C SC 620
Advanced Topics in Natural Language Processing

Lecture 18

3/30

Reading List

- *Readings in Machine Translation*, Eds. Nirenburg, S. *et al*. MIT Press 2003.
- Reading list:
  - 12. Correlational Analysis and Mechanical Translation. Ceccato, S.
  - 13. Automatic Translation: Some Theoretical Aspects and the Design of a Translation System. Kulagina, O. and I. Mel’cuk
  - **16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.**
  - 17. The Proper Place of Men and Machines in Language Translation. Kay, M.

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- Concept of sublanguage
  - language of X is a *sublanguage* of English
    - where X = (physics, aeronautics, electronics, etc.)
  - It is within the domain of sublanguages that automatic translation appears to be practical
    - Example:
      - *Taum-meteo*: English -> French for weather reports
      - aviation maintenance manuals

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 2 Description of a Particular Sublanguage
- 2.1 The Corpus
  - TAUM = Traduction Automatique Université de Montréal
  - Instructions for aircraft maintenance
    - 70,000 words in English
    - 3,548 different words
    - nouns 1714 verbs 667 adjectives 664 adverbs 168
    - prepositions 134 numerals 63 quantifiers 46 pronouns 35
    - 571 idioms: 443 of which are technical

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 2.2 Restrictions
- 2.2.1 Lexical Restrictions
  - 4,876 different lexical items in 70,000 words
  - Estimate for full set of texts: 40,000 lexical items
  - Compare to Webster’s 3rd: 450,000
  - Vocabulary of this sublanguage is highly restricted
    - contains: aileron, motor, compressor, jack, filter, check, axial, quick-disconnect
    - not present: parsley, meson, seduce, endocrine, hope, think, believe, personal pronouns (I, me, we, us, he, she)
    - categories noun, verb, adjective and adverb are the most limited
    - all articles and coordinate conjunctions present
    - 80% of one-word prepositions: not apropos, notwithstanding

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 2.2.1 Syntactic Restrictions
  - Direct questions do not occur at all
    - Do you have your tool kit?
    - Is the motor turned off?
  - Tag questions inappropriate
    - Check the batteries, won’t you?
    - The switch should not be on, should it?
  - No simple past tense
    - The engine stopped
    - High temperatures caused buckling
  - No exclamatory sentences
    - How powerful the engine is!
    - What a complex hydraulic system this plane has!

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- Full range of constructions present:
  - passive, restrictions and non-restrictive relative clauses, extraposition, nominalization
- Long and complicated sentences common:
  - “This unit contains the fuel metering section. shutoff valve, and a mechanical governor that functions as either an over speed governor for the high pressure rotor or provides manual control when the electronic computer section of the fuel control system is deactivated.”
  - “… as lightweight, two-spool geared transonic-stage, front-fan, jet propulsion engine.”

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- Difficult problems
  - conjunction scope
    - “Disconnect pressure and return lines from pump”
  - compound bracketing
    - “The stability augmentor pitch axis actuator housing support”

- 2.2.3 Semantic Restrictions
- 2.2.3.1 Categorization and Subcategorization
  - Reduction in polysemy: word classes

case (N)	*case the joint
lug (N)	*they lugged the equipment from the plane
cake (V)	*the pilot likes banana cake
jerky (A)	*carry a pound of jerky on long flights
fine (A)	*fine them for smoking *a fine for smoking
cable (N)	*cable the forward compartment

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- Reduction in polysemy: senses within classes
  - eccentric (A) [-animate] *eccentric pilot
  - ball (N) *the annual ball
  - check (N) [+abstract] *cash this check
  - bore (V) [-animate object] *inaction may bore the crew
  - bore (N) *the pilot is a bore
    - cylindrical hole, inside diameter of a cylinder
- Categorial ambiguity

check	pump	case	drain	fitting
N	N	N	N	N
V	V	V	V	V

  - 2^5=32, but *case* is N onlyin corpus => 16

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

Case	ejection	door	locks	immediately
N	V	N	N	Adv
V			V	

- *Case* N only =>
- subject: *case ejection door* =>
- *locks* only candidate for a verb

- Semantic range reduction
  - A small heat exchanger uses engine fuel for cooling purposes
    - *cooling* modifies *purposes*
    - *cooling* takes *purposes* as object
    - only concrete objects are cooled in corpus (not tempers etc.)
    - *cool* (V): direct object [+concrete]

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 2.2.3.2 Specificity
  - the + N specific only
    - the oil tank is not a component of the engine
    - the computer provides increased fuel scheduling
  - no generic reference as in:
    - the dolphin is a mammal
    - the invention of the wheel was a crucial step
  - differs from a textbook
    - the motor is a machine that converts electrical into mechanical energy vs.
    - the motor is a constant-displacement piston type
  - Omission of articles
    - clean (the) reservoir system
    - French translation requires a definite article

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 2.2.3.3 Semantic Features
  - [+concrete] use only in this sublanguge
    - air, battery, dirt, machine, flap, flash, post, rod, solution, speed, spring, tool, net, web, race
  - [-human] use only
    - agent, body, boss, buffer, crank, elbow, governor, joint, nut, page, selector, starter
  - Subject/object restrictions

charge	object [+concrete]
circulate	subject [+fluid] (intransitive)
divert	object [+fluid]
function	subject [+part] (part of aircraft or related equip.)
top	object [+concrete]
die	subject [-animate]

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 2.2.3.3 Semantic Features
  - [male] [female] not relevant in corpus
  - [+human] used only on a few nouns
  - [+human] used mainly in signaling implied subjects for imperatives
    - check fan blade clearance
    - adjust pump pressure control valve

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 2.2.3.3 Semantic Features
  - Representation
    - *F (unary)
      - *concrete *abstract
    - +F, -F (binary)
      - air, oil, water, etc. [+fluid]
      - all other nouns must be marked [-fluid]
      - all verb argument positions which do not accept [+fluid] arguments
    - Unary representation used

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 2.3 Reduction
- 2.3.1 Omission of Definite Article
  - Optional in sublanguage
      - check indicator rod extension
      - check the ground test system
    - no attempt made to predict its presence or omission
    - *the* still the most frequent word in the corpus (2,925)
- 2.3.2 Omission of Copula
  - Check reservoir full
    - check that the reservoir is full
  - Check fluid level above REFILL mark
    - check that the fluild level is above REFILL mark
  - Check that fuel systems are full
  - Check fluid level indicator is registering correctly
  - Pump not delivering fuel (progressive)

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 2.3.3 Omission of That Complementizer
    - Check that fuel systems are full
    - Check fluid level indicator is registering correctly
  - Standard English
    - *we are checking the indicator is working
- 2.4 Frequently Occurring Forms
- 2.4.1 Imperative
  - maintenance manual is like a cook book
  - imperatives occurs very often

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 2.4.2 Non-Predicative Adjectives
  - marked as ATRIB in the parsing dictionary
  - 25% of all adjectives
    - John is proud (predicative)/the proud father (attributive)
    - (A) actual, chief, consequent, entire, respective
    - (B) nickel-cadmium, piston-type, pressure-regulating, anti-stall, single-point, non-priority
  - (B) is productive
    - X-type
    - X-Ving
    - anti-X
    - Xnum-Y
    - non-X
  - Do not inflect
    - *chiefer *pressure-regulatingest

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 2.4.2 Non-Predicative Adjectives
  - numerical expression + measure unit/noun
    - (A) 115/200-volt 0.0045 inch, 10-micron
    - (B) 3-phase, 19-cell, 2-stage, two-lobe
  - in (B)
    - *phase* is a measure unit wrt *generator*
    - *cell* wrt *battery*
    - *lobe* wrt *cam*
  - should not be entered as individual lexical items in the dictionary
  - convention
    - hyphen place between a numerical expression and measure unit when the compound is used as a prenominal modifier and to write the measure unit in the singular
      - three stage turbine
    - otherwise, no hyphen and measure unit is pluralized
      - the turbine has three stages

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 2.4.3 Noun Sequences
    - external hydraulic power ground test quick-disconnect fittings
    - fuselage aft section flight control and utility hydraulic system filter elements
    - fan nozzle discharge static pressure water manometer
  - A need to give highly descriptive names to parts of the aircraft in terms of their function and their relation to other parts
  - Likely to occur in any texts describing very complex machinery containing a large number of specialized parts
  - *empilage*: sequence from 1st adjective or noun to last noun
    - 4,400 different empilages in the corpus

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 2.4.3 Noun Sequences
  - bracketing problem
  - need to understand the syntactic and semantic relations involved
    - main fuel system drain valve
      - *main* applies to *fuel system*
      - valve’s function is to drain the main fuel system
  - 50 basic syntactic/semantic relations
    - have
    - whole-part
    - place
    - subject
    - object

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

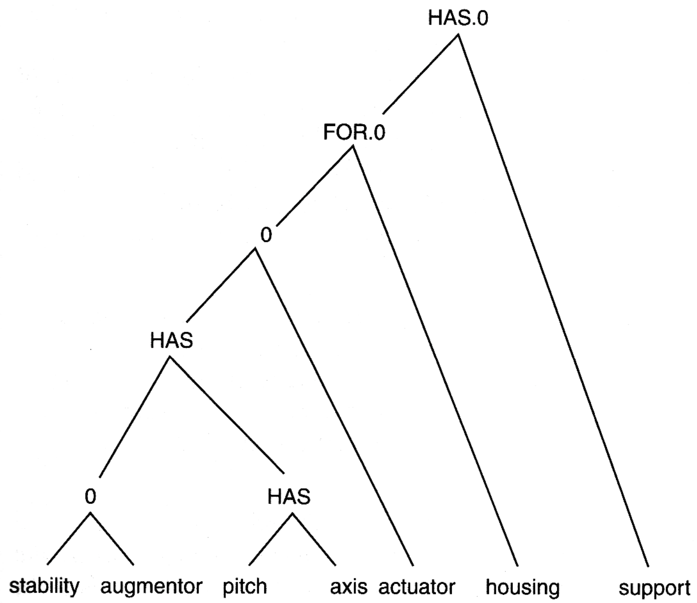

Figure 16.1
Stability is grammatical OBJECT of *augment*; *pitch* HAS an *axis*; *stability augmentor* HAS a *pitch axis*; *stability augmentor pitch axis is* OBJECT of *actuate*; the *housing* is FOR the *stability augmentor pitch axis actuator* (which is also OBJECT of *housing*); the *stability augmentor pitch axis actuator housing* HAS a *support* (and is also OBJECT of *support*).

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- Comment on noun sequences
  - Finite number of aircraft parts - just list them, see discussion in section 3.4
  - Suggestion to reference part number instead of tackling semantic decomposition

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 2.5 Idioms
  - technical term for any multiword expression which is entered into the dictionary
  - (I) meaning of expression is not predictable from the meanings of its components
    - with respect to
    - nose gear
    - finger tight
  - (II) translation idioms

• aspect ratio	allongement
• DC power	courant continu
• buttock line	section longitudinale

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 2.5 Idioms
  - (III) frequent item that “feels like” a compound word
    - landing gear
    - filter element
    - relief valve
  - (IV) rare expression and parsing it would require undesirable changes in parsing strategy
    - right and left of center
    - right [of something]
    - left [of something]

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 2.6 Text Structure
- 2.6.1 Gross Structure
  - texts divided into numbered sections each of which deals with a specific part of the aircraft
  - occurrence of a polysemous word may signal a particular meaning
    - capacity
      - volume in the hydraulic system
      - farads in the electrical system
    - valve
      - clapet (French)   hydraulics
      - soupape or valve   motors

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 2.6.2 Linking Devices
  - Discourse
  - Repetition
    - Install rotor on shaft, then align index marks on inner races of bearing. Position bearing on shaft with vendor identification marking on outer race on same side as puller groove on inner races
  - Shortening of nouns
    - Remove jumper hose from pressure in line. Cap line and hose

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 2.6.2 Linking Devices
  - Discourse
  - Pronouns
    - The main system relief *valve* is located on the left side of the engine compartment, just forward of the hydraulic reservoir. *It* is adjusted …
    - Candidates for antecedent of *it*:
      - main system relief valve feature [adjustable]
      - left side of the engine compartment
      - hydraulic reservoir
  - Nominalization
    - Vent manifold may be *leaking*. This *leakage* will allow …

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 2.6.2 Linking Devices
  - Discourse
  - Implicit Reference
    - Remove and inspect the fuselage aft section flight control and utility hydraulic system filter elements. If **found** to be highly contamin[at]ed, **clean** and **reinstall**, then remove and inspect all flight control actuator filter elements. If **found** to be highly contaminated, **clean** and **reinstall**, then remove and inspect all hydraulic system restrictors. If restrictors are found to be highly contaminated, **clean** and **reinstall**.
  - List Context
    - Correct wiring
      - Remedy
    - Bleed fittings on brake assembly (Imperative or Location)
  - [Present system does not handle discourse. Rejected for reasons of economy.]

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 2.7 Odds and Ends
- 2.7.1 Numerical Expressions and Reference
    - secure with two attaching bolts
    - gauge should read 1000 PSI
  - all numerical expressions represented by Arabic numerals after parsing since this is more convenient at the transfer stage
- 2.7.2 Labels
  - Corpus word that should be not translated - refers to a label on a part of the aircraft or related test equipment
  - All caps used
    - Set switch to ON
    - Ensure that the PITCH CONT switch is ON

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 2.7.3 N-V*ing*, N-V*ed* Compounds
    - gear-driven
    - cockpit-mounted
    - seat-adjusting
    - spring-loaded

      air-separating
      motor-operated
      spring-adjusting
  - N = name of part, V = operation acting on part
  - represented as adjectives when there is no corresponding verb
    - A spinner hub and an axial flow fan are gear-driven by the low pressure spool
    - *X gear-drive Y (in corpus)
  - Construction
    - N be A by[agentive] N

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 3 Practicability of Automatic Translation
- 3.1 Formal Grammars for Natural Languages
  - It is precisely when we try to formalize our knowledge of a language that the difficulties begin. Generative grammarians have put an enormous amount of effect into the formalization of rules of grammar.
  - Their lack of success in producing a set of rules that will generate all and only the sentences of a natural language in its entirety hardlys seems encouraging
  - Generative grammarians usually aim only for a description of the “standard language” or the language of an “ideal speaker in an ideal community”

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- Automatic translation from L1 to L2 does not require complete grammars of L1 and L2, only context sensitive transfer rules to obtain the proper lexical items in L2 and some rules for restructuring the resulting string of lexical items in L2.
- Experience at TAUM with a transfer-based approach: even with a very limited corpus, extremely fine grammatical analysis of both languages is required in order to translate 80% of the number of sentences in a text
- The solution seems to lie in restricting one's attention to sublanguages

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 3.2 Text Norms
  - Authors of maintenance manuals, cook books, articles in scientific journals, etc. are generally guided by norms in writing in their particular field
  - Norms do not themselves constitute a grammar, but they do indicate certain regularities not present throughout the whole language, thus simplifying the task of writing formal grammars for texts in specialized fields

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 3.2 Text Norms
  - Claim:
    - Norms + Reduction in polysemy from semantic restrictions + limited vocabulary + syntactic restrictions = practical automatic translation for sub-languages
  - Difficulty of Automatic Translation
    - Are we talking about just a question of scale here?
    - What are the ingredients for success?

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 3.3 TAUM-METEO
  - System for automatic translation of weather reports from English to French
  - Sublanguage has a very small vocabulary
  - Telegraphic (concise) style
  - Syntax is highly restricted
    - no relative clauses or passives, omission of copula, no use of articles, etc.
  - Syntactic analysis depends very much on semantic subcategorization

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- Fixed Format
- (i) place names preceding the forecast
  - RED RIVER
  - INTERLAKE
- (ii) meteorological conditions for the day
  - MAINLY SUNNY TODAY
  - WINDS 25KM PER HOUR
- (iii) statements of maxima and minima
  - HIGHS TODAY 15 TO 18
  - LOWS TONIGHT NEAR 3
- (iv) outlook for next day
  - OUTLOOK FOR THURSDAY …
  - CONTINUING MAINLY SUNNY
- (v) heading of bulletin indicating origin

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 3.4 Idioms (and noun sequences)
  - Parse them or simply list in dictionary?
  - How large a dictionary will be needed?
    - Seems to be no limit
  - Idiom list cannot pre-empty parsing
      - Locate all check points
      - Check points for pitting
    - *Check* as a verb as well as a noun
    - Polysemy of *points* (map onto different words in French)

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- Few word sequences are idioms in all contexts in a sublanguage
    - How to check if word sequence should be interpreted as an idiom?
    - Idiom may be split
        - He acted without malice in spite and because of her threat
        - in spite … of
    - However, unlikely in a maintenance manual to have
        - spite malicious intent sense
        - in spite as a prepositional phrase (PP)

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 3.5 Recognition and Generation
  - Assumption that input is grammatical reduces the problem
    - strategies to locate verb and complements, assigning words to various categories depending on context, assigning constituent structure, etc.
  - Generation is easier from output of parser
    - cf. semantic representations, deep structures, or other abstract objects currently employed in many generative grammars
  - If source sentence can be parsed, it’s a fair bet that the corresponding target sentence can be generated

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 4 The Concept of Sublanguage
- 4.1 Characteristics
  - Sublanguage is not simply an arbitrary subset of the set of sentences of a language
  - Characteristics:
    - (i) limited subject matter
    - (ii) lexical, syntactic and semantic restrictions
    - (iii) deviant rules of grammar
    - (iv) high frequency of certain constructions
    - (v) text structure
    - (vi) use of special symbols

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- (iii) deviant rules of grammar
  - co-occurrence restrictions not present in the standard language
    - *eccentric pilot [-animate]
  - article drop
  - sublanguage grammar is not a subgrammar of the standard language

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 4.2 Cooccurrence and Subcategorization
  - Relations: function, part-of, subject and object
    - installation kit
    - installation procedure
    - installation difficulty
  - Lexical entry: *installation*:
    - abstract:
      - **F** (function) to subclass of nouns to its right
        - » installation kit installation procedure
      - **object** to subclass of nouns to its left
        - » pump installation filter installation

Paper 16. Automatic Translation and the Concept of Sublanguage. Lehrberger, J.

- 4.3 Sublanguage and the Language as a Whole
  - Sublanguages are worthy of study
    - the language of sports-casting
    - the language of biophysics
  - deviant forms can be paraphrased in standard language
    - check reservoir full
    - check to ensure that the reservoir is full