Chapter NLP:I

I. Introduction to Linguistics

- Goals of Language Technology
- Examples of NLP Systems
- NLP Problems
- Linguistic Levels & Terminology
- Historical Background

Linguistic Levels – Example

What language is it? What do you notice?

Vuonna 1982 Nokia toimitti ensimmäisen täysin digitaalisen puhelinkeskuksen Euroopassa ja samana vuonna esittelimme maailman ensimmäisen autopuhelimen analogiselle NMT-standardille. 1980-luvulla kehitetty GSM-standardi mahdollisti korkealaatuiset äänipuhelut, hyödynsi entistä tehokkaammin radiotaajuuksia ja tarjosi korkealaatuisemman äänentoiston. Ensimmäinen GSM-puhelu soitettiin Nokian matkapuhelimella yhtiön Radiolinja-operaattorille rakentamassa verkossa vuonna 1991. . . .

Linguistic Levels – Example

What language is it? What do you notice?

Vuonna 1982 Nokia toimitti ensimmäisen täysin digitaalisen puhelinkeskuksen Euroopassa ja samana vuonna esittelimme maailman ensimmäisen autopuhelimen analogiselle NMT-standardille. 1980-luvulla kehitetty GSM-standardi mahdollisti korkealaatuiset äänipuhelut, hyödynsi entistä tehokkaammin radiotaajuuksia ja tarjosi korkealaatuisemman äänentoiston. Ensimmäinen GSM-puhelu soitettiin Nokian matkapuhelimella yhtiön Radiolinja-operaattorille rakentamassa verkossa vuonna 1991. . . .

What can we see?

- ❑ Characters
- ❑ Letter combinations
- ❑ Wordforms
- ❑ Sentences
- ❑ . . .

Linguistic Levels – Example

What language is it? What do you notice?

Vuonna 1982 Nokia toimitti ensimmäisen täysin digitaalisen puhelinkeskuksen Euroopassa ja samana vuonna esittelimme maailman ensimmäisen autopuhelimen analogiselle NMT-standardille. 1980-luvulla kehitetty GSM-standardi mahdollisti korkealaatuiset äänipuhelut, hyödynsi entistä tehokkaammin radiotaajuuksia ja tarjosi korkealaatuisemman äänentoiston. Ensimmäinen GSM-puhelu soitettiin Nokian matkapuhelimella yhtiön Radiolinja-operaattorille rakentamassa verkossa vuonna 1991. . . .

What can we see?

- Characters
- Letter combinations
- Wordforms
- Sentences
- . . .

- Individual word forms are probably proper names

Linguistic Levels – Example

What language is it? What do you notice?

Vuonna 1982 Nokia toimitti ensimmäisen täysin digitaalisen puhelinkeskuksen Euroopassa ja samana vuonna esittelimme maailman ensimmäisen autopuhelimen analogiselle NMT-standardille. 1980-luvulla kehitetty GSM-standardi mahdollisti korkealaatuiset äänipuhelut, hyödynsi entistä tehokkaammin radiotaajuuksia ja tarjosi korkealaatuisemman äänentoiston. Ensimmäinen GSM-puhelu soitettiin Nokian matkapuhelimella yhtiön Radiolinja-operaattorille rakentamassa verkossa vuonna 1991. . . .

What can we see?

- Characters
- Letter combinations
- Wordforms
- Sentences
- . . .

- Individual word forms are probably proper names
- We also recognize numbers, presumably years

Linguistic Levels – Example (now well understandable)

In 1982, Nokia introduced both the first fully-digital local telephone exchange in Europe and the world?s first car phone for the Nordic Mobile Telephone analog standard. The breakthrough of GSM (global system for mobile communications) in the 1980s introduced more efficient use of radio frequencies and higher-quality sound. The first GSM call was made with a Nokia phone over the Nokia-built network of a Finnish operator called Radiolinjain 1991.

It was around this time that Nokia made the strategic decision to make telecommunications and mobile our core business. Our other businesses, including aluminum, cable, chemicals, paper, rubber, power plant, and television businesses were divested.

By 1998, Nokia was the world leader in mobile phones, a position it enjoyed for more than a decade.And still, the business and technology worlds would continue to evolve, as would Nokia.

Nokia History

Linguistic Levels

Explanans	***Explanandum***
Phonetics	*Sounds (Tokens)* How humans produce and perceive sounds (e.g. Words) Different "sounds" in the same phonetic "environment" distinguish/discriminate between two different words
Phonology	*Groups of sounds:* Phone, Phoneme (meaning-discriminating unit, distinguishing words) e.g. a speech sound in a language
Morphology	*Groups of phonemes:* Morpheme – meaning-bearing unit, minimal unit of grammatical analysis from which words are composed Allomorph – meaning equivalent morphemes e.g. speech, speak, spoken
Lexicon	*Groups of morphemes:* Wordform – inflected form of a word Word, lexeme – Equivalence class of wordforms
Syntax	*Groups of words:* Phrases – valid combination of word forms Sentences – grammatically complete sequence of phrases
Semantics	Proposition (Statement) – true sentence
Pragmatics	Speech act: state-changing

Linguistic Levels

Linguistic levels in the decomposition of a sentence: The sentence "The Gewandhaus zu Leipzig is located at Augustusplatz" consists of

1. A concatenation of letters:
   T–h–e–G–e–w–a–n–d–h–a–u–s–z–u–L–e–i–p–z–i–g–i–s–l–o–c–a–t–e–d–a–t–A–u–g–u–s–t–u–s–p–l–a–t–z

2. A concatenation of morphemes:
   The–Gewand–haus–zu–Leipzig–is–**locat–ed**–at–Augustusplatz

3. A concatenation of word forms:
   The–Gewandhaus–zu–Leipzig–is–located–at–Augustusplatz
   locate_ed (Composition); location, locate (Abstraction)

4. A concatenation of phrases:
   The Gewandhaus zu Leipzig–is located–at Augustusplatz

5. Paragraphs, Texts, Documents

Linguistic Levels

How this insight contributes to methods from Computer Science?

- Features in the sense of machine learning can be found at all linguistic levels
- In addition to the purely "linguistic" features, statistical or pattern-based features are often useful for language processing applications.

Examples

- Uni-, Bi-, Tri-, N-Gram
  `"The", "quick", "brown", "fox", "jumps" "over", "the", "lazy", "dog", "The_quick", "quick_brown", "brown_fox", "fox_jumps", "jumps_over", "over_the", "the_lazy", "lazy_dog", "The_quick_brown" "quick_brown_fox" "brown_fox_jumps", "fox_jumps_over", "jumps_over_the", "over_the_lazy", "the_lazy_dog"`
- (weighted) Word and/or N-Gram occurrences (Usage statistics)
- Co–occurrences, Concordances (Word Combination)
- Metadata, Timestamps

Basic definitions I

1. **Alphabet**
   Let NL be a natural language and let A be a set of characters, $A = \{l_1, l_2, \ldots, l_k\}$. We call A an alphabet of NL of size k.
   Example: $A_E = \{a, b, \ldots, z\}$ $k_E = 26$

2. **String**
   Let $l_1, l_2, \ldots, l_n$ be letters from A. The tuple t with $t =< l_1, l_2, ..., l_n >$ is called a string and n is the length of t.

3. **Set of strings**
   Let A^n be the Cartesian product of the alphabet A. A^n is called the set of strings of length n.

   Example: $A^3 = \{(a, a, a), (a, a, b), \ldots, (a, a, z), (b, a, a), \ldots, (z, z, z)\}$

Basic definitions II

4. **Lexicon of a language**
   Let NL be a natural language and L a subset of A^+.
   $A^+ = \cup_{n>o} A^n)$. We call $L \leq A^+$ a lexicon of NL.

5. **Wordform**, Set of wordforms of length n
   Each element W of the lexicon L is called a (natural) word form. W^n is the intersection of A^n with L and is called the set of word forms of length n.

6. **Token**
   Occurrence of a string (word form) in a text.
   (Total number of tokens in a text = text volume).

7. **Type**
   Equivalence class of identical strings (word forms) in a text.
   (Total number of types in a text = vocabulary range).

How many words

“I like to buy the newspaper from time to time, but I bought it yesterday.”

- Wordform
  - inflected form as it occurs in the text
  - buy and bought ... different wordforms

- Lemma
  - Word forms with the same stem, word category and meaning
  - buy and bought ... share the same lemma (buy)

- Token
  - Actual occurrence of a word form
  - 15 tokens(without punctuation)

- Type
  - Pattern of a token
  - 12 types(without punctuation)

Tokenization

Decomposition of strings (of a language!) into word forms. **See slides about Tokenization in Section Words**

Specifications are required for:

- Special characters
- punctuation
- Word combinations
- Hyphen

NLTK Tokenisierung

```
The quick brown fox jumps over the lazy dog.
```

```
→ "The", "quick", "brown", "fox", "jumps" "over", "the",
"lazy", "dog"
```

Type Token Ratio

Mark Twain's Tom Sawyer
71,370 tokens
8,018 word types
tokens/type Verhältnis= 8.9

Alle Werke Shakespeares
884,647 word tokens
29,066 word types
tokens/type Verhältnis= 30.4

How should this measure be interpreted?

Basic definitions III

8. **Trigrams of a word form**

   Let t be A^+ with $t = (l_1, l_2, \dots l_n)$, 0 = empty element.

   The set T of trigrams from t is the set of 3-tuples such that

$$T = \{< 0, 0, l_1 >, < 0, l_1, l_2 >, < l_1, l_2, l_3 >, < l_2, l_3, l_4 >, \dots, \\ < l_{n-2}, l_{n-1}, l_n >, < l_{n-1}, l_n, 0 >, < l_n, 0, 0 >\}$$

Bigrams

char–ngram	Freq	char–ngram	Freq	char–ngram	Freq
th	2013245	ha	687260	li	472659
in	1802779	et	673609	ic	460445
he	1757953	se	668187	rt	459903
er	1501930	ve	666796	so	429841
an	1485564	ro	655277	fo	426557
re	1383661	le	642876	la	425211
es	1188314	of	616689	il	421744
on	1165201	as	614021	rs	412002
en	1068260	de	561689	di	411161
nd	1060671	si	558970	na	405482
or	1045185	ta	554442	ee	403451
nt	1018046	ra	550079	be	399326
st	1016248	me	544768	ch	394534
to	1009349	ur	541086	ss	385240
ti	977423	sa	539431	ca	385084
at	964675	ne	534519	ns	383171
ou	936776	ll	534219	ac	379258
ea	933603	ec	527795	ho	377851
ng	896029	ri	527121	yo	376484
ar	881728	co	525482	ma	372809
ed	837447	ce	490385	wi	371163
te	811945	io	483325	ot	370311
it	796320	om	479779	tt	355777
al	785150	hi	478751	us	352414
is	745662	el	477841	ts	344573

Trigrams

char–ngram	Freq	char–ngram	Freq	char–ngram	Freq
the	1257720	wit	194521	thi	153208
ing	743278	eth	192176	sta	151584
and	728494	pro	191579	con	150659
ent	432733	sto	191226	tth	149463
ion	424512	ort	190923	ted	147614
for	347474	res	187858	eve	145414
tio	342008	ear	185947	ect	142350
you	291652	sin	185937	sth	138682
ati	276216	tin	184006	out	138153
her	262058	The	179563	eco	137409
our	260975	din	171973	ome	137304
ere	257130	san	171607	hes	136994
tha	252811	ons	170654	ore	135407
est	238535	men	170115	ave	134711
are	237052	ess	169828	ean	134335
ers	227373	ill	166129	rth	134067
nth	226344	ont	164985	per	133078
int	226179	his	163927	dth	132744
rea	219602	oft	162648	ngt	132301
ter	215809	ive	158980	ist	131523
ith	212533	oth	158494	eto	131434
ate	210591	â??	158121	oun	131370
ver	209986	nce	157934	ide	131017
all	197738	com	156796	eof	127857
hat	197671	fth	153966	edt	127309

Künstliche Sprache (Kupfmüller)

Diphone, Triphone

Einergruppen (Buchstabenhäufigkeit)

EME GKNEET ERS TITBL BTZENFNDGBGD EAI E LASZ
BETEATR IASMIRCH EGEOM

Zweiergruppen (Paarhäufigkeit)

AUSZ KEINU WONDINGLIN DUFRN ISAR STEISBERER ITEHM
ANORER

Dreiergruppen

PLANZEUNDGES PHIN INE UNDEN ÜBBEICHT GES AUF ES SO
UNG GAN DICH WANDERSO

Vierergruppen

ICH FOLGEMÄSZIG BIS STEHEN DISPONIN SEELE NAMEN

String Similarities

- N-gram similarity often interesting feature

- Useful for e.g.
  - Spell check
  - citation and plagiarism detection
  - (semi-)automatic learning of linguistic structures

- Methods (examples – See Section Similarities in Text Models for details)
  - Number of identical trigrams, e.g. Dice coefficient

$$d_w(a, b) = \frac{2|T(a) \cap T(b)|}{|T(a)| + |T(b)|}$$

  - Editing distance, cost of transforming one string into another, e.g. Levenshtein matrix [Wikipeda Entry]

Basic definitions IV

9. **Substring**

   Let t, u be A^+ strings with $t = (t_1, t_2, \ldots, t_n)$ and $u = (u_1, u_2, \ldots u_m)$.

   We call u a substring of t, if $1 \leq m \leq n$ and there is an i and a j such that $u = t_{i\ to\ j}$ for all $1 \leq j \leq m$

   Example: "nun" is substring of "pronunciation".

   `p r o n u n c i a t i o n`

   i= 1 2 3 **4 5 6** 7 8 9 10 11 12 13

   j= 1 2 3

Basic definitions V

10. **Word form combinations of length r**
    Let L be a tuple of word forms, $L = (W_1, W_2, ...W_r)$ with $W_i \in L$. We call L a word form combination of length r.

11. **Set of wordform combinations**
    If L^r is the Cartesian product of L. L^+ is called a set of wordform combinations of length r. $(L^+ = \cup_{r>o} L^r)$

12. **Set of sentences**
    Let SYN be a set of syntactic restrictions. The set S, with $S \leq L^+$ following SYN, is called the set of sentences.

13. **Word**
    Equivalence class of morphologically related word forms.

14. **Concept**
    equivalence class of semantically related words (e.g. global context)

Acquisition of linguistic knowledge

Given a possibility space of N-grams (letters and word forms) of a language:

- Determine the lexicon of word forms
- Determine the lexicon of words
- Determine the set of syntactic restrictions SYN

Why does natural language processing require linguistic knowledge?

Natural languages are not static like formal languages, but develop according to their own laws and dynamics.

- Expressions in natural languages can be ambiguous
  - Lexical
  - Structural (word, phrase, sentence)
- Texts are subject to their own laws of linguistic statistics
  - Zipf's law
- Texts reflect the linguistic dynamics of a language
  - Neologisms, extinction of forms, regional or milieu-related usages

What means language in this sense?

Language vs. Language ability

- **Language:** English, Finnish, German, . . .
  Empirically, we find *realizations* (utterances, texts) of single languages.
  People possess *knowledge* of one (or more) single languages.

- **Language ability:** Possibility specific to humans to learn (understand, use) a single language.
  Language in the sense of linguistic ability: *abstract system*

Singular Languages

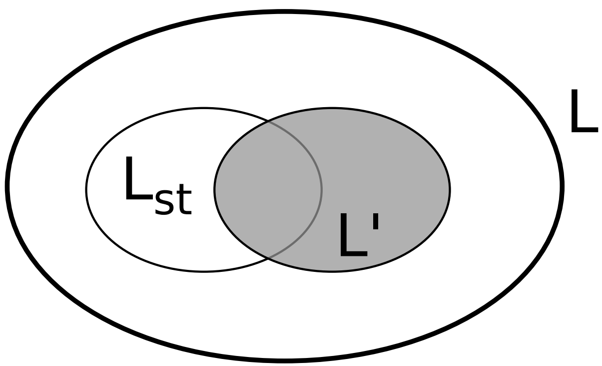
Individual languages and their realizations

- Everyday language
- newspapers
- scientific essays, reference books, encyclopedias
- Internet
- correspondence and email
- technical documentation, user manuals and product descriptions
- standards, laws, commentaries and contracts
- organizational instructions
- etc.

Sub Languages

[Z. Harris (1968)]: Subset of structures that can be generated by the language system.

- syntactic and semantic constraints
  - deviant grammar
  - high probability of certain constructions
- lexical constraints
  - medicine, weather reports, legal texts, technical instructions, ...
- characteristic morphemes
- – medicine, chemistry, technical manuals



Example – the weather report

Example – the weather report

With depression “Fritz” over the Bay of Biscay, humid but gradually milder air reaches Central Europe. In Spain and France, the depression brings heavy downpours. Over the Baltic States and in large parts of Scandinavia, High "Birgit" ensures dry but cold winter weather. Another low-pressure area with rain lies in the eastern Mediterranean.

Example – the weather report

Example – the weather report

With depression “Fritz” over the Bay of Biscay, humid but gradually milder air reaches Central Europe. In Spain and France, the depression brings heavy downpours. Over the Baltic States and in large parts of Scandinavia, High "Birgit" ensures dry but cold winter weather. Another low-pressure area with rain lies in the eastern Mediterranean.

What can we see?

- Syntactic and semantic constraints
- Lexical constraints

Example – the weather report

Example – the weather report

With depression “Fritz” over the Bay of Biscay, humid but gradually milder air reaches Central Europe. In Spain and France, the depression brings heavy downpours. Over the Baltic States and in large parts of Scandinavia, High "Birgit" ensures dry but cold winter weather. Another low-pressure area with rain lies in the eastern Mediterranean.

What can we see?

- Syntactic and semantic constraints
- Lexical constraints

Example – the weather report

Example – the weather report

With depression “Fritz” over the Bay of Biscay, humid but gradually milder air reaches Central Europe. In Spain and France, the depression brings heavy downpours. Over the Baltic States and in large parts of Scandinavia, High "Birgit" ensures dry but cold winter weather. Another low-pressure area with rain lies in the eastern Mediterranean.

What can we see?

- Syntactic and semantic constraints
- Lexical constraints

Properties of Sub Languages

1. SL (= Sublanguage) form thematic groups
   - Constant lexicons
   - Constant syntax

2. SL features can remain identical across different languages.
   - Passive voice in technical instructions
   - Frequencies of sentence structures and terms
   - Omission of the article

3. SL Characteristics are gradable and change
   - [N. Sager 1967], “LinguisticString Project”

Language registers

Language registers refer to linguistic variations that correlate with different *situations* of language use

Different situations

1. Channel factors (e.g. mobile phone, Internet, telephone,...) restricting the mode and quantity of linguistic action

2. Different social roles requiring specific modes of linguistic action (e.g. politeness)

Linguistic variations

1. Constraints

2. Amendmends
   - Syntax
   - Lexicon

Example Computer Talk

Example – Language register *Computer Talk*

- When people communicate with a computer system they adapt their language register assuming that the system will require “simpler” input

- Computer talk comparable to *baby talk* or *foreigner talk*

- Empirical validation (Krause et. al. – Computer Talk (9-783-487-095-691)) using simulations of 4 systems:

System 1: optimal human-human information system
System 2: optimal human-computer IS
System 3: restricted human-computer IS
System 4: strongly restricted human-computer IS

Example Computer Talk

Computer Talk – Share of 10 most frequent sentence patterns at total input

	Typed input	***Spoken input***
S_1	63,04 %	35,75 %
S_2	88,57 %	77,12 %
S_3	97,41 %	82,14 %
S_4	99,26 %	99,24 %

Organon Model

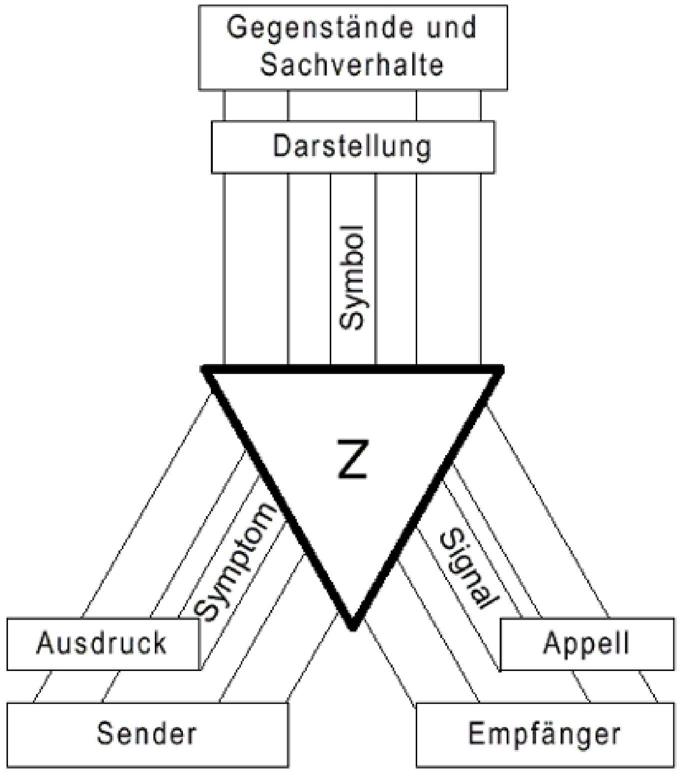
The organon model of Bühler 1934 [Wikipedia] (Language Theory – ISBN 3-8252-1159-2)

- The basis of linguistic communication are linguistic expressions

- Linguistic expressions have three dimensions:
  - the sender (speaker, writer)
  - the receiver (listener, reader)
  - the referenced thing (objects and events, properties, facts, . . .)

- In relation to a sender, (intended) receiver and the referenced thing, linguistic expressions therefore have a threefold function:
  - the expressive function (Ausdrucksfunktion, Symptom)
  - the conative function (Appellfunktion, i.e. appealing function).
  - the representation function (Darstellungsfunktion, Symbol)

Organon Model – Scheme



Linguistic Aspects

We therefore find linguistic expressions in the form of

- the expressive function (Ausdrucksfunktion, Symptom)
- the conative function (Appellfunktion, i.e. appealing function).
- the representation function (Darstellungsfunktion, Symbol)

Example (The Big Lebowski):

Linguistic Aspects

Example (The Big Lebowski – Script): *It is the directors job to make up a perfect mix of Expression, Conative(ness) and Representation*

...The paper bag hugged to his chest explodes milk as it hits the toilet rim and the satchel pulverizes tile as it crashes to the floor. Context – Stage instructions

The Dude blows bubbles. Context – Acting instructions

VOICE

We want that money, Lebowski. Bunny said you were good for it.

Hands haul the Dude out of the toilet. The Dude blubbers and gasps for air. Context – Acting instructions

VOICE

Where's the money, Lebowski!

His head is plunged back into the toilet. Context – Acting instructions

VOICE

Where's the money, Lebowski!

The hands haul him out again, dripping and gasping.

VOICE

WHERE'S THE FUCKING MONEY, SHITHEAD! Expression, Appealing

DUDE

It's uh, it's down there somewhere. Lemme take another look. Expression, Appealing – **Humor???, Sarcasm???**

His head is plunged back in. Context – Acting instructions

VOICE

Don't fuck with us. ...

[Original Scipt]

Literature

Biemann, Heyer, Quasthoff; **Wissensrohstoff Text - Eine Einführung in das Text Mining** , Springer Vieweg, 2022.

Krause, Hitzenberger, **Computer Talk**, Olms 1992

Bühler, Karl; **The Theory of Language: The Representational Function of Language (Sprachtheorie)**, UTB, 1934/1990

Küpfmüller, K.: **Die Entropie der deutschen Sprache**. Fernmeldetechnische Zeitung 7, 1954, S. 265-272.

Harris, Z. **Mathematical structures of language**. Interscience tracts in pure and applied mathematics., 1968