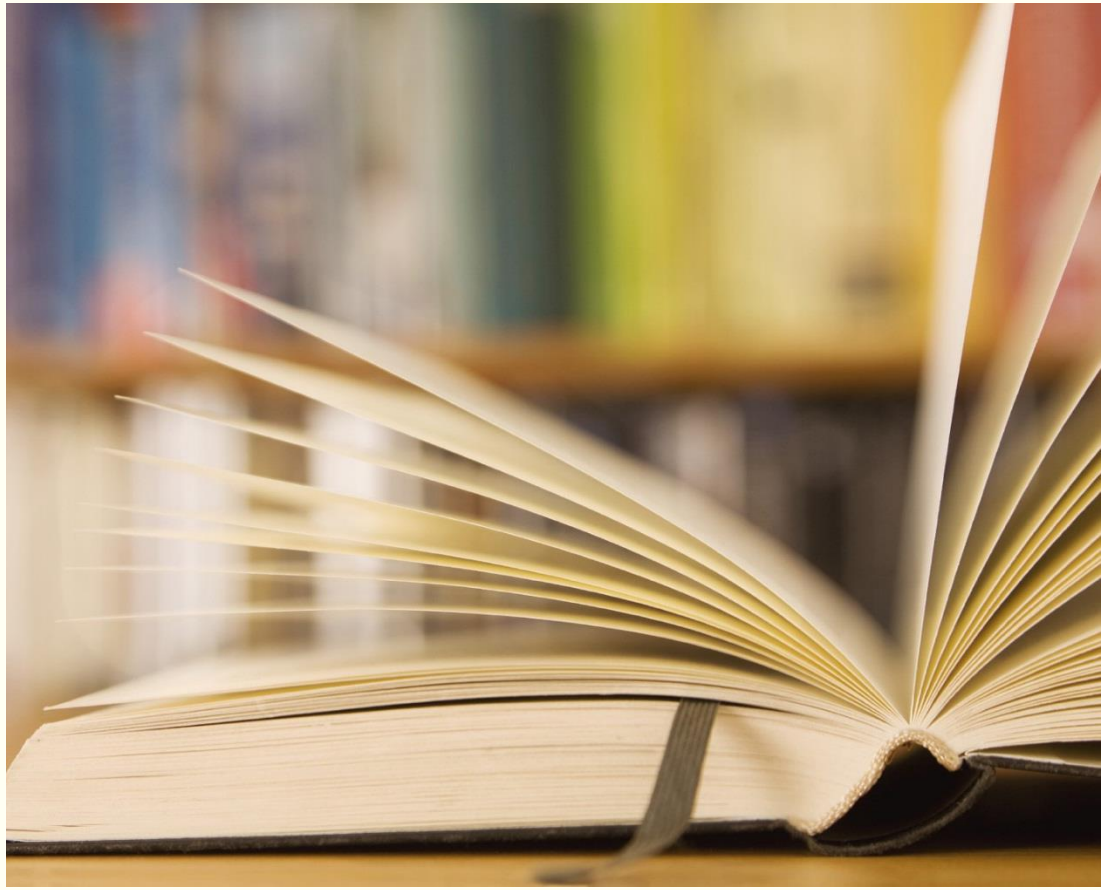
NATURAL LANGUAGE PROCESSING

Dr. G. Bharadwaja Kumar



NATURAL?

- In computing, a **character encoding** is used to represent a repertoire of characters by some kind of encoding system.

ASCII

Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040	&#32;	Space	64	40	100	&#64;	@	96	60	140	&#96;	`
1	1	001	SOH (start of heading)	33	21	041	&#33;	!	65	41	101	&#65;	A	97	61	141	&#97;	a
2	2	002	STX (start of text)	34	22	042	&#34;	"	66	42	102	&#66;	B	98	62	142	&#98;	b
3	3	003	ETX (end of text)	35	23	043	&#35;	#	67	43	103	&#67;	C	99	63	143	&#99;	c
4	4	004	EOT (end of transmission)	36	24	044	&#36;	$	68	44	104	&#68;	D	100	64	144	&#100;	d
5	5	005	ENQ (enquiry)	37	25	045	&#37;	%	69	45	105	&#69;	E	101	65	145	&#101;	e
6	6	006	ACK (acknowledge)	38	26	046	&#38;	&	70	46	106	&#70;	F	102	66	146	&#102;	f
7	7	007	BEL (bell)	39	27	047	&#39;	'	71	47	107	&#71;	G	103	67	147	&#103;	g
8	8	010	BS (backspace)	40	28	050	&#40;	(	72	48	110	&#72;	H	104	68	150	&#104;	h
9	9	011	TAB (horizontal tab)	41	29	051	&#41;	)	73	49	111	&#73;	I	105	69	151	&#105;	i
10	A	012	LF (NL line feed, new line)	42	2A	052	&#42;	*	74	4A	112	&#74;	J	106	6A	152	&#106;	j
11	B	013	VT (vertical tab)	43	2B	053	&#43;	+	75	4B	113	&#75;	K	107	6B	153	&#107;	k
12	C	014	FF (NP form feed, new page)	44	2C	054	&#44;	,	76	4C	114	&#76;	L	108	6C	154	&#108;	l
13	D	015	CR (carriage return)	45	2D	055	&#45;	-	77	4D	115	&#77;	M	109	6D	155	&#109;	m
14	E	016	SO (shift out)	46	2E	056	&#46;	.	78	4E	116	&#78;	N	110	6E	156	&#110;	n
15	F	017	SI (shift in)	47	2F	057	&#47;	/	79	4F	117	&#79;	O	111	6F	157	&#111;	o
16	10	020	DLE (data link escape)	48	30	060	&#48;	0	80	50	120	&#80;	P	112	70	160	&#112;	p
17	11	021	DC1 (device control 1)	49	31	061	&#49;	1	81	51	121	&#81;	Q	113	71	161	&#113;	q
18	12	022	DC2 (device control 2)	50	32	062	&#50;	2	82	52	122	&#82;	R	114	72	162	&#114;	r
19	13	023	DC3 (device control 3)	51	33	063	&#51;	3	83	53	123	&#83;	S	115	73	163	&#115;	s
20	14	024	DC4 (device control 4)	52	34	064	&#52;	4	84	54	124	&#84;	T	116	74	164	&#116;	t
21	15	025	NAK (negative acknowledge)	53	35	065	&#53;	5	85	55	125	&#85;	U	117	75	165	&#117;	u
22	16	026	SYN (synchronous idle)	54	36	066	&#54;	6	86	56	126	&#86;	V	118	76	166	&#118;	v
23	17	027	ETB (end of trans. block)	55	37	067	&#55;	7	87	57	127	&#87;	W	119	77	167	&#119;	w
24	18	030	CAN (cancel)	56	38	070	&#56;	8	88	58	130	&#88;	X	120	78	170	&#120;	x
25	19	031	EM (end of medium)	57	39	071	&#57;	9	89	59	131	&#89;	Y	121	79	171	&#121;	y
26	1A	032	SUB (substitute)	58	3A	072	&#58;	:	90	5A	132	&#90;	Z	122	7A	172	&#122;	z
27	1B	033	ESC (escape)	59	3B	073	&#59;	;	91	5B	133	&#91;	[	123	7B	173	&#123;	{
28	1C	034	FS (file separator)	60	3C	074	&#60;	<	92	5C	134	&#92;	\	124	7C	174	&#124;	|
29	1D	035	GS (group separator)	61	3D	075	&#61;	=	93	5D	135	&#93;	]	125	7D	175	&#125;	}
30	1E	036	RS (record separator)	62	3E	076	&#62;	>	94	5E	136	&#94;	^	126	7E	176	&#126;	~
31	1F	037	US (unit separator)	63	3F	077	&#63;	?	95	5F	137	&#95;	_	127	7F	177	&#127;	DEL

Source: www.LookupTables.com

ISCII

Hex	Hex Dec	0 0	1 16	2 32	3 48	4 64	5 80	6 96	7 112	8 128	9 144	A 160	B 176	C 192	D 208	E 224	F 240
0	0	NUL	DLE	SP	0	@	P	`	p				ओ	ढ	ॲ	ॅ	EXT
1	1	SOH	DC1	!	1	A	Q	a	q			ॅ	औ	ण	ल	ॆ	०
2	2	STX	DC2	"	2	B	R	b	r			ॆ	ऑ	त	ळ	े	१
3	3	ETX	DC3	#	3	C	S	c	s			ः	क	थ	क्ष	ॉ	२
4	4	EOT	DC4	$	4	D	T	d	t			अ	ख	द	ज्ञ	ॊ	३
5	5	ENQ	NAK	%	5	E	U	e	u			आ	ग	ध	श्र	ो	४
6	6	ACK	SYN	&	6	F	V	f	v			इ	घ	न	ष	ौ	५
7	7	BEL	ETB	'	7	G	W	g	w			ई	ङ	ऩ	स	्	६
8	8	BS	CAN	(	8	H	X	h	x			उ	च	प	ह	़	७
9	9	HT	EM	)	9	I	Y	i	y			ऊ	छ	फ	INV	़	८
A	10	LF	SUB	*	:	J	Z	j	z			ऋ	ज	ब	ा	ा	९
B	11	VT	ESC	+	;	K	[	k	{			ऌ	झ	भ	ि		
C	12	FF	FS	,	<	L	\	l	|		ऍ	ञ	म	ी		
D	13	CR	GS	-	=	M	]	m	}			ऎ	ट	य	ु		
E	14	SO	RS	.	>	N	^	n	~			ए	ठ	र	ू		
F	15	SI	US	/	?	O	_	o	DEL			ऐ	ड	ऱ	ृ	ATR	

What is a Character?

- Indian Languages:
  - No ‘alphabet’, not letters, no spellings
  - Phoneme-based
  - Units are syllable-like: called ‘akshara’-s
- akshara-s very large in number
  - Corpus studies not sufficient
- Made up of vowels, consonants etc.
- Not all sequences valid

Character Encoding

Any file has to go through encoding/decoding in order to be properly stored as file or displayed on screen.

Your computer needs a way to translate the character set of your language's writing system into a sequence of 1s and 0s.

This transformation is called Character encoding.

- **ASCII** is a 7-bit encoding based on the English alphabet.
- **8-bit encodings** are extensions to ASCII that add a potpourri of useful, non-standard characters like é and æ. They can only add 127 characters, so usually only support one script at a time. When you see a page on the web, chances are it's encoded in one of these encodings.

Encoding Systems

- There are many encoding systems. The most popular encoding systems used today are:
  - ASCII. For English. Most widely used before year 2000.
  - UTF-8 of Unicode (used in Linux by default, and much of the Internet)
  - UTF-16 of Unicode (used by Microsoft Windows and Mac OS X's file systems, Java programing language, ...)
  - GB 18030 (Used in China, contains all Unicode chars).
  - EUC (Extended Unix Code). Used in Japan.
  - IEC 8859 series (used for most European langs)

- **Unicode-based encodings** implement the Unicode standard and include UTF-8, UTF-16 and UTF-32/UCS-4.

- They go beyond 8-bits and support almost every language in the world.

- UTF-8 is gaining traction as the dominant international encoding of the web.

- Unicode defines several encoding system. UTF-8 and UTF-16 are the two most popular Unicode encoding systems.

Unicode

- Unicode's character set includes ALL human language's written symbols.

- Each character in Unicode is given a unique ID. This id is a number (integer), and is called the char's code point.

- For example, the code point for the greek alpha α char is 945. In hexadecimal it's “3b1”. In the standard Unicode notation it is written as “U+03B1”.

Fonts

- A font matches a number in encoding system to a text character which is displayed on a monitor or printed out.
- On a computer, the **keyboard** is the usual device that allows a user to input the numeric codes that are translated as text by the font.
- Different fonts allow for different character styles to be displayed. Some example fonts:

- **Times New Roman**
  Georgia
  Arial
  Arial Black
  Verdana

Orthography -- Word & Sentence Segmentation

- When a computer has decoded a file, it then needs to display the characters as glyphs on the screen.
- This set of glyphs is a font.
- computer needs to map the Unicode code points to a font.

- **WX notation** is a transliteration scheme for representing Indian languages in ASCII. This scheme originated at IIT Kanpur for computational processing of Indian languages, and is widely used among the natural language processing (NLP) community in India.

Vowels [edit]

अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ
a	A	i	I	u	U	e	E	o	O

Sonorants [edit]

ऋ	ॠ	ऌ
q	Q	L

Anusvāra and visarga [edit]

अं	अः
M	H

Anunasika is represented by 'z'. For example, अँ = az.

Consonants [edit]

क्	ख्	ग्	घ्	ङ्	Velar
k	K	g	G	f	
च्	छ्	ज्	झ्	ञ्	Palatal
c	C	j	J	F	
ट्	ठ्	ड्	ढ्	ण्	Retroflex
t	T	d	D	N	
त्	थ्	द्	ध्	न्	Dental
w	W	x	X	n	
प्	फ्	ब्	भ्	म्	Labial
p	P	b	B	m	
य्	र्	ल्	व्		Semi-vowel
y	r	l	v		
श्	ष्	स्	ह्		Fricative
S	R	s	h		

"A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language"

(Sinclair 1996)

What is a CORPUS?

"the term *corpus* as used in modern linguistics can best be defined as a collection of sampled texts, written or spoken, in machine-readable form which may be annotated with various forms of linguistic information"

(McEnery, Xiao and Tono 2006)

What is a corpus?

A corpus can be defined as a *collection of texts* assumed to be *representative* of a given language put together so that it can be used for *linguistic analysis*.

Usually the assumption is that the language stored in a corpus is *naturally-occurring*, that is gathered according to *explicit design criteria*, with a specific purpose in mind, and with a claim to represent natural chunks of language *selected according to specific typology*

What is a corpus for?

- A corpus is made for the study of language in a broad sense
  - To test existing linguistic theory and hypotheses
  - To generate and verify new linguistic hypotheses
  - Beyond linguistics, to provide textual evidence in text-based humanities and social sciences subjects

- The purpose is reflected in a well-designed corpus

Why use corpora?

- Even expert speakers have only a partial knowledge of a language
  - **A corpus can be more comprehensive and balanced**
- Even expert speakers tend to notice the unusual and can't think of what is possible
  - **A corpus can show us what is common and typical**
- Even expert speakers cannot quantify their knowledge of language
  - **A corpus can readily give us accurate statistics**

Why use corpora?

- Even expert speakers cannot remember everything they know
  - **A corpus can store and recall all the information that has been stored in it**
- Even experts speakers cannot make up natural examples
  - **A corpus can provide us with a vast number of examples in real communication context**
- Even expert speakers have prejudices and preferences and every language has cultural connotations and underlying ideology
  - **A corpus can give you more objective evidence**

Why use corpora?

- Even expert speakers are not always available to be consulted
  - **A corpus can be made permanently accessible to all**
- Even expert speakers cannot keep up with language change
  - **A constantly updated corpus can reflect even recent changes in the language**
- Even expert speakers lack authority: they can be challenged by other expert speakers
  - **A corpus can encompass the actual language use of many expert speakers**

Benefits of corpus data

- Corpus data is more reliable
  - A corpus pools together linguistic intuitions of a range of language speakers, which offsets the potential biases in intuitions of individual speakers
- Corpus data is more natural
  - It is used in real communications instead of being invented specifically for linguistic analysis
- Corpus data is contextualized
  - Attested language use which has already occurred in real linguistic context
- Corpus data is quantitative
  - Corpora can provide frequencies and statistics readily
- Corpus data can find differences that intuitions alone cannot perceive
  - E.g. synonyms *totally, absolutely, utterly, completely, entirely*

What corpora cannot do

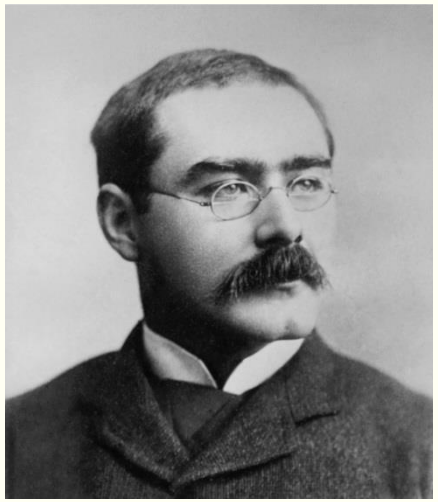
- Corpora do not provide negative evidence
  - Cannot tell us what is possible or not possible
  - Can show what is central and typical in language
- Corpora can yield findings but rarely provide explanations for what is observed
  - Interfacing other methodologies
- The use of corpora as a methodology also defines the boundaries of any given study
  - Importance of amenable research questions
- The findings based on a particular corpus only tell us what is true in that corpus
  - Generalisation vs. representativeness

Corpus-based learner dictionaries

- First 'fully corpus-based' dictionary
  - *Collins Cobuild English Dictionary* (1987)
- Some corpus-based learner dictionaries
  - *Longman Dictionary of Contemporary English* (3rd edition)
  - *Oxford Advanced Learner's Dictionary* (OALD, 5th edition)
  - *Cambridge International Dictionary of English* (1st edition)

“Words are, of course, the most powerful drug used by mankind.”

— Rudyard Kipling



Representativeness

Essential feature of a corpus.

Balance (*the range of genres included in a corpus*) and sampling (*how the text chunks for each genre are selected*) ensure representativeness.

Representativeness

A corpus is representative if…

…the findings based on its contents cane be generalized to the said language variety (Leech 1991);

…its samples include the full range of variability in a population (Biber 1993)

It changes over time (Hunston 2002): if a corpus is not regularly updated, it rapidly becomes unrepresentative.

Representativeness

2 main types (for the range of text categories represented):

- General corpora – a basis for an overall description of a language (variety); their representativeness depends on the sampling from a broad range of genres.
- Specialized corpora – domain- or genre specific corpora; their representativeness can be measured by the degree of closure or saturation (lexical features).

Balance

The range of text categories included in the corpus:

The acceptable balance is determined by the intended uses.

A balanced corpus covers a wide range of text categories which are supposed to be representative of the language (variety) under consideration.

Sampling

A corpus is a sample of a given population

A sample is representative if what we find for the sample holds for the general population

Samples are scaled-down versions of a larger population

Sampling

Sampling techniques:

- Simple random sampling: all sampling units within the sampling frame are numbered and the sample is chosen by use of a table or random numbers; rare features could not be accounted for.
- Stratified random sampling: the population is divided in relatively homogeneous groups, i.e. the strata, and then these latter are sampled at random; never less representative than the former method.

Sampling

Proportion and number of samples:

The number of samples across text categories should be proportional to their frequencies and/or weights in the target population in order for the resulting corpus to be considered as representative

Data collection

Spoken data must be transcribed from audio recordings.

Written text must be rendered machine-readable by keyboarding or OCR (Optical Character Recognition) scanning.

Language data so collected form a RAW CORPUS.

Corpus Mark-up

System of standard codes inserted into a document stored in electronic form to provide information about the text itself and govern formatting, printing and other processes.

Most widely used mark-up schemes:

- TEI (Text Encoding Initiative)
- CES (Corpus Encoding Standard)

Corpus Mark-up

It is essential in corpus-building because…

…sampled texts are out of context and it allows to recover contextual information

…it provides more information than the file names alone (recover text types, sociolinguistic variables, textual information – structure)

…it ads value to the corpus because it allows for a broader range of questions
to be addressed

…it allows to insert editorial comments during the corpus building process.

Corpus Mark-up

Extra-textual and textual information must be kept **separate** from the corpus data.

Examples:

COCOA mark-up scheme

<A WILLIAM SHAKESPEARE>

A= author, *attribute name*

WILLIAM SHAKESPEARE= *attribute value*

TEI Mark-up Scheme

Each individual text is a *document* consisting in a *header* and a *body*, in turn composed of different *elements*.

Ex. in the header there are 4 main elements:

- A file description <fileDesc>
- An encoding description <encodingDesc>
- A text profile <profileDesc>
- A revision history <revisionDesc>

Tags can be nested, i.e. they can appear inside other elements.

TEI Mark-up Scheme

It can be expressed using a number of different formal languages.

SGML (Standard Generalized Mark-up Language – used by the BNC)

XML (Extensible Mark-up Language)

Corpus Annotation

Necessary in order to extract relevant information from corpora.

“The process of adding interpretive, linguistic information to an electronic corpus of spoken and/or written language data”

(Leech 1997)

Annotation vs. Mark-up

Corpus mark-up provides objective, verifiable information.

Annotation is concerned with interpretive linguistic information.

The advantages of annotation

1. It makes extracting information easier, faster and enables human analysts to exploit and retrieve analyses of which they are not themselves capable.
2. Annotated corpora are reusable resources.
3. Corpus annotation records a linguistic analysis explicitly.
4. Corpus annotation provides a standard reference resource, a stable base of linguistic analyses, so that successive studies can be compared and contrasted on a common basis.

How are corpora annotated?

- Automatic annotation
- Computer-assisted annotation
- Manual annotation

Sinclair (1992): the introduction of the human element in corpus annotation reduces consistency.

Types of annotation

Corpora can be annotated at different levels of linguistic analysis.

- Phonological level
  - Syllable boundaries (phonetic/phonemic annotation)
  - Prosodic features (prosodic annotation)

- Morphological level
  - Prefixes
  - Suffixes
  - Stems

  (morphological annotation)

Types of annotation

- Lexical level
  - Part of speech (POS Tagging)
  - Lemmas (lemmatization)
  - Semantic fields (semantic annotation)
- Syntactic level
  - parsing
  - treebanking
  - bracketing

Types of annotation

- Discourse level
  - Anaphoric relations (coreference annotation)
  - Speech acts (pragmatic annotation)
  - Stylistic features such as speech and thought in presentation (stylistic annotation).

POS Tagging

POS is the most common type of annotation.

Also known as grammatical tagging or morpho-syntactic annotation.

It provides the basis of further forms of analysis such as parsing and semantic annotation.

Many linguistic analyses, e.g. the collocates of a word depend heavily on POS tagging.

POS Tagging

It can be performed automatically with taggers like CLAWS

http://www.comp.lancs.ac.uk/ucrel/claws/

You can try it for free online.

Examples of tags: NN1 (noun), VVZ (verb in the third person of the simple present tense), VVD (verb in the simple past form), ADJ0 (adjective in the basic form), etc.

POS Tagging

Problems:

- Word segmentation (tokenization)
  - Multiwords (so that, inspite of)
  - Mergers (can’t, gonna)
  - Variably spelled compounds (noticeboard, notice-board, notice board)

Lemmatization

Type of annotation that reduces the inflextional variants of words to their respective lexemes or lemmas as they appear in dictionary entries:

Do, does, did, done, doing= DO

Corpus, corpora= CORPUS

Small capital letters are the convention.

Lemmatization

It is important in vocabulary studies and lexicography, e.g. in studying the distribution pattern of lexemes and improving dictionaries and computer lexicons.

It can be automatically performed.

Parsing

Once a corpus is POS tagged, it is possible to bring these morpho-syntactic categories into higher level syntactic relationships with one another, that is, to analyse the sentences in a corpus into their constituents.

Parsing consists in bracketing.

It can be automated but with a low precision rate.

Parsing

Example:

```
(S    (NP    Mary)
        (VP      visited)
                   (NP    a
                            (ADJP very nice)
                            boy)))
```

Semantic annotation

It assigns codes indicating the semantic features of the semantic fields of the words in a text. It is knowledge-based so it needs to be manual most of the time.

Two types:

- One marks the semantic relationships between the constituents in a sentence
- One marks the semantic features of words in a text

Coreference annotation

- Pronouns
- Repetition
- Substitution
- Ellipsis

Computer-assisted at best.

Pragmatic annotation

- Speech/dialogue acts in domain-specific dialogue.

  The most coherent system is DRI (Discourse Representation Initiative).

  3 layers of coding:
  - Segmentation (dividing dialogue in textual units, utterances)
  - Functional annotation (dialogue act annotation)
  - Utterance tags (applying utterance tags that characterize the role of the utterance as a dialogue act)

Pragmatic annotation

Utterance tags:

- Communicative status (intelligible, complete, etc.)
- Information level and status (indicating the semantic content of the utterance and how it relates to the task in question)
- Forward-looking communicative function (utterances that may constrain or affect the discourse, e.g. assert, request, question and offer)
- Backwarding-looking communicative function (utterances that relate to previous parts of the discourse, e.g. accept, backchannelling, answer)

Types of corpora

- Multilingual
- Monolingual

Multilingual Corpora

- Parallel corpora (source texts plus translations): Canadian Hansard
- Comparable corpora (monolingual subcorpora designed using the same sampling techniques): Aahrus corpus of contract law
  - Multilingual
  - Bilingual

Multilingual Corpora

Important resources for translation and contrastive studies.

Multilingual corpora…

- …give new insight into the language compared
- …can be used to study language specific and universal features
- …illuminate differences between source texts and translations
- …can be used for a number of practical applications, in lexicography, language teaching, translation, etc.

Parallel Corpora

- Bilingual vs.Multilingual
- **Unidirectional** (from L_a to L_b or from L_b to L_c alone) vs. **Bidirectional** (from L_a to L_b and from L_b to L_a) vs. **Multidirectional** (from L_a to L_b, L_c etc.)

Comparable corpora

A corpus containing components that are collected using the same sampling techniques and similar balance and representativeness, e.g. the *same proportions* of the texts of the *same genres* in the *same domains* in a range of *different languages* in the *same sampling period*.

Comparable vs. parallel corpora

The sampling frame is essential for comparable corpora but not for parallel corpora because the texts are exact translations of
each other.

Corpus Alignment

In order for us to be able to fully exploit parallel corpora, they need to be aligned.

Different types of alignment:

- Word-level alignment
- Sentence-level alignment
- Paragraph alignment