bar one tous sauf un seul; bar none sans exception.

III transitive verb (p prés etc -rr-)

- (block) barrer [way, path]; to bar sb's way barrer le passage à qn;
- (ban) exclude [person] (from sth de qch); interdire [activity]; journalists were barred l'accès était interdit aux journalistes; to bar sb from doing interdire à qn de faire; his religion bars him from marrying sa religion lui interdit de se marier;
- (fasten) mettre la barre à [gate, shutter]; the gate was barred on avait mis la barre au portail.

IV barred past participle adjective

- [window] à barreaux;
- (striped) barred with barré de [colour, mud].

V -barred (dans composés) four five-barred gate portail à quatre/cinq barreaux.

Idioms

a no holds barred contest une lutte où tous les coups sont permis;
it was a divorce battle with no holds barred le divorce a été une lutte où tous les coups semblaient permis.

