Langue et Informatique

Pierre R. Mercuriali

CRIT – UMLP – Besançon

Spring 2026

References – to go further

- *Language & Computers*, Lelia Glass, Markus Dickinson, Chris Brew, and Detmar Meurers
- *Speech and Language Processing*, Dan Jurafsky, James H. Martin
- *Artificial Intelligence: A Modern Approach*, Stuart Russell, Peter Norvig
- *Verbal Behavior*, Burrhus Frederic Skinner
- (Illustrations from Wikipedia unless specified)

Outline

Introduction

Encodings

Writer's Aid

What is language?

UNIVERSITÉ DE FRANCHE-COMTÉ

C.R.I.T. UR3224

Some propositions? (Just terms!)

- ...
- ...

Some caracteristics

Anthropologist Charles Hockett (1960)

- Produced with vocal apparatus
- Transitory, ephemeral (barring recording)
- Discrete units?
- Learnable
- Transmitted culturally
- Reflexive (use language to talk about language)

Verbal behavior

Psychologist Burrhus Skinner (60s)

- Behavior: what we can observe.
- Verbal: concerned with *mediation* through someone else
- Example
  - Individual is thirsty (stimulus);
  - There is an audience (someone else);
  - *"Gimme a drink"*;
  - Individual gets the drink.

Some languages?...

- French?
- English?
- Swiss Sign Language?
- Music?
- Heavy Metal music?
- Python?
- Mathematics?

Language: vocal vs written

VOCAL

- Can do amazing things already! (Tell stories, discuss, etc.)
- All human societies use it
- Date? Problem: *no trace!*
  - Johanna Nichols (statistical arguments): language differenciation at least 100k y.o.
  - Stone tools:-3.4M, earliest fire:-1.7M, Neanderthals: -500k...

WRITTEN

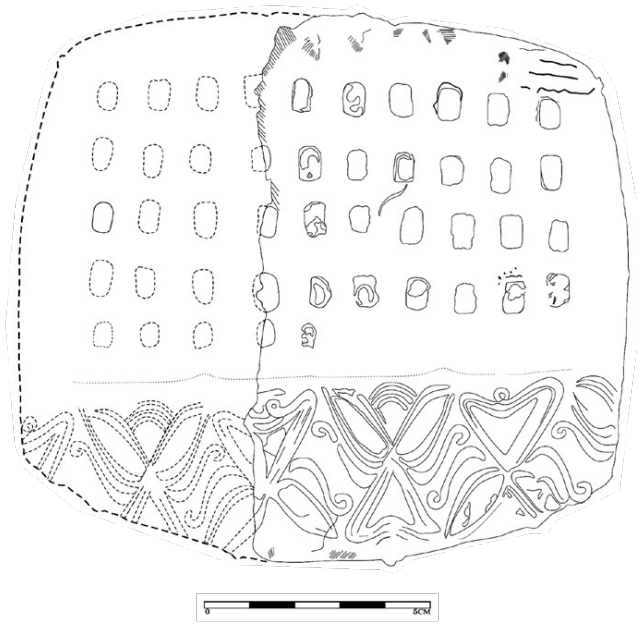
- Records of the ephemeral
- Not all humans use it
- Not all societies use it ("oral tradition", Homeric poetry)
- 3 to 4k years old (archeological), maybe more (Lebombo bone, tally stick -42k)

Early writing: tally sticks

Lebombo bone, tally stick from -42k B.C.

Tally stick from Germany (1550s at least). archaeology.org

Early writing: clay tablets

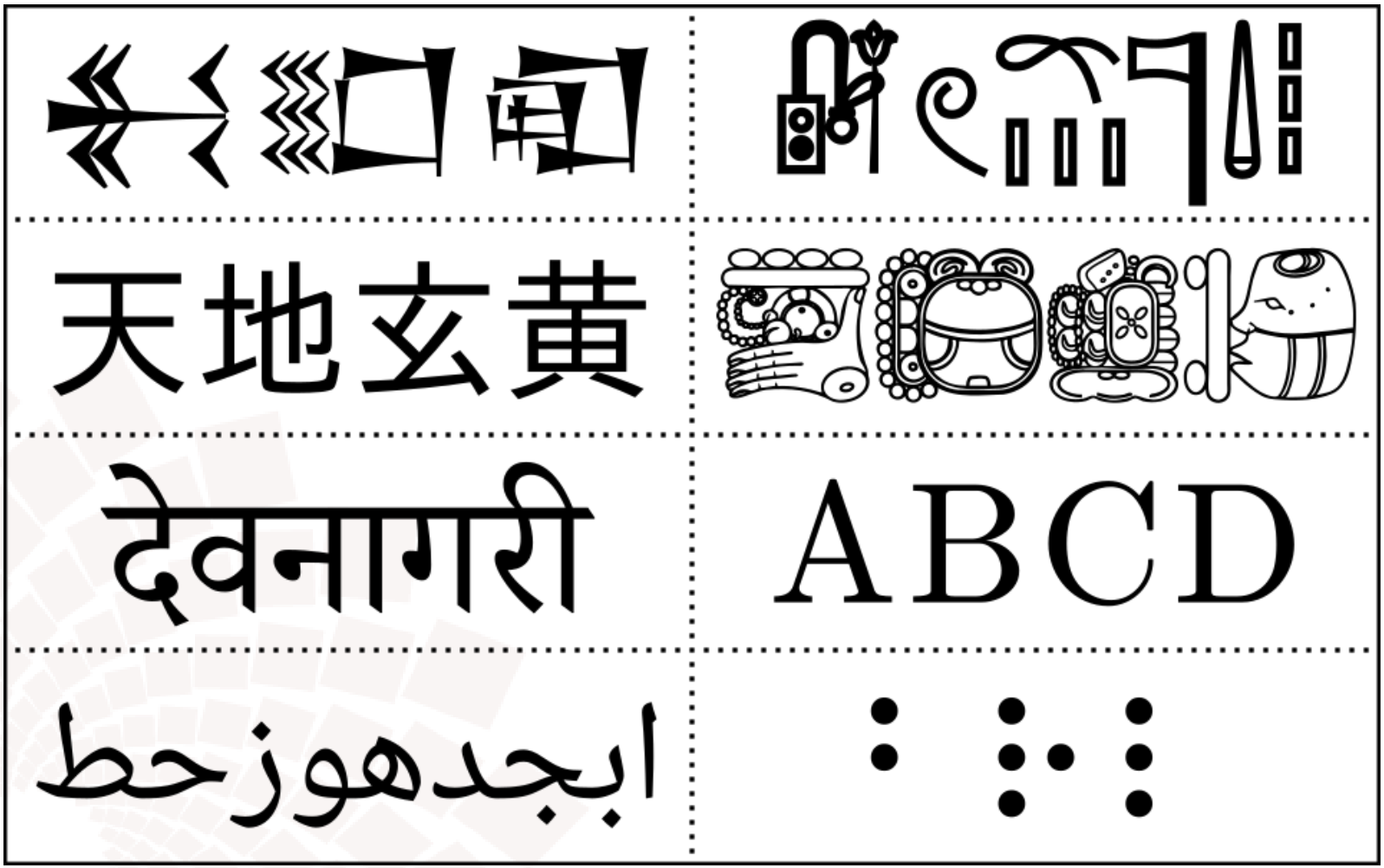


Proto-Elamite Tablets from Shahr-i Sokhta (Iran/Persia), c. 3100 – c. 2900 BC.

Early writing: papyrii

Heart weighting, from the *Book of the Dead*. Papyrus of Hunefer, c. 1275 BC.

Writing: overview



Ian Remsen: survey of writing systems. Cuneiform, hieroglyphs, Chinese, Mayan, Devanagari, Latin, Arabic, Braille.

An Evolution of Language

- *Parcimony*: small gradual changes more likely than one big change.
- Behavior: what we observe. Actions in organism.
- Operant behavior: has effect on environment.
- Operant control: the operant behavior is controlled by environment (reaction)

1. Decisive step: vocal musculature under operant control (genetic change)
   - Environment now has effect on vocal behavior
   - Result: coordination of all organs for speech production
   - Natural selection advantages: sound efficient in the dark, hidden, hands busy
2. Generalization of consequences: same answer:
   - in other environments,
   - with other consequences,
   - under exclusive control of a certain stimulus
3. Verbal answer modified and maintained by verbal environment, maintained from generation to generation (a *language*)

Outline

Introduction

Encodings

Writer's Aid

Writing systems

- How to process language in NLP?
- Encoding language: writing behavior
- These slides: 26 letters (Latin alphabet)
  - 26 letters, upper and lowercase
  - Punctuation
  - About 100 symbols. Simple?

But we want to deal with *all* languages!

Writing systems versus languages

Same writing system, different languages

French, English, German, Vietnamese... Latin alphabet

Same language, different writing systems

- Chinese: traditional characters, simplified characters, Pinyin
- Turkish: Arabic vs Latin
- Japanese: one language, 3 different writing systems

A misconception?

...

"French is written in the English alphabet."

What is or isn't encoded in writing?

- Language: (often) sounds.
- Alphabetic system: each symbol roughly represents a sound.
- Syllabic system: each symbol represents a syllable.
- Logographic system: each symbol represents an abstraction (not sound).

Alphabetic writing system

- Each character: one sound/articulatory gesture
- Some exceptions (in English, French, etc.)
  - Silent letters: knee, dept
  - Multiple letters, one sound: running, revolution
  - Multiple sounds, one letter: tax
  - Homophones: Colonel, kernel. River bank, financial bank.
- Latin, Greek, Cyrillic alphabets, etc.

Exploring writing systems

UNIVERSITÉ DE FRANCHE-COMTÉ

C.R.I.T. UR3224

Omniglot

https://www.omniglot.com/

International Phonetic Alphabet

UNIVERSITÉ DE FRANCHE-COMTÉ

C.R.I.T. UR3224

How to (more or less) accurately report vocal behavior?

`https://www.ipachart.com`

- Each character unambiguously represents exactly one sound.
- Periodic table of sounds: meaningful groupings.
  - Row: how air is stopped/ articulated
  - Columns: front to back of mouth
- Used by linguists
- Represent sounds of all languages
- Not actually used to write any language! (Why not?)

IPA for French (consonants)

UNIVERSITÉ DE FRANCHE-COMTÉ
C.R.I.T. UR3224

Consonant phonemes of French

		Labial	Dental/ Alveolar	Palatal/ Postalv.	Velar/ Uvular
Nasal		m	n	ɲ	(ŋ)
Plosive	voiceless	p	t		k
	voiced	b	d		g
Fricative	voiceless	f	s	ʃ	
	voiced	v	z	ʒ	ʁ
Approximant	plain		l	j	
	labial			ɥ	w

French IPA for consonants.

Abjads: consonant alphabets

- Only consonants are written
- Vowels are inferred from context
- Hebrew, Arabic

(Note: sometimes languages are written right to left, in columns...)

Syllabic writing systems (I)

- Each symbol: a syllable
- Abudiga (e.g. Burmese): organized into meaningful rows, columnns
- Regular syllabaries (e.g. Vai in Sierra Leone): no meaningful grouping

Syllabic writing systems (II)

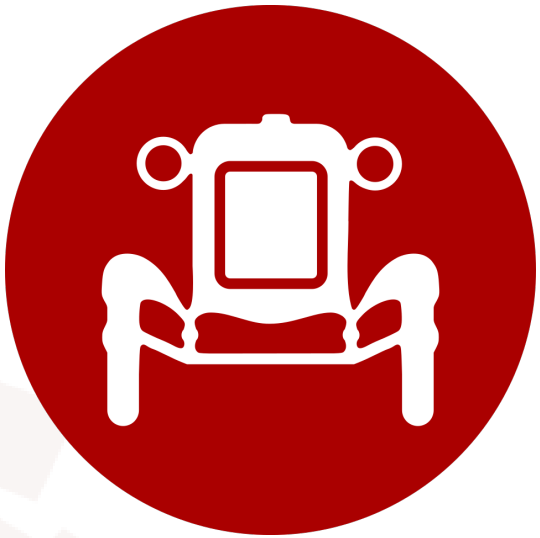
Some tendencies can be observed in vocal communities.

- Syllable structure varies across languages
- Hawaiian: only open and CV syllables. *Aloha!*
- Mandarin: syllables can only end in a vowel. Nasals.
- English: allows closed syllables (CVC: as in *top*).
- English: consonant clusters (CCVCC: *stork*)

Large number of possible syllables in English: syllabary less practical.

Logographic writing systems (I)

- No human language written in a pure logographic system
- Road symbols: logographs
- Do they have standard phonetic realizations?



"Interdit aux automobiles," c. 1920.

Logographic writing systems (II)

- Chinese characters:
  - Syllables
  - Logographic and phonetic elements
  - "semantic-phonetic compounds"
- Over time they are used in an increasing variety of contexts ("abstraction")

Changes (and abstraction) over time

From left to right: Sun, Moon, Vehicle, Horse.
From *Chinese character recognition: History, status and prospects* (Dai, Liu, Xiao, 2007)

Abstraction

"Moon" radical: used as "moon" but also "month."

Hybrid systems

- Chinese "semantic-phonetic" compounds
- Korean hangeul: each block represents a syllable with alphabetic elements

Some features of a writing system

- Are the basic symbols enhanced with *diacritics*?
- How are the words separated?
- Are there words?
- How are sentences separated?
- Paragraphs?
- Punctuation? Quotation marks? Italics?
- Upper-case? Lower-case? No case?
- Left to right? Right to left? Top to bottom? Left to right, then right to left, then left to right, etc? (Boustrophedon)

Emoji

- Emoji are logographic
- Are they a writing system?
  - In what context are they used?
  - What do they "stand in"?
  - Are they "expressive" enough?
- Example: *Emoji Dick* (Fred Benenson, 2010). *Moby Dick* translated into emoji.
- Recovering the original text is impossible. (Or is it?)

Storing things in the computer

- Bit: 0 or 1.
- Bytes: sequences of bits.
- Bytes can represent (decimal) numbers in binary notation.
- Examples ("Big Endian" notation):
  - 00000000: 0
  - 00000001: 1
  - 00000010: 2
  - 00000100: 4
  - 00000101: 5
  - 01001010: 74

Storing characters in the computer

- 8 bits in a byte.
- We can represent $2^8 = 256$ characters!
- With 7 bits, we can represent $2^7 = 128$ characters.
- Enough for American characters.

ASCII: American Standard Code for Information Exchange.

The Entire ASCII Table

UNIVERSITÉ DE FRANCHE-COMTÉ

C.R.I.T. UR3224

ASCII TABLE

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

Entire ASCII table (128 characters).

Unicode

- ASCII is enough for English.
- For all languages we use *UNICODE*.
- Unicode : 8 *bytes* : 2^{32} : *4 294 967 296 characters!*
- UTF-8: 256 characters.

`https://shapecatcher.com/`
`http://xahlee.info/comp/unicode_index.html`

Consequences

- Digital writing allows language to be disseminated.
- Humans use language in many different contexts,
  as reactions to specific stimuli.
- Computers are artificial.
  a different representation of language.

Review (practice your "intraverbals")

- Give examples of alphabetic, syllabic, logographic writing systems.
- Explain the meaning of rows and columns in the IPA for consonants.
- Discuss why a language can be written in severan different writing systems.
- Discuss what this shows about the relationship between written and spoken forms of a language.
- Recognize the numbers can be represented in different ways.
- Explain what Unicode is.

Activity: exploring a writing system

- Choose a writing system in Omniglot.
- Is it alphabetic, syllabic, logographic?
- How many symbols does it contain?
- Is it available in Unicode?
- Can you write your name in it?

Examples: Georgian, Armenian, Mayan, Gothic, Fraser, Egyptian hieroglyphics, Japanese hiragana, Mongolian, Cherokee...

Outline

Introduction

Encodings

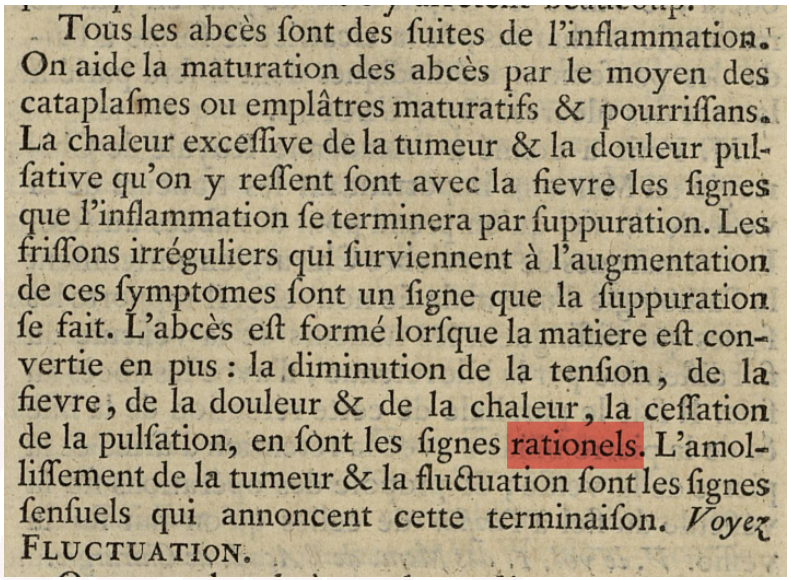
Writer's Aid

A History of Spelling

- Latin alphabet: 700 to 600 B.C. (inspired by Greek + Etruscan alphabets)
- Printing press: 1470s. Bibles, dictionaries.
- ASCII: 1960s
- English spelling: formalized... in the 1700s!
- How did William Shakespeare (1564-1616) himself write his name?
  - Willm Shakp
  - William Shaksper
  - Wm Shakspe
  - William Shakspere
  - Willm Shakspere
  - William Shakspeare

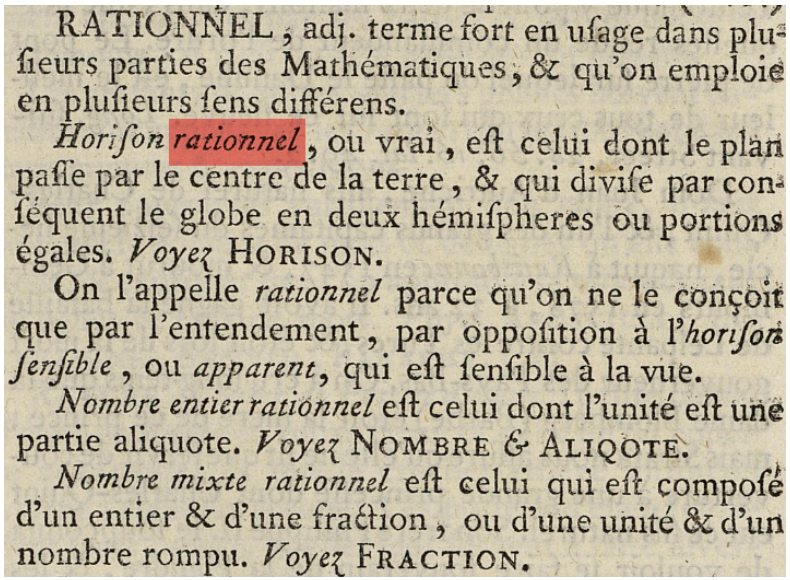
See also: `www.agecrofthall.org/single-post/shakespeare-s-name-and-handwriting`

An example in French



Tous les abcès ſont des ſuites de l'inflammation. On aide la maturation des abcès par le moyen des cataplaſmes ou emplâtres maturatifs & pourriſſans. La chaleur exceſſive de la tumeur & la douleur pulſative qu'on y reſſent ſont avec la fievre les ſignes que l'inflammation ſe terminera par ſuppuration. Les friſſons irréguliers qui ſurviennent à l'augmentation de ces ſymptomes ſont un ſigne que la ſuppuration ſe fait. L'abcès eſt formé lorſque la matiere eſt convertie en pus : la diminution de la tenſion, de la fievre, de la douleur & de la chaleur, la ceſſation de la pulſation, en ſont les ſignes rationels. L'amolliſſement de la tumeur & la fluctuation ſont les ſignes ſenſuels qui annoncent cette terminaiſon. *Voyez* FLUCTUATION.

Entry ***abcès***.



RATIONNEL, adj. terme fort en uſage dans pluſieurs parties des Mathématiques, & qu'on emploie en pluſieurs ſens différens.

Horiſon rationnel, ou vrai, eſt celui dont le plan paſſe par le centre de la terre, & qui diviſe par conſéquent le globe en deux hémiſpheres ou portions égales. *Voyez* HORISON.

On l'appelle *rationnel* parce qu'on ne le conçoit que par l'entendement, par oppoſition à l'*horiſon ſenſible*, ou *apparent*, qui eſt ſenſible à la vue.

Nombre entier rationnel eſt celui dont l'unité eſt une partie aliquote. *Voyez* NOMBRE & ALIQOTE.

Nombre mixte rationnel eſt celui qui eſt compoſé d'un entier & d'une fraction, ou d'une unité & d'un nombre rompu. *Voyez* FRACTION.

Entry ***rationnel***.

l'Encyclopédie, **Diderot et d'Alembert 1751.**

Why standardized spelling?

- 2 extremes:
  - Everyone writes however they want (Shakespeare)
  - Everyone writes phonetically (IPA: unambiguous)
- Meet in the middle?
- Consequences for record-keeping?
  (How to search for specific things?)
- Consequences for communication?
  (What about accent variability?)

Accents

- Français "moderne", uniformisé : `le /poisson/`
- Meusien (lorrain) : `eul' /pisson/, /posson/`

Check out `www.youtube.com/watch?v=ubGjasm63Y0`, and `atlas.lisn.upsaclay.fr/` !