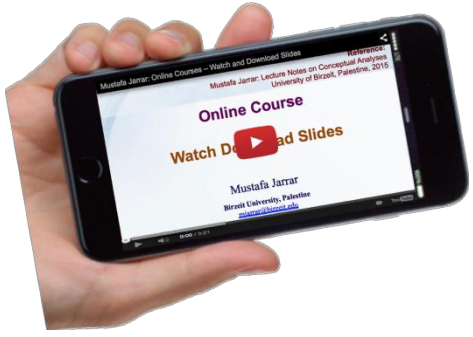
Natural Language Processing

Character Encoding

Mustafa Jarrar

Watch this lecture and download the slides



Course Page: http://www.jarrar.info/courses/NLP/

More Online Courses at: http://www.jarrar.info

Natural Language Processing
Character Encoding

In this lecture:

- Part 1: **Encoding Schemes**
- Part 2: Locales and Word Order

ASCII

- **American Standard Code for Information Interchange** (published 1963)
- based on the English alphabet, encodes 128 characters into seven-bit integers

- 95 printable chars
  0-9 a-z A-Z …

- 33 non-printing control chars
  return, tab, escape, …
  Most are obsolete - for old typers

- **For example**, lowercase (i):
  1101001 binary
  69 hexadecimal
  105 decimal

b_7 b_6 b_5 → / Bits / b_4 b_3 b_2 b_1 ↓ / Row ↓ / Column →	0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
	0	1	2	3	4	5	6	7
0 0 0 0	0	NUL	DLE	SP	0	@	P	`	p
0 0 0 1	1	SOH	DC1	!	1	A	Q	a	q
0 0 1 0	2	STX	DC2	"	2	B	R	b	r
0 0 1 1	3	ETX	DC3	#	3	C	S	c	s
0 1 0 0	4	EOT	DC4	$	4	D	T	d	t
0 1 0 1	5	ENQ	NAK	%	5	E	U	e	u
0 1 1 0	6	ACK	SYN	&	6	F	V	f	v
0 1 1 1	7	BEL	ETB	'	7	G	W	g	w
1 0 0 0	8	BS	CAN	(	8	H	X	h	x
1 0 0 1	9	HT	EM	)	9	I	Y	i	y
1 0 1 0	10	LF	SUB	*	:	J	Z	j	z
1 0 1 1	11	VT	ESC	+	;	K	[	k	{
1 1 0 0	12	FF	FS	,	<	L	\	l	|
1 1 0 1	13	CR	GS	-	=	M	]	m	}
1 1 1 0	14	SO	RS	.	>	N	^	n	~
1 1 1 1	15	SI	US	/	?	O	_	o	DEL

Extended ASCII

eight-bit or larger character encodings - includes the 7-bit chars, and more

128	Ç	144	É	160	á	176	▒	192	└	208	╨	224	α	240	≡
129	ü	145	æ	161	í	177	▒	193	┴	209	╤	225	ß	241	±
130	é	146	Æ	162	ó	178	▓	194	┬	210	╥	226	Γ	242	≥
131	â	147	ô	163	ú	179	│	195	├	211	╙	227	π	243	≤
132	ä	148	ö	164	ñ	180	┤	196	─	212	╘	228	Σ	244	⌠
133	à	149	ò	165	Ñ	181	╡	197	┼	213	╒	229	σ	245	⌡
134	å	150	û	166	ª	182	╢	198	╞	214	╓	230	µ	246	÷
135	ç	151	ù	167	º	183	╖	199	╟	215	╫	231	τ	247	≈
136	ê	152	ÿ	168	¿	184	╕	200	╚	216	╪	232	Φ	248	°
137	ë	153	Ö	169	⌐	185	╣	201	╔	217	┘	233	Θ	249	∙
138	è	154	Ü	170	¬	186	║	202	╩	218	┌	234	Ω	250	·
139	ï	155	¢	171	½	187	╗	203	╦	219	█	235	δ	251	√
140	î	156	£	172	¼	188	╝	204	╠	220	▄	236	∞	252	ⁿ
141	ì	157	¥	173	¡	189	╜	205	═	221	▌	237	φ	253	²
142	Ä	158	₧	174	«	190	╛	206	╬	222	▐	238	ε	254	■
143	Å	159	ƒ	175	»	191	┐	207	╧	223	▀	239	∩	255	

For example, lowercase (i):

1101001 binary

69 hexadecimal

105 decimal

Unicode

Universal standard for encoding, representing, and handling text in computers.

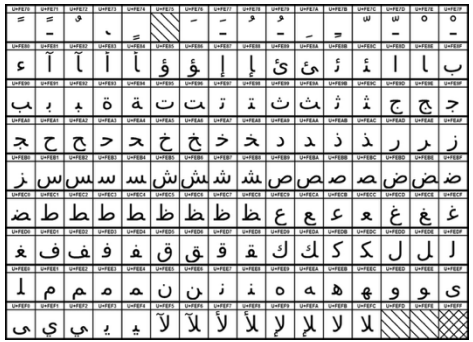
To ensure every character from every writing system has a unique code, no matter the platform, program, or language.

- Unique number for each character (e.g., U+0627 = Arabic letter “ا”)
- Covers 150+ scripts (Latin, Arabic, Chinese, Emoji, etc.).
- Enables global text processing and exchange across systems.

Without Unicode → text may appear garbled or inconsistent.
With Unicode → seamless communication across languages and platforms.

Examples

U+0041 → A (Latin)
U+0627 → ا (Arabic)
U+4E2D → 中 (Chinese)
U+1F600 → 😀 (Emoji)



Natural Language Processing
Character Encoding

In this lecture:

- Part 1: Encoding Schemes
- Part 2: **Locales and Word Order**

Locales and Word Order
(الترتيب الأبجدي)

A locale defines regional & cultural preferences for software:

- Language & script (Arabic, English, Chinese…)
- Date, time, and number formats
- Collation rules (how words are sorted and compared)

Different languages sort words differently:

- English: alphabetical order (A → Z)
- Swedish: å, ä, ö are separate letters at the end of the alphabet
- Arabic: (أ ب ت...) or (أ ب ج د..)
- Chinese/Japanese: sorted by radicals, stroke count, or phonetics
- Case sensitivity, accents, and diacritics **change order**:
- resume ≠ résumé in some locales, but treated the same in others

Unicode Collation

Collation = rules for comparing and ordering text.
Determines how strings are **sorted, matched, or searched** in different languages.
Without collation → “alphabetical order” varies and may confuse users.

Unicode Collation Algorithm (UCA)
A **standard method** for comparing Unicode strings across languages.
Provides a **default order** for all characters.
Allows **locale-specific tailoring** (e.g., Arabic, Swedish, Chinese).

Examples
English: resume < résumé < resumé
Swedish: z < å < ä < ö
Arabic: Sorts by base letters, ignoring diacritics ا = أ
Chinese: Sorted by **pinyin** or **radical + stroke order**.

Letter ordering systems in Arabic

1. Alphabetical Order (الألفبائي/الترتيب الهجائي)

أ، ب، ت، ث، ج، ح، خ، د، ذ، ر، ز، س، ش، ص، ض، ط، ظ، ع، غ، ف، ق، ك، ل، م، ن، هـ، و، ي

- Widely used today, especially in dictionaries, indexes, and reference works.
- Based on the similarity of letters in shape and sound
- By the linguist Naṣr ibn Asim al-Laythī (نصر بن عاصم الليثي), order of al-Ḥajjāj, to distinguish it from the Abjad order.

2. Abjad Order (الترتيب الأبجدي)

أ، ب، ج، د، هـ، و، ز، ح، ط، ي، ك، ل، م، ن، س، ع، ف، ص، ق، ر، ش، ت، ث، خ، ذ، ض، ظ، غ

- Oldest ordering of the Arabic script
- Primarily used in Abjad numerals (ḥisāb al-jummal), where each letter is assigned a numerical value - historically used in other Semitic languages.

3. Phonetic Order (المخرجي/الترتيب الصوتي)

ع، ح، هـ، خ، ق، ج، ش، ض، ص، س، ز، ط، د، ت، ظ، ث، ذ، ر، ل، ن، ف، ب، م، و، ي، أ، ء

- Less common than the other two systems. It is based on the articulation points of letters in the speech organs, beginning from the throat and moving outward [1, 4].
- by al-Khalīl ibn Aḥmad al-Farāhīdī (الخليل بن زحمد)

References

1. Mustafa Jarrar, Tymaa Hammouda: Qabas: An Open-Source Arabic Lexicographic Database. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 13363–13370, Torino, Italia. ELRA and ICCL.
2. Mustafa Jarrar, Diyam Akra, Tymaa Hammouda: **ALMA: Fast Lemmatizer and POS Tagger for Arabic.** Proceedings of the 2024 AI in Computational Linguistics (ACLING 2024), Procedia Computer Science.ELSEVIER.
3. Tymaa Hammouda, Mustafa Jarrar, Mohammed Khalilia: **SinaTools: Open Source Toolkit for Arabic Natural Language Understanding.** In Proceedings of the 2024 AI in Computational Linguistics (ACLING 2024), Procedia Computer Science, Dubai. ELSEVIER.
4. Jarrar, M. (2021). **The Arabic Ontology - An Arabic Wordnet with Ontologically Clean Content.** Applied Ontology Journal, 16:1, 1-26. IOS Press.
5. Jarrar, M., & Amayreh, H. (2019). **An Arabic-Multilingual Database with a Lexicographic Search Engine**. In Proceedings – 24th International Conference on Applications of Natural Language to Information Systems (NLDB 2019). Lecture Notes in Computer Science (vol. 11608, pp. 234-246). Springer. Doi:10.1007/978-3-030-23281-8_19
6. Karim El Haff, Mustafa Jarrar, Tymaa Hammouda, Fadi Zaraket: **Curras + Baladi: Towards a Levantine Corpus.** In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2022), Marseille, France. 2022
7. Amal Nayouf, Tymaa Hammouda, Mustafa Jarrar, Fadi zaraket, Mohamad-Bassam Kurdy: **Nâbra: Syrian Arabic Dialects with Morphological Annotations.** In Proceedings of the 1st Arabic Natural Language Processing Conference (ArabicNLP), Part of the TKM24 2023. ACL.
8. Mustafa Jarrar, Fadi Zaraket, Tymaa Hammouda, Daanish Masood, Martin Waehlisch: **Lisan: Yemeni, Iraqi, Libyan, and Sudanese Arabic Dialect Corpora with Morphological Annotations.** The 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA). Pages(1-7). IEEE. Egypt. DOI 10.1109/AICCSA59173.2023.10479250. 2023
9. Mustafa Jarrar, Sanad Malaysha, Tymaa Hammouda, Mohammed Khalilia: **SALMA: Arabic Sense-Annotated Corpus and WSD Benchmarks.** Proceedings the 1st ArabicNLP, Part of the ACL 2023. ACL.
10. Sana Ghanem, Mustafa Jarrar, Radi Jarrar, Ibrahim Bounhas: **A Benchmark and Scoring Algorithm for Enriching Arabic Synonyms.** In Proceedings of GWC2023, (pp.274-283). Spain, 2023
11. Moustafa Al-Hajj, Mustafa Jarrar: **ArabGlossBERT: Fine-Tuning BERT on Context-Gloss Pairs for WSD.** In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). PP 40--48, 2021
12. Moustafa Al-Hajj, Mustafa Jarrar: **LU-BZU at SemEval-2021 Task 2: Word2Vec and Lemma2Vec performance in Arabic Word-in-Context disambiguation.** In Proceedings of the Fifteenth Workshop on Semantic Evaluation (SemEval2021) Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC). PP 748--755, Association for Computational Linguistics. 2021
13. Mustafa Jarrar, Hamzeh Amayreh, John P. McCrae: **Representing Arabic Lexicons in Lemon - a Preliminary Study.** The 2nd Conference on Language, Data and Knowledge (LDK 2019). Pages(29-33). CEUR, Volume 2402. ISSN:1613-0073. Leipzig, Germany. 2019
14. Diana Alhafi, Anton Deik, Mustafa Jarrar: **Usability Evaluation of Lexicographic e-Services.** 16th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA). Pages(1-7). IEEE. UAE. 2019
15. Mustafa Jarrar, Fadi Zaraket, Rami Asia, Hamzeh Amayreh: **Diacritic-Based Matching of Arabic Words.** ACM Asian and Low-Resource Language Information Processing. Volume 18, No 2, Pages(10:1-10:21), ACM, ISSN:2375-4699. December, 2018
16. Mustafa Jarrar: **Building a Formal Arabic Ontology (Invited Paper)**. In proceedings of the Experts Meeting on Arabic Ontologies and Semantic Networks. Alecso, Arab League. Tunis, July 26-28, 2011.
17. Mustafa Jarrar: **Towards the notion of gloss, and the adoption of linguistic resources in formal ontology engineering.** In proceedings of the 15th International World Wide Web Conference (WWW2006). Edinburgh, Scotland. Pages 497-503. ACM Press. ISBN: 1595933239. May 2006.
18. Mohammed Khalilia, Sanad Malaysha, Reem Suwaileh, Mustafa Jarrar, Alaa Aljabari, Tamer Elsayed, Imed Zitouni: **ArabicNLU 2024: The First Arabic Natural Language Understanding Shared Task.** In Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024), Bangkok, Thailand. Association for Computational Linguistics.
19. Mustafa Jarrar, Nagham Hamad, Mohammed Khalilia, Bashar Talafha, AbdelRahim Elmadany, Muhammad Abdul-Mageed: **WojoodNER 2024: The Second Arabic Named Entity Recognition Shared Task.** In Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024), Bangkok, Thailand. Association for Computational Linguistics.