Codage de l'information

Laurent Noé-Léopold Weinberg
paternité : Francesco De Comité

Licence 1 Mathématiques Informatique Semestre 1
Université de Lille
Faculté des Sciences et Technologies

10 septembre 2025

Représentation des caractères

Un peu de philosophie

- Les premières utilisations des ordinateurs étaient purement numériques : calculs scientifiques, commandes de moteur...
- On ne concevait pas de communiquer avec des ordinateurs, ni de se servir des ordinateurs pour communiquer entre humains.
- Il existait déjà des méthodes de communication entre êtres humains répondant aux besoins.

Allez, encore un peu de philosophie

- Lorsqu'on a eu besoin de produire ou utiliser des messages en langage naturel, on a utilisé les techniques et les codages existants.
- Les différentes contraintes liées à l'informatique ont amené les concepteurs à modifier ou préciser ces codages.
- Ce cours commencera par regarder les méthodes d'échanges de messages entre humains.
- On continuera ensuite, chronologiquement, à décrire les différentes solutions informatiques.

Communication humaine

Les langages

- Deux grandes familles de langages :
  - Les langages alphabétiques : on dispose d'un ensemble de symboles représentant des sons, qu'on associe pour former des mots ...
  - Les idéogrammes : un symbole représente une idée, un concept. On les associe pour créer un message.
- Dans les deux cas, on dispose d'un ensemble de symboles de base (plus ou moins étendu, mais fini), avec lesquels on peut communiquer.

Communication humaine

Les langages

- Ces outils de communication ont fait leurs preuves (discours, livres, contrats ...)
- Pratiques et précis quand la longueur des messages n'est pas un problème...
- ...ou quand la durée de transmission du message ne joue pas sur le coût de cette transmission.
- Idéal quand la longueur et les informations véhiculées par le message permettent d'éliminer les imprécisions et les ambiguïtés.

Exemples

- Décisions de justice, contrats, diffusion des lois et édits gouvernementaux...
- Informations plus générales (victoires, vie des dirigeants ...)

Communication humaine

L'apparition des contraintes

- Comment aller plus vite que les moyens de transports physiques ?
- Combien ça coûte ?
- Comment limiter ou éviter les erreurs de transmission ?

Télé-communication

Différents moyens

- Le coureur de Marathon, les coursiers.
- Optique : Tours génoises, signaux de fumée, signaux maritimes, télégraphe de Chappe.
- Electrique : Télégraphe Morse.
- Radio : Télégraphie sans fil.
- Informatique.

Tours génoises

La tour de Parata (Corse)

- Construites au XVème siècle sur les côtes des iles dépendant de la République de Gênes.
- Leur rôle : prévenir des attaques des pirates barbaresques.
- Chaque tour, bâtie sur la côte, est à portée de vue de deux autres tours (une de chaque coté).
- En cas d'attaque, un feu est allumé sur la tour, ainsi que sur les tours voisines de proche en proche.
- Un seul signal, un seul message.

Crédits : Wikipedia

Signaux de fumée

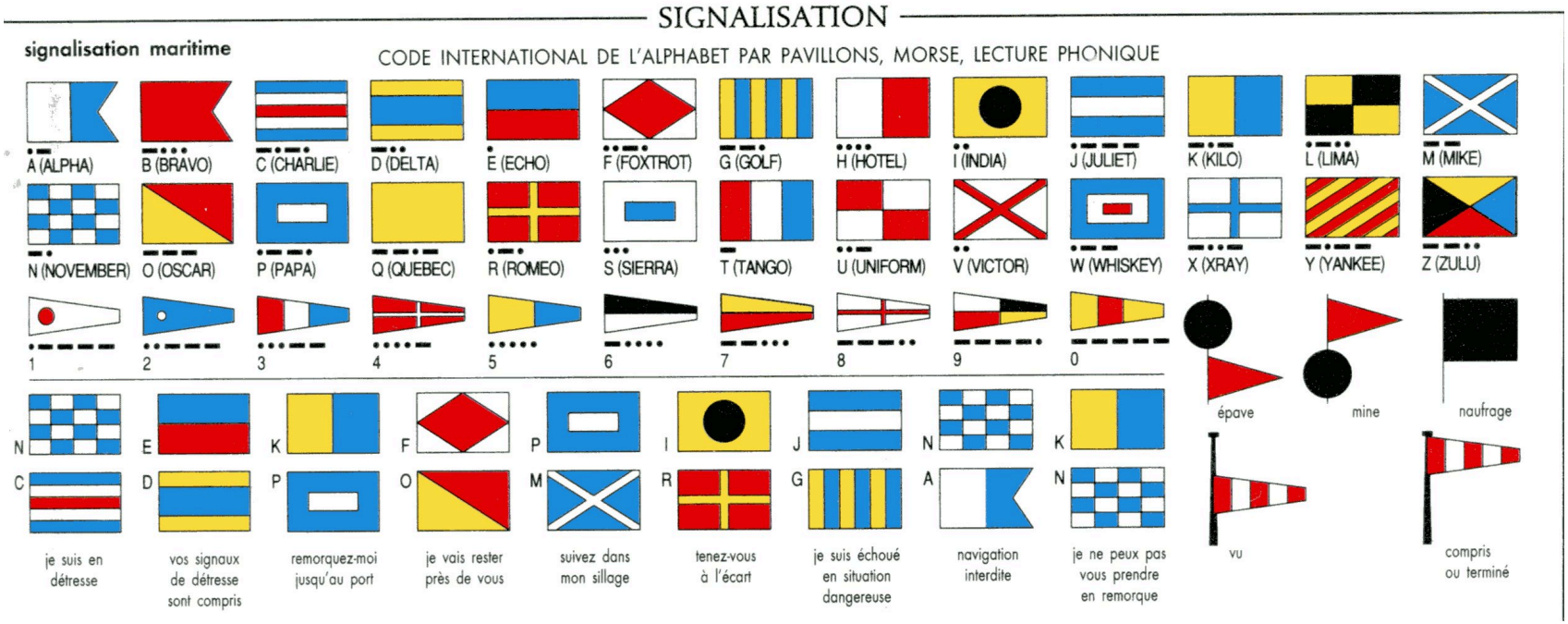
- Pas plus de deux ou trois signaux distincts.
- Difficile de moduler les signaux : tout ou rien.
- Les messages ne sont pas confidentiels.

Crédits : Wikipedia, F. Remington

Signaux maritimes

Crédits : Studios Ghibli

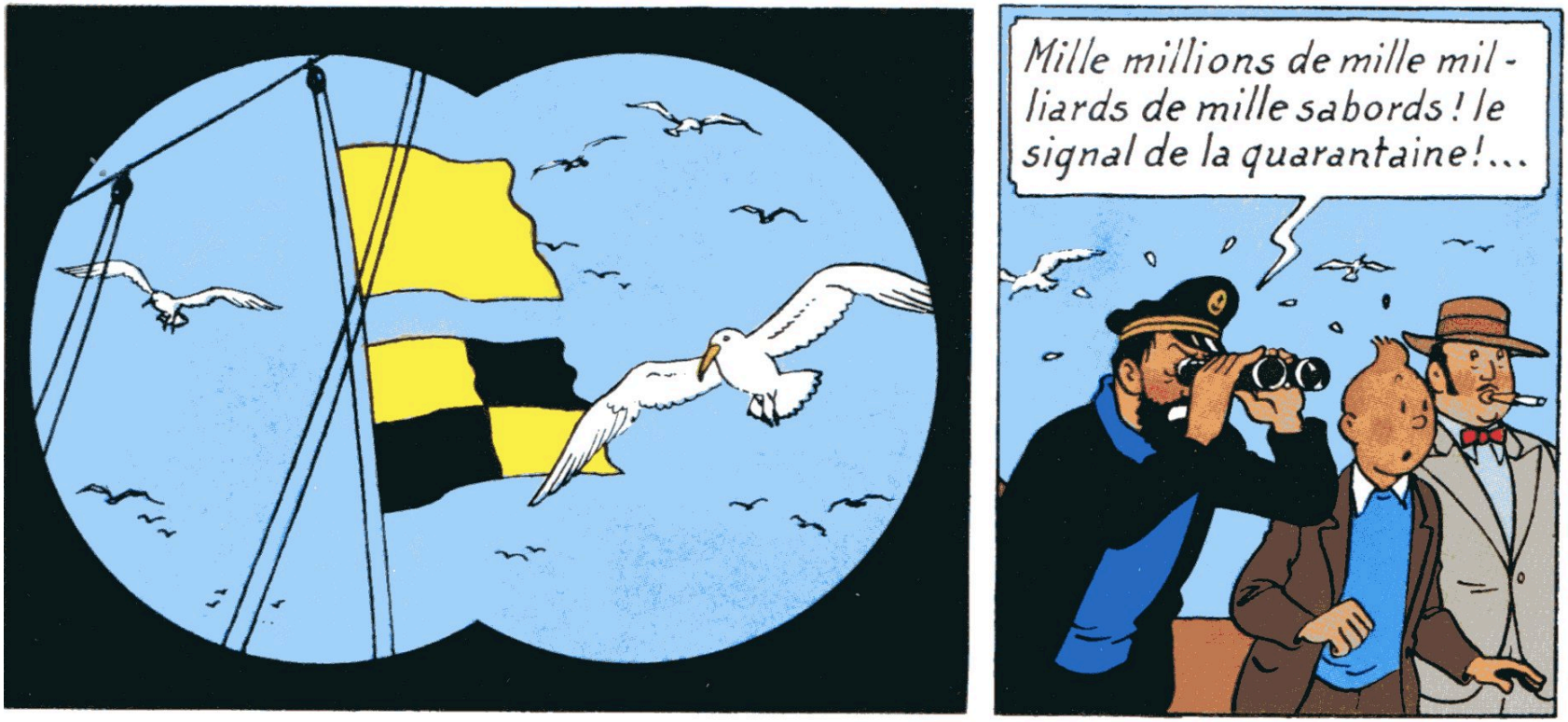
Signaux maritimes



- Un pavillon par lettre et chiffre.
- Chaque pavillon a une signification non alphabétique.
- Les combinaisons de deux ou trois pavillons ont une signification spéciale.
- Peu adapté pour afficher de longs mots.
- Pas de confidentialité, mais ce n'est pas le but.

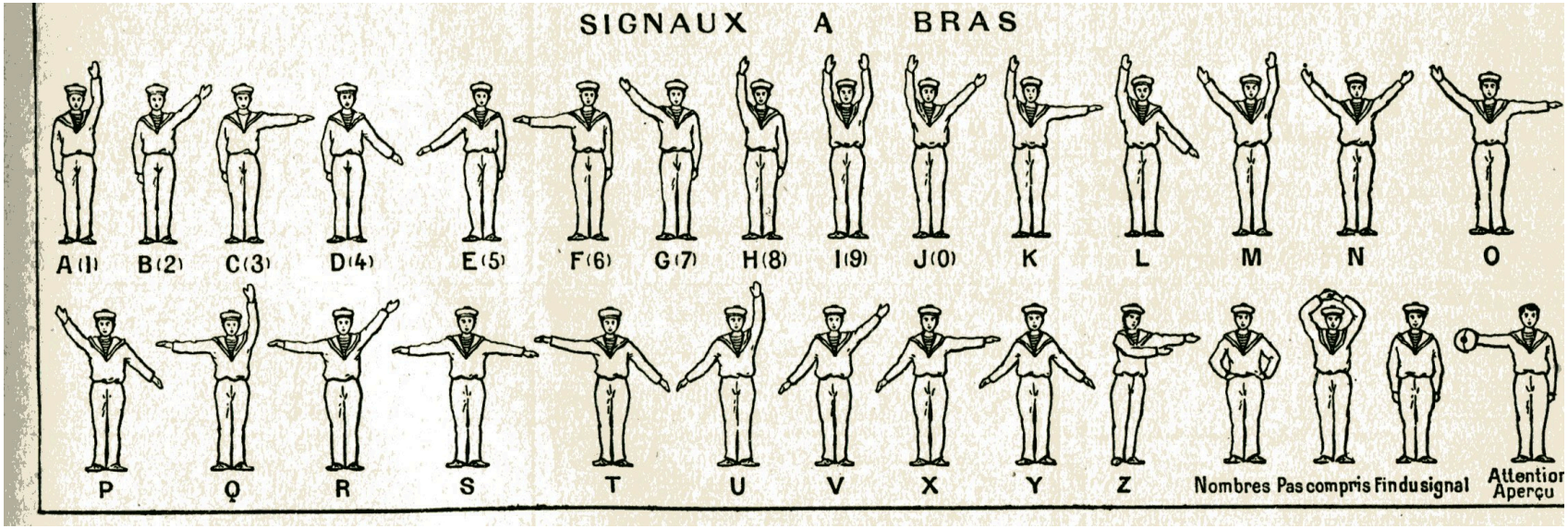
Crédits : Larousse

Composition de signaux



Crédits : Casterman

Signaux maritimes



- Codage alphabétique.
- Sans doute plus rapide que les pavillons.
- Un dispositif de commutation entre lettres et chiffres.
- Les associations entre une lettre et une position sont faites de façon systématique : voyez-vous comment ?
- Pourquoi le Z a-t-il cette configuration bizarre ?

Crédits : Larousse

Signaux maritimes

Crédits : Inconnus

Sémaphore

Lettre	bras droit	bras gauche	Lettre	bras droit	bras gauche
A	Sud	Nord	N	Nord-Ouest	Nord-Est
B	Sud	Nord-Est	O	Nord-Ouest	Est
C	Sud	Est	P	Nord-Ouest	Sud-Est
D	Sud	Sud-Est	Q	Ouest	Nord
E	Sud-Ouest	Sud	R	Ouest	Nord-Est
F	Ouest	Sud	S	Ouest	Est
G	Nord-Ouest	Sud	T	Ouest	Sud-Est
H	Nord	Sud	U	Sud-Ouest	Nord
I	Nord	Nord	V	Sud-Ouest	Nord-Est
J	Nord	Nord-Est	W		
K	Nord	Est	X	Sud-Ouest	Est
L	Nord	Sud-Est	Y	Sud-Ouest	Sud-Est
M	Nord-Ouest	Nord	Z	?	?

Sémaphore

Codage :Sud=0, Sud-Est ou Sud-Ouest=1, Est ou Ouest=2
Nord–Est ou Nord-Ouest =3, Nord=4

Lettre	D	G	Lettre	D	G
A	0	4	N	3	3
B	0	3	O	3	2
C	0	2	P	3	1
D	0	1	Q	2	4
E	1	0	R	2	3
F	2	0	S	2	2
G	3	0	T	2	1
H	4	0	U	1	4
I	4	4	V	1	3
J	4	3	W		
K	4	2	X	1	2
L	4	1	Y	1	1
M	3	4	Z	?	?

Sémaphore

Analyse

- On utilise tous les codes de '01' jusqu'à '44', avec les chiffres de 0 à 4
- Equivalent à un codage de base 5 sur deux chiffres.
- On peut coder 25 situations (5*5).
- Insuffisant pour coder les 26 lettres de l'alphabet.
- On choisit de ne pas représenter le 'W' : il reste 25 lettres à coder.
- On ne peut pas utiliser le code Sud-Sud : on doit utiliser un type de codage différent pour un des caractères : Z

Télégraphe de Chappe

Source gallica.bnf.fr / Bibliothèque nationale de France

- Claude Chappe (1763-1805)
- Système de communication optique.
- Déployé à partir de 1794.
- Utilisé jusqu'à l'apparition du télégraphe électrique (≈ 1870).

Crédits : gallica.bnf.fr

Principe du télégraphe de Chappe

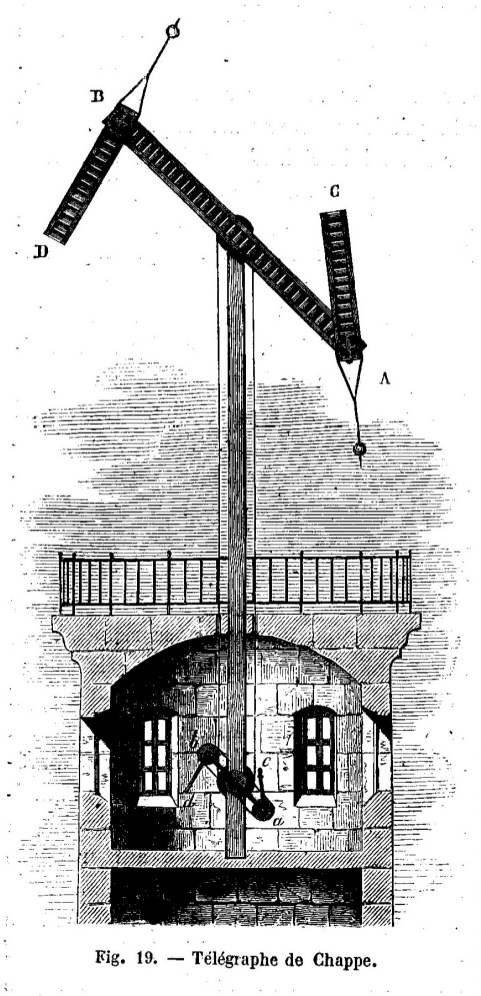
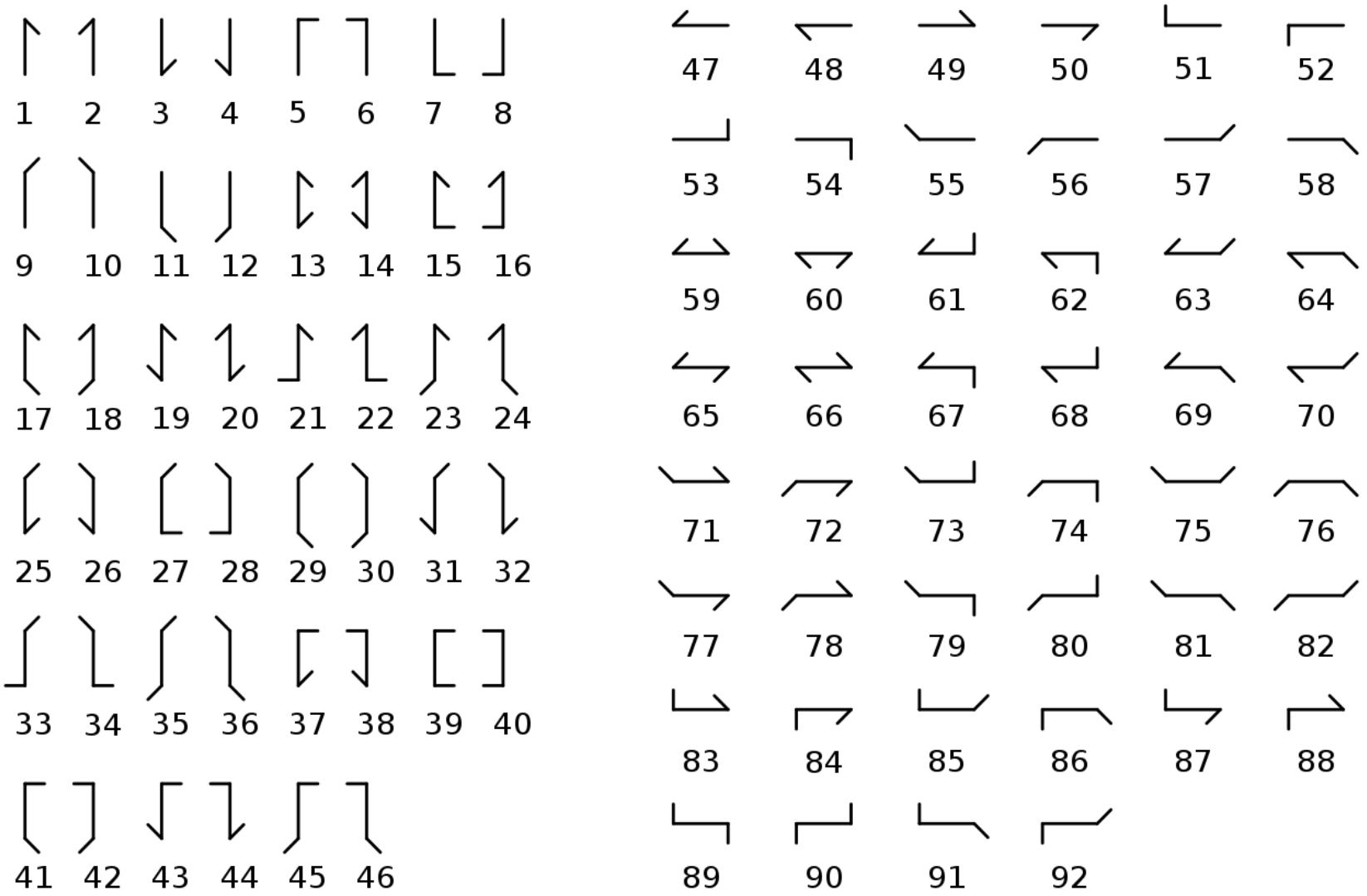


Fig. 19. — Télégraphe de Chappe.

- Trois bras articulés.
- A chaque position des bras correspond un nombre compris entre 1 et 98 ...
- 6 codes sont réservés au service (codes de contrôle), il reste 92 codes disponibles.
- Un mot est codé par deux positions : soit $92 \times 92 = 8464$ mots.
- Les codes sont consignés dans le livre des codes.
- Seuls les directeurs en début et fin de ligne possèdent un exemplaire de ce livre.
- Les opérateurs transmettent les codes sans en déchiffrer le sens.

Crédits : wikipedia

Codage Chappe



Crédits : wikipedia

Livres des codes (extrait)

67	02	Boni
27	64	Bonification, bonifier
94	06	Bonnet
21	98	Bonté
48	05	Bord
13	92	A bord
21	25	*Bordeaux*
97	18	Bordée
64	83	Border, bordage
55	49	Bordereau
09	79	Bordure
36	99	Borgne
28	56	Borner, borne
31	12	Bornez-vous à
58	60	*Bosphore*
02	14	Bossoir
44	34	Bosse, bossu
38	12	Bout
33	61	Bouteille
86	69	Boutique
15	74	Bouton, boutonner
05	68	Boutonnière
59	50	Boyau
18	87	Braconnier
73	85	Brancard
20	04	Branche
77	23	Braquer
17	03	Bras
58	82	Brasser, brasse
92	88	Brasserie
22	19	Brasseur
57	98	Brave, bravement
15	28	Braver, bravade
82	79	Bravo

Codage

32	Commandant de la rade	82	Conseil de discipline
33	Vestibule	83	Président du Conseil d'État
34	Littérateur	84	Graver, gravure
35	Ayant dû	85	Ennemi
36	Paternité, paternel, paternellement	86	Ration de
37	De ce que	87	Braconnier
38	À ses	88	Sachez
39	Intercepter	89	Huis clos
40	Cylindre, cylindrique	90	Premier écuyer de l'Empereur
41	Ni	91	Résigner, résignation
42	Vin	92	Présentable
43	Consommer, consommation	93	Arbitrer, arbitrage

Extrait de la page 18

Décodage

Attention : l'exemple ci-dessus n'est pas vraiment un code de Chappe, car des codes supérieurs à 92 sont utilisés ...

Crédits : Musée des transmissions Cesson-Sévigné

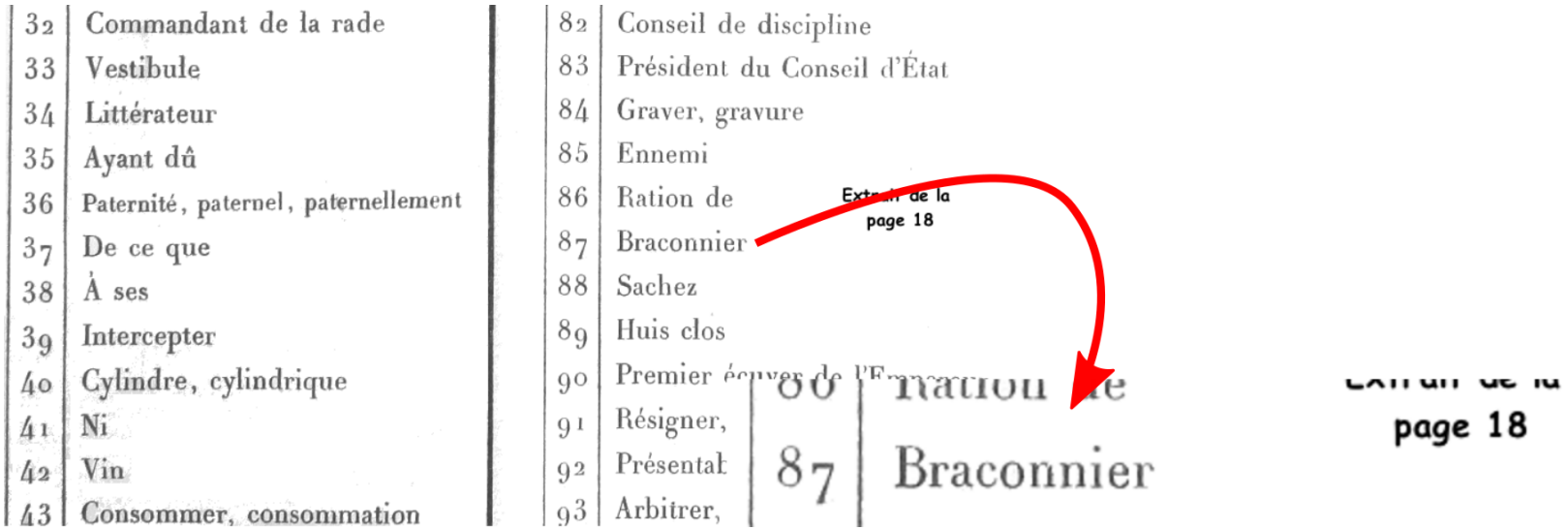
Chappe - codage

67	02	Boni
27	64	Bonification, bonifier
94	06	Bonnet
21	98	Bonté
48	05	Bord
13	92	A bord
21	25	*Bordeaux*
97	18	Bordée
64	83	Border, bordage
55	49	Bordereau
09	79	Bordure
36	99	Borgne
28	56	Borner, borne
31	12	Bornez-vous à
58	60	*Bosphore*
02	14	Bossoir
44	34	Bosse, bossu
38	12	Bout
33	61	Bouteille
86	69	Boutique
15	74	Bouton, boutonner
05	68	Boutonnière
59	50	Boyau
18	87	Braconnier
73	85	Brancard
20	04	Branche
77	23	Braquer
17	03	Bras
58	82	Brasser, brasse
92	88	Brasserie
22	19	Brasseur
57	98	Brave, bravement
15	28	Braver, bravade
82	79	Bravo

18	87	Braconnier

- L'émetteur cherche le mot à transmettre dans un dictionnaire, où les mots sont classés par ordre alphabétique.
- A "Braconnier" sont associés les codes 18 et 87.

Chappe - décodage



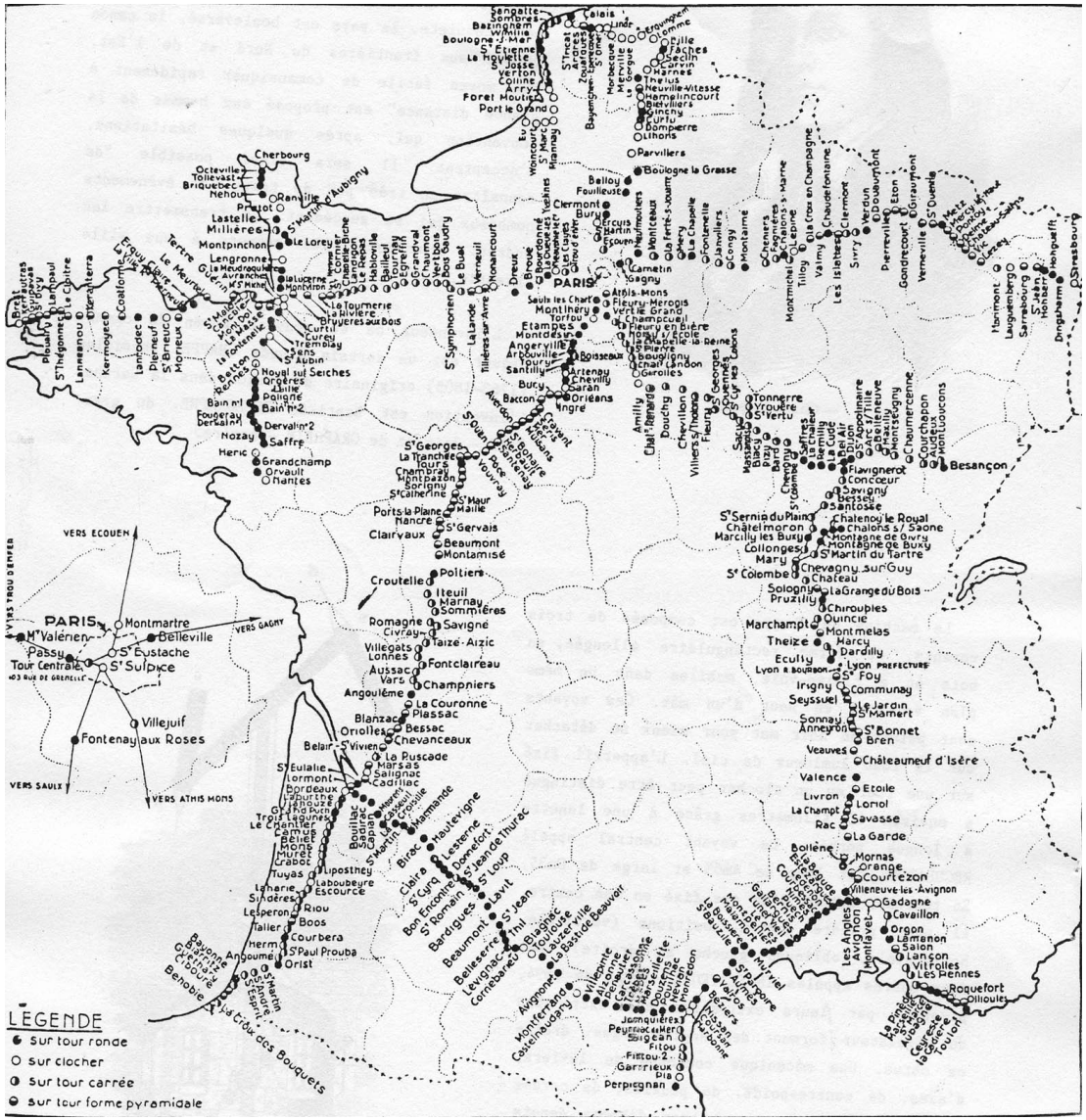
- Le récepteur se sert du premier code (18) pour ouvrir le livre à la bonne page.
- La deuxième partie du code (87) lui indique la position du mot dans la page.

Divers

Chronologie

- La ligne Paris-Lille est opérationnelle en 1794.
- Une quinzaine de tours, 9 minutes pour un symbole.
- Premier message : la prise du Quesnoy, le 15 août 1794 annoncée à la Convention.
- de 1834 à 1836 : piratage du télégraphe par les frères Blanc.

Extension du réseau (1852)



Crédits : Cité des Télécoms, Pleumeur-Bodou

Piratage du télégraphe

Histoire

- Connaître les cours de la bourse de Paris avant les autres procure un avantage certain.
- Les informations circulaient à vitesse de diligence, soit deux ou trois jours vers 1830 pour relier Paris à Bordeaux.
- Un message télégraphique est transmis en quelques heures.
- Mais ...le télégraphe est réservé aux communications officielles, et inacessible aux particuliers.
- Difficile de corrompre à la fois les responsables en tête et en queue de ligne.

Piratage du télégraphe

Points faibles, point d'entrée

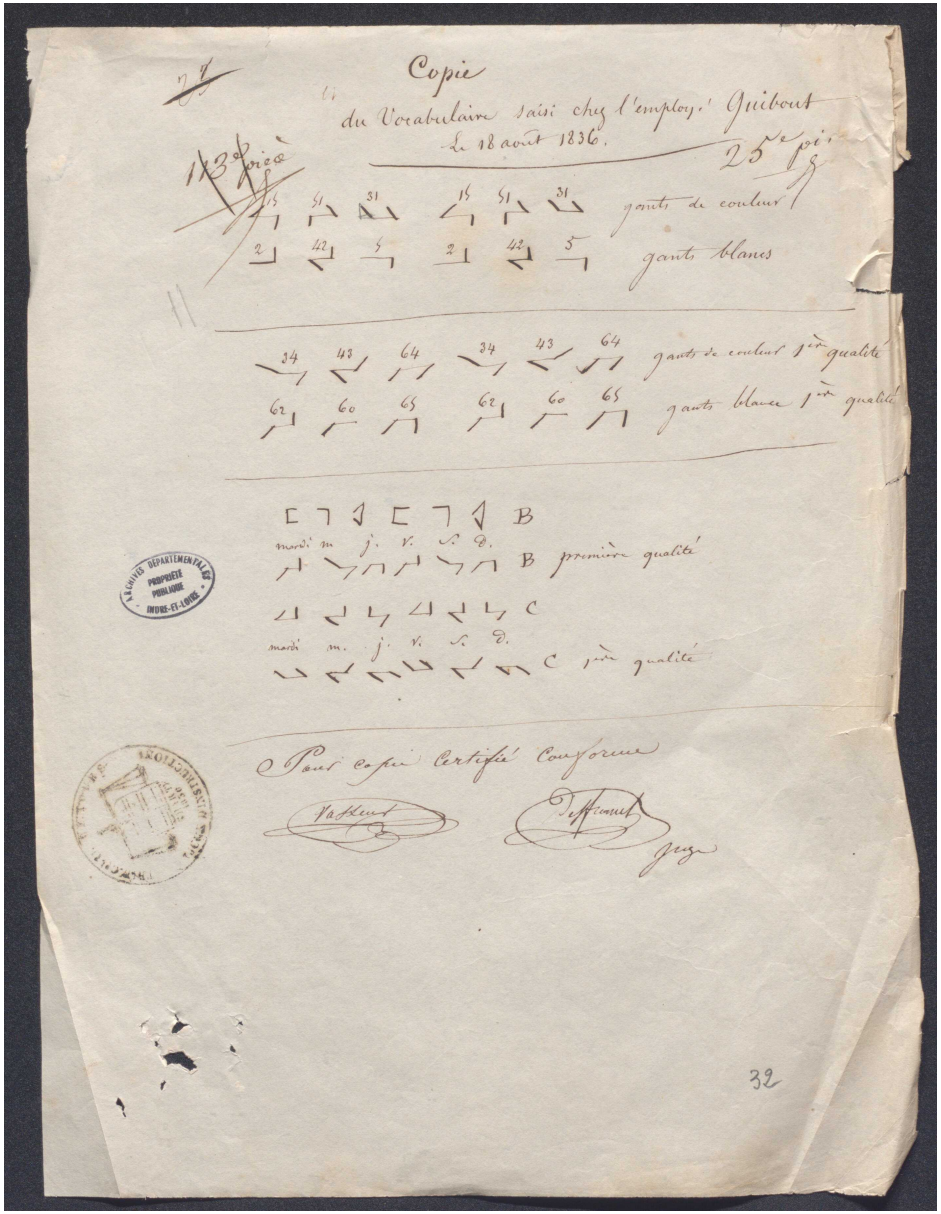
- L'information à transmettre ne peut pas être insérée dans le corps du message :
  - Les pirates n'ont pas accès aux livres de code.
  - Le décodeur aurait connaissance du message.
- La seule façon pour un opérateur intermédiaire de transmettre une information est par l'introduction de messages de service.

Piratage du télégraphe

Mode opératoire

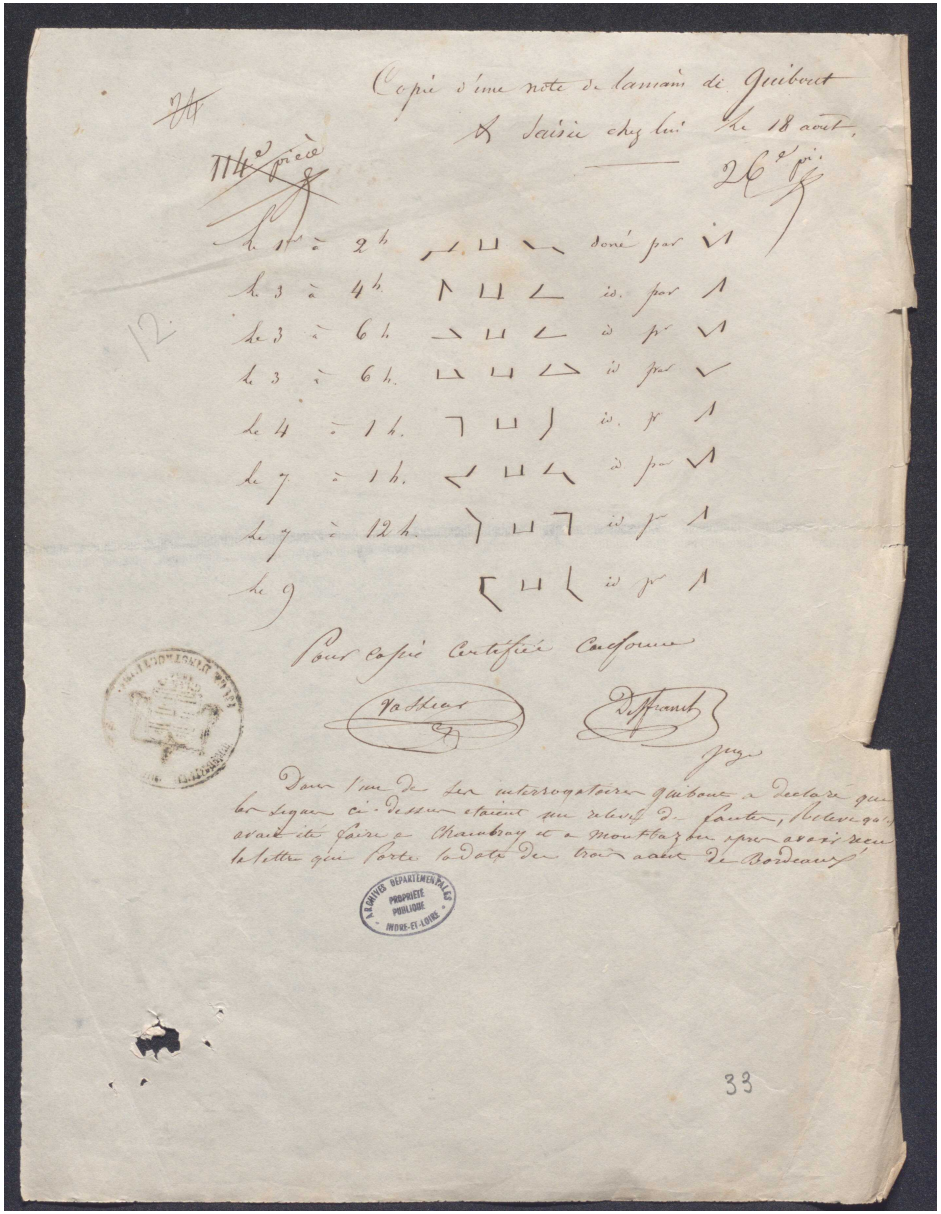
- Nombre de messages différents très limités : *le cours monte*, *le cours descend*
- Les frères Blanc ont mis au point un protocole avec l'employé Guibout :
  - Lorsqu'ils doivent transmettre un message, ils lui envoient une paire de gants.
  - Suivant la couleur des gants, Guibout transmet l'un ou l'autre message.
- Un complice relevait les messages depuis une chambre d'hotel à Bordeaux avec vue sur la tour du télégraphe.

Piratage du télégraphe



Crédits : Archives d'Indre et Loire

Piratage du télégraphe



Crédits : Archives d'Indre et Loire

Piratage du télégraphe

Epilogue

- Les frères Blanc sont condamnés à une amende pour corruption de fonctionnaires.
- Ils fondent le casino de Monte-Carlo.
- Toutes les pièces du procès ont été digitalisées par les Archives d'Indre et Loire et sont disponibles en ligne.
- Les liens seront publiés sur Moodle.

Télégraphie électrique

Petite chronologie

- 1832 : code Morse.
- 1838 : Première ligne entre Londres et Birmingham.
- 1851 : Câble sous-marin France-Angleterre.
- 1858 : Câble transatlantique.
- 1901 : Première liaison transatlantique sans fil (G.Marconi : prix Nobel 1909).
- 1904 : Communications maritimes sans fil.
- 1914 : Première utilisation du signal S.O.S. par le Titanic.

Télégraphie électrique

Principe

- Les signaux sont propagés par l'intermédiaire de fils électriques.
- Un signal est défini par la présence ou l'absence de courant.
- Les prémices du binaire.

Télégraphe à cadran (1850) : émetteur

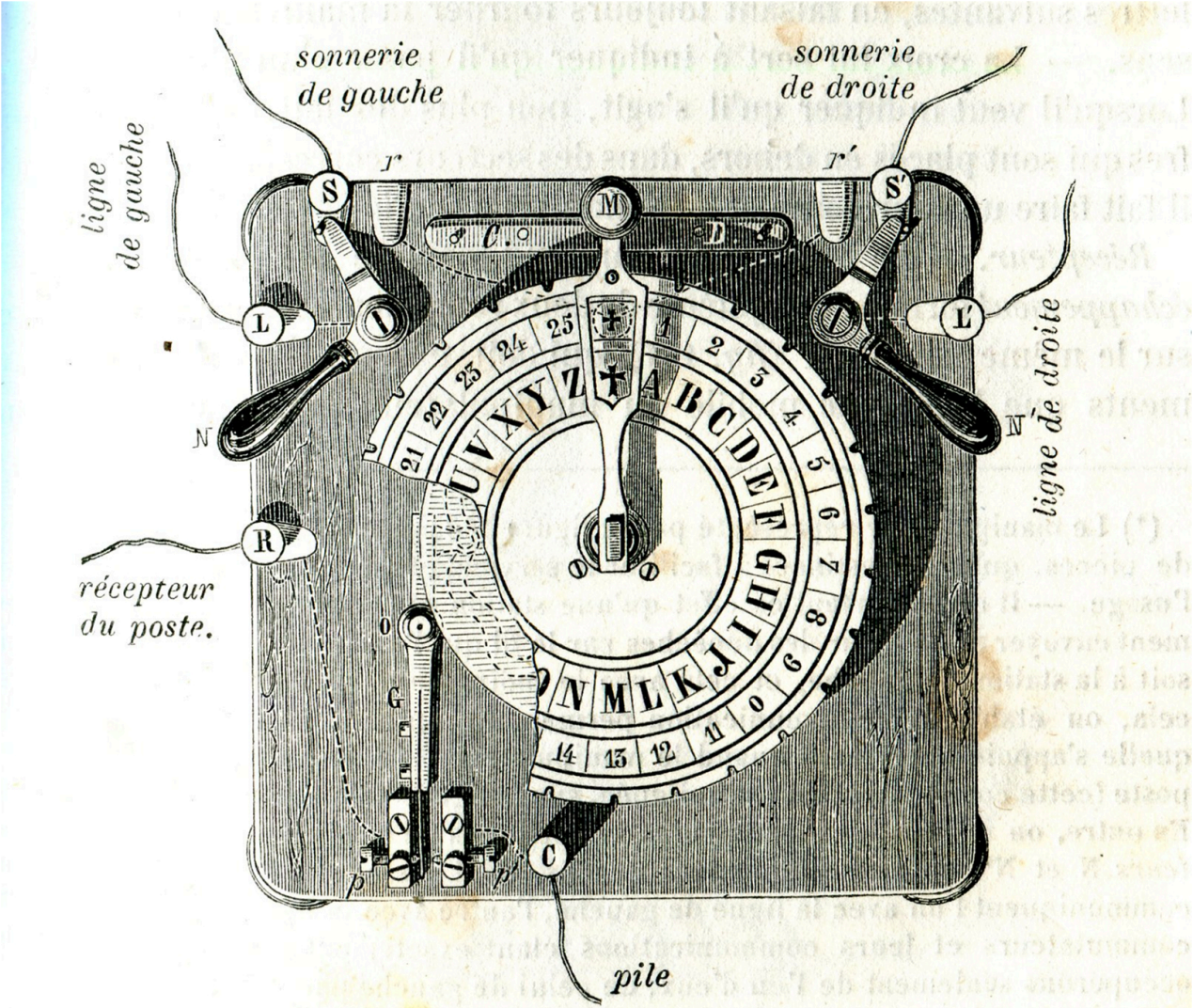


Fig. 460. — Manipulateur du télégraphe à cadran.

Crédits : Physique Drion et Fernet

Télégraphe à cadran : récepteur

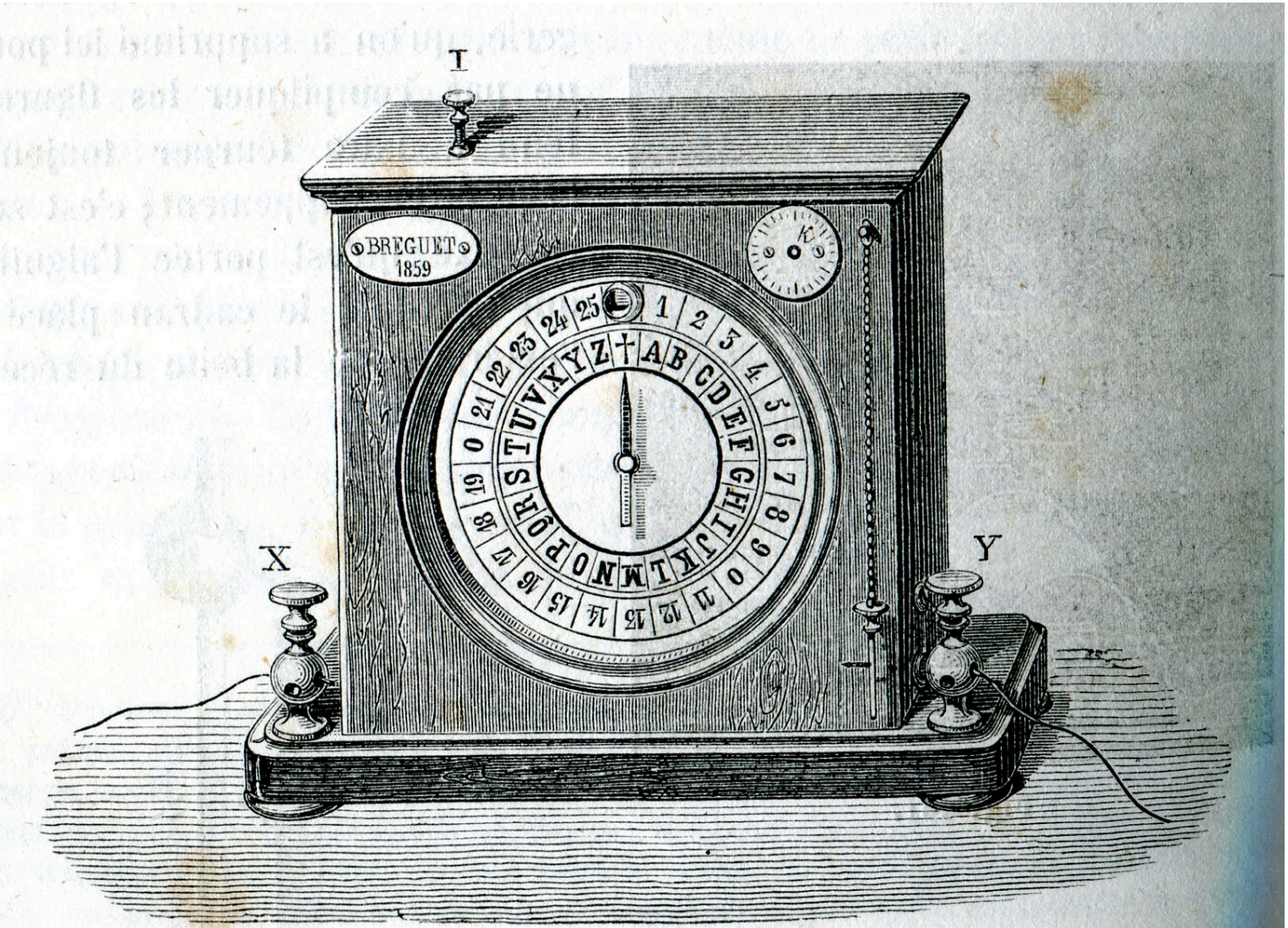


Fig. 463. — Récepteur du télégraphe à cadran.

Crédits : Physique Drion et Fernet

Télégraphe à cadran

Principe

- L'émetteur place la fenêtre de la manivelle en face de la lettre qu'il veut transmettre.
- Pour chaque mouvement de la manivelle, un signal électrique est transmis.
- Ces signaux électriques font bouger l'aiguille du récepteur.
- Ce qui est transmis, c'est la distance entre deux lettres consécutives (un peu comme le jpeg...).
- On aurait pu positionner ces lettres pour limiter le nombre moyen de mouvements ...
- Basculement lettres/nombres : faire un tour complet.

Code Morse (1832 ...)

Description

Les lettres et les chiffres sont codés par une suite de *points* et de *traits* ou bien de signaux *courts* ou *longs*.
L'espace entre le codage d'un caractère et celui du caractère suivant est plus long que celui entre deux signaux d'un même caractère.

Avantages

- Facile à transmettre :
  - Résistance au bruit
  - La version filaire n'a besoin que d'un fil.
- Rapidité de transmission (après apprentissage), tant en émission qu'en réception.

Télégraphe Morse

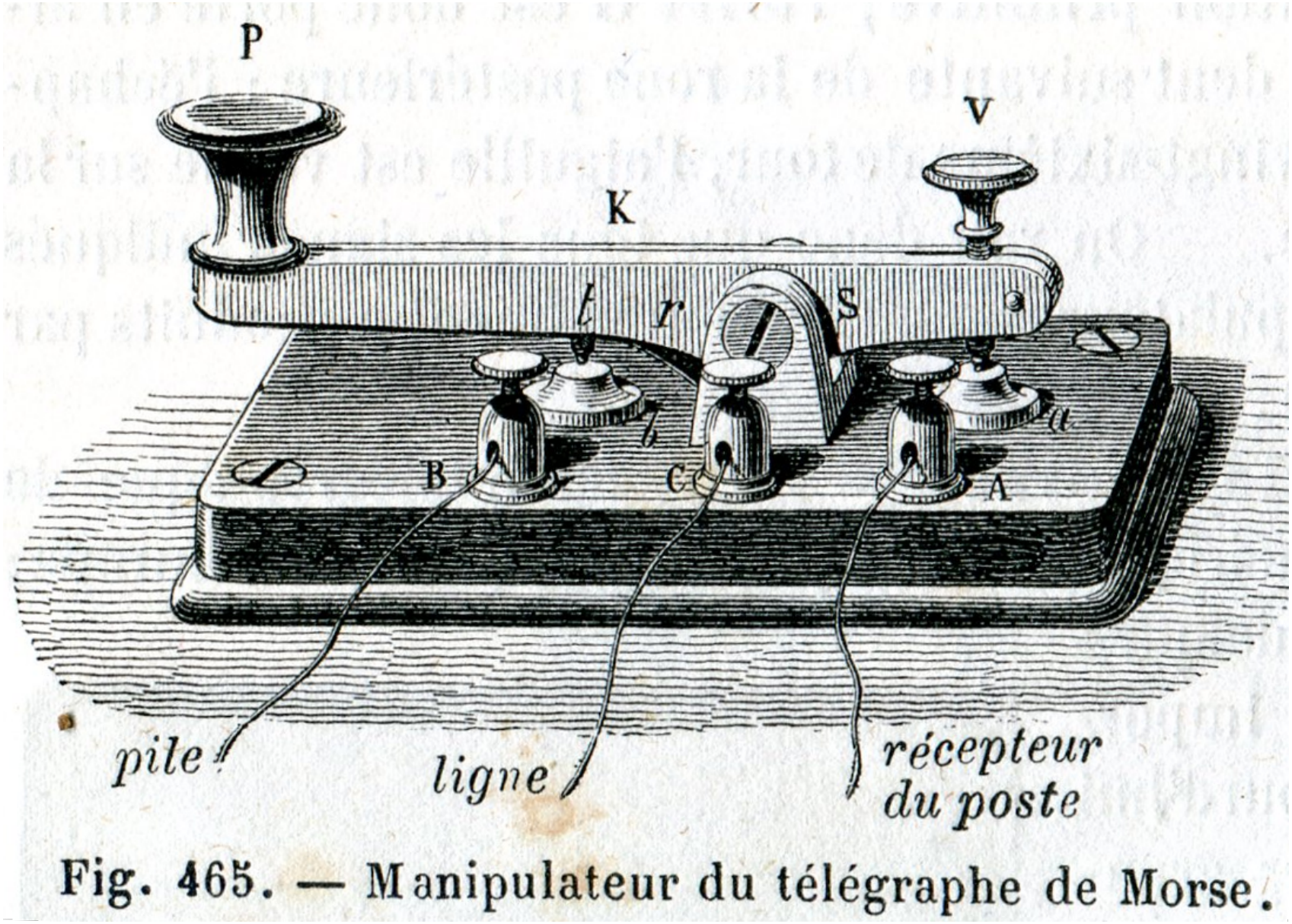


Fig. 465. — Manipulateur du télégraphe de Morse.

Crédits : Physique Drion et Fernet

Télégraphe Morse : récepteur

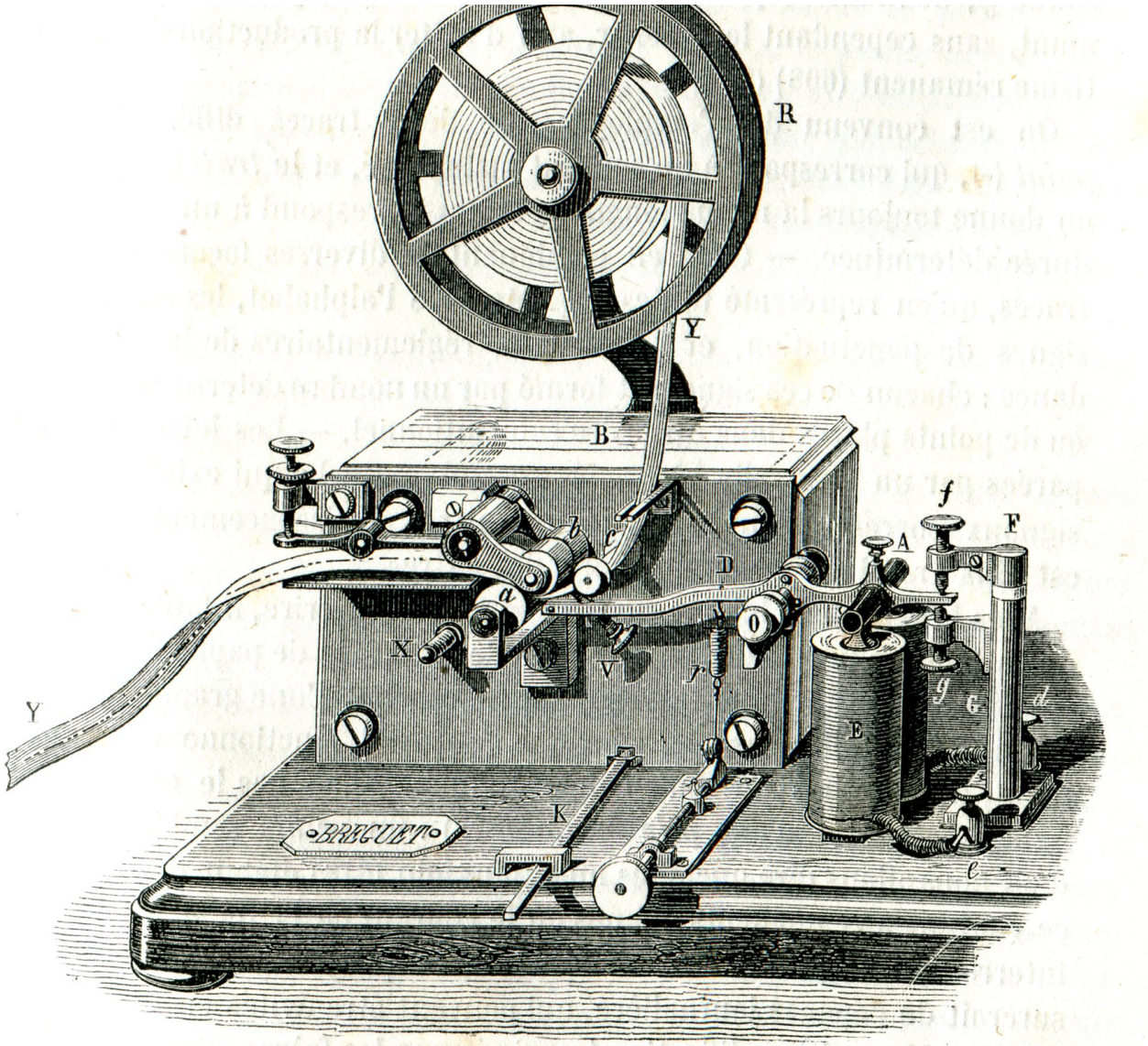
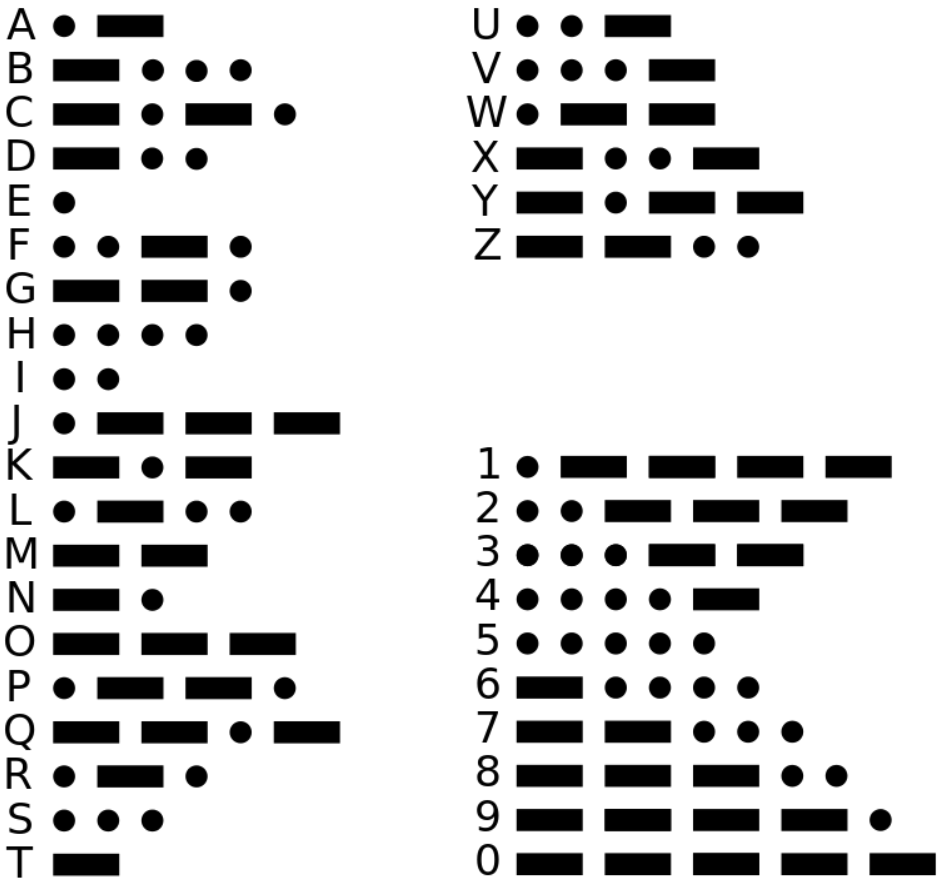


Fig. 467. — Récepteur du télégraphe de Morse.

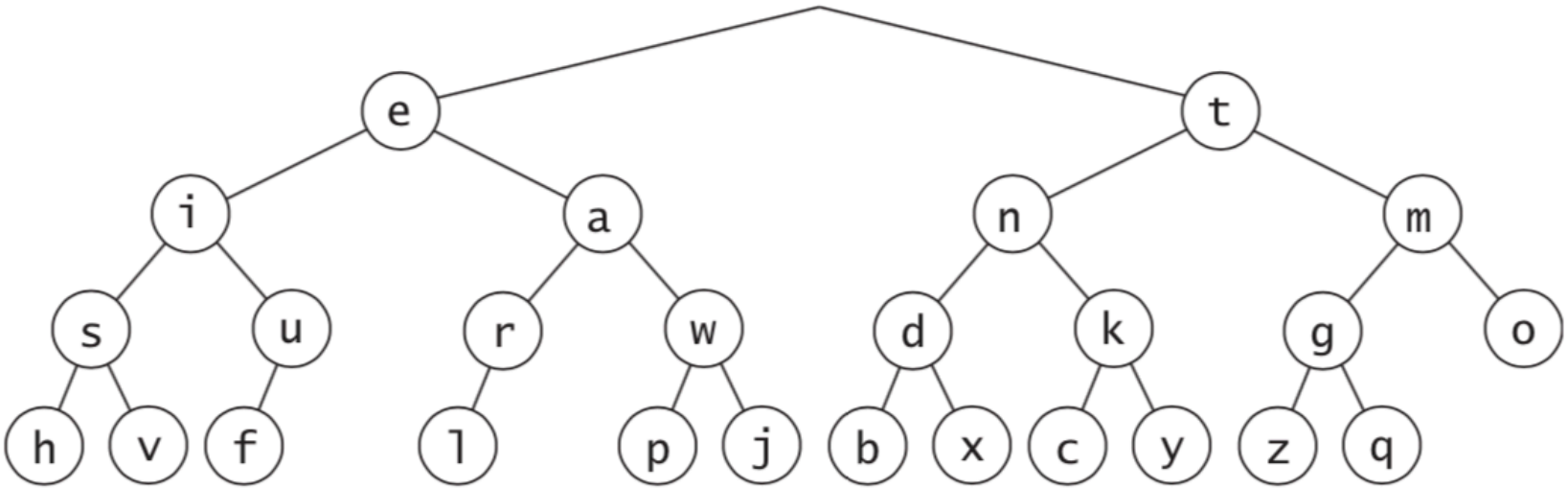
Crédits : Physique Drion et Fernet

Code morse international

1. Un tiret est égal à trois points.
2. L'espacement entre deux éléments d'une même lettre est égal à un point.
3. L'espacement entre deux lettres est égal à trois points.
4. L'espacement entre deux mots est égal à sept points.



Code Morse : sous forme d'arbre



- Descendre vers la gauche : écrire un point.
- Descendre vers la droite : écrire un trait.
- Fréquence des lettres en anglais : E,T,A,O,I,N,S,R,H,D,L,U,C,M,F,Y,W,G,P,B,V,K,X,Q,J,Z
- Les lettres fréquentes ont une écriture plus courte (sauf le M ???).

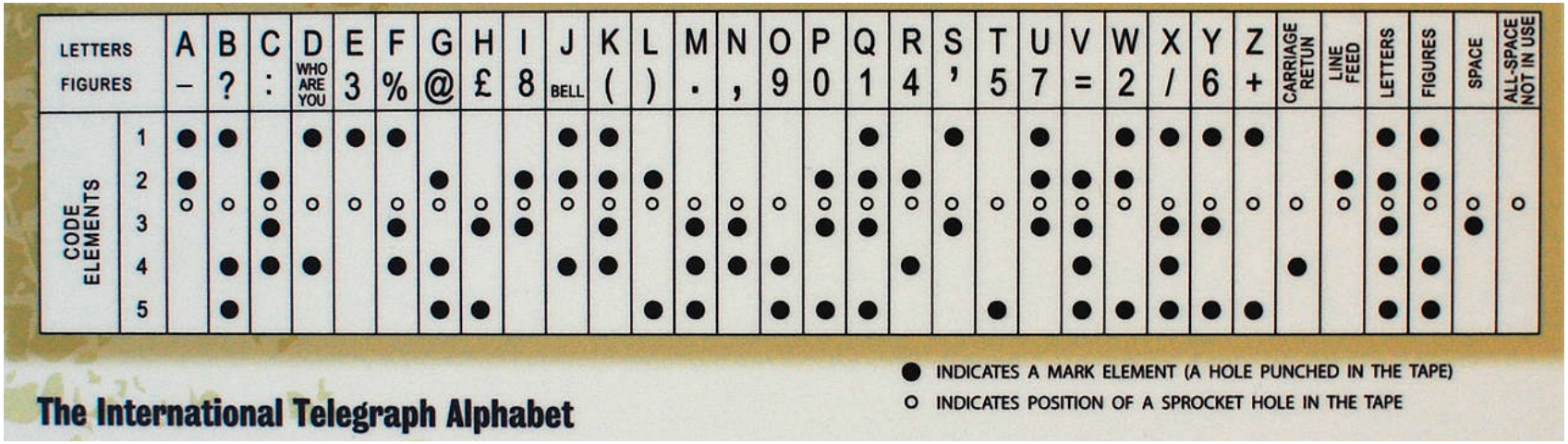
Crédits : Zsuzsanna Dianovics

Code Baudot et Telex

Contexte

- La bande passante devenant plus large, on peut envoyer plusieurs signaux en même temps.
- On peut alors passer d'un signal de largeur 1 (opérateur Morse), à quelque chose de plus large (5 signaux à la fois).
- Initialement, les caractères étaient tapés sur un clavier de machine à écrire, et retranscrits sur un ruban perforé, à décoder à l'arrivée.
- Le système a été amélioré en plaçant une seconde machine à écrire à la destination.
- On définit un nouveau codage : le code Baudot (1874).

Code Baudot



LETTERS / FIGURES	A –	B ?	C :	D WHO ARE YOU	E 3	F %	G @	H £	I 8	J BELL	K (	L)	M .	N ,	O 9	P 0	Q 1	R 4	S '	T 5	U 7	V =	W 2	X /	Y 6	Z +	CARRIAGE RETUN	LINE FEED	LETTERS	FIGURES	SPACE	ALL-SPACE NOT IN USE
CODE ELEMENTS 1	●	●		●	●	●				●	●						●		●		●		●	●	●	●			●	●		
CODE ELEMENTS 2	●	○	●	○	○	○	●	○	●	●	●	●	○	○	○	●	●	●	○	○	●	●	●	○	●	○	○	○	●	●	○	○
CODE ELEMENTS 3	○	○	●	○	○	●	○	●	●	○	●	○	●	●	○	●	●	○	●	○	●	●	○	●	○	●	○	○	○	●	●	○
CODE ELEMENTS 4		●	●	●	○	●	●	○	○	●	●	○	●	●	●	○	○	●	○	○	○	●	○	●	○	○	●	○	○	●	○	○
CODE ELEMENTS 5		●		○	○	○	●	●	○	○	○	●	●	○	●	●	●	○	○	●	○	●	●	●	●	●	○	○	○	○	○	○

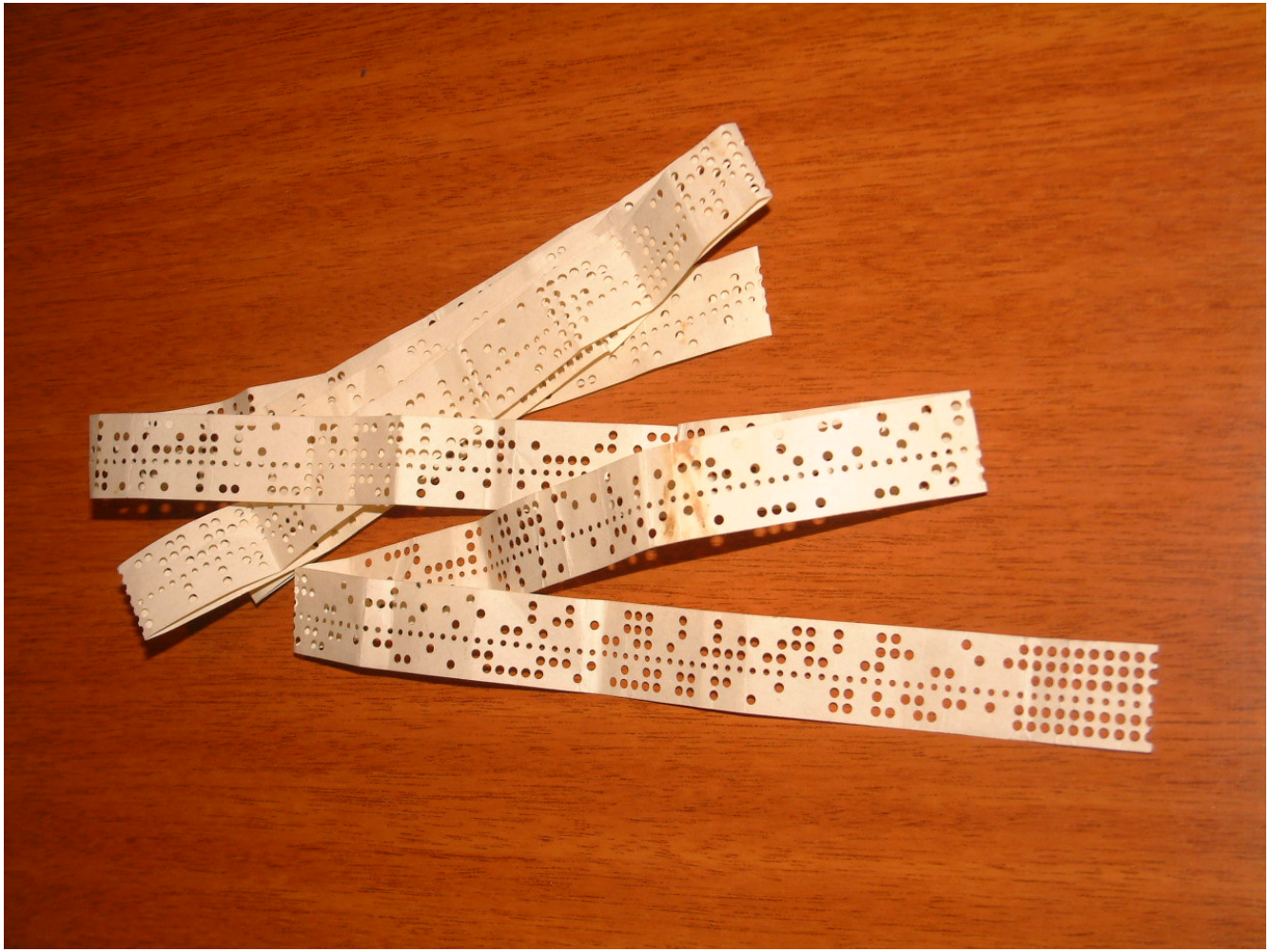
● INDICATES A MARK ELEMENT (A HOLE PUNCHED IN THE TAPE)
○ INDICATES POSITION OF A SPROCKET HOLE IN THE TAPE

The International Telegraph Alphabet

Code	Lettre	Symbole
11000	A	-
...	...	...
01101	P	0
...	...	...
11100	U	7
...	...	...
11111	Mode lettre	
11011	Mode symbole	

Crédits : Inconnu

Code Baudot

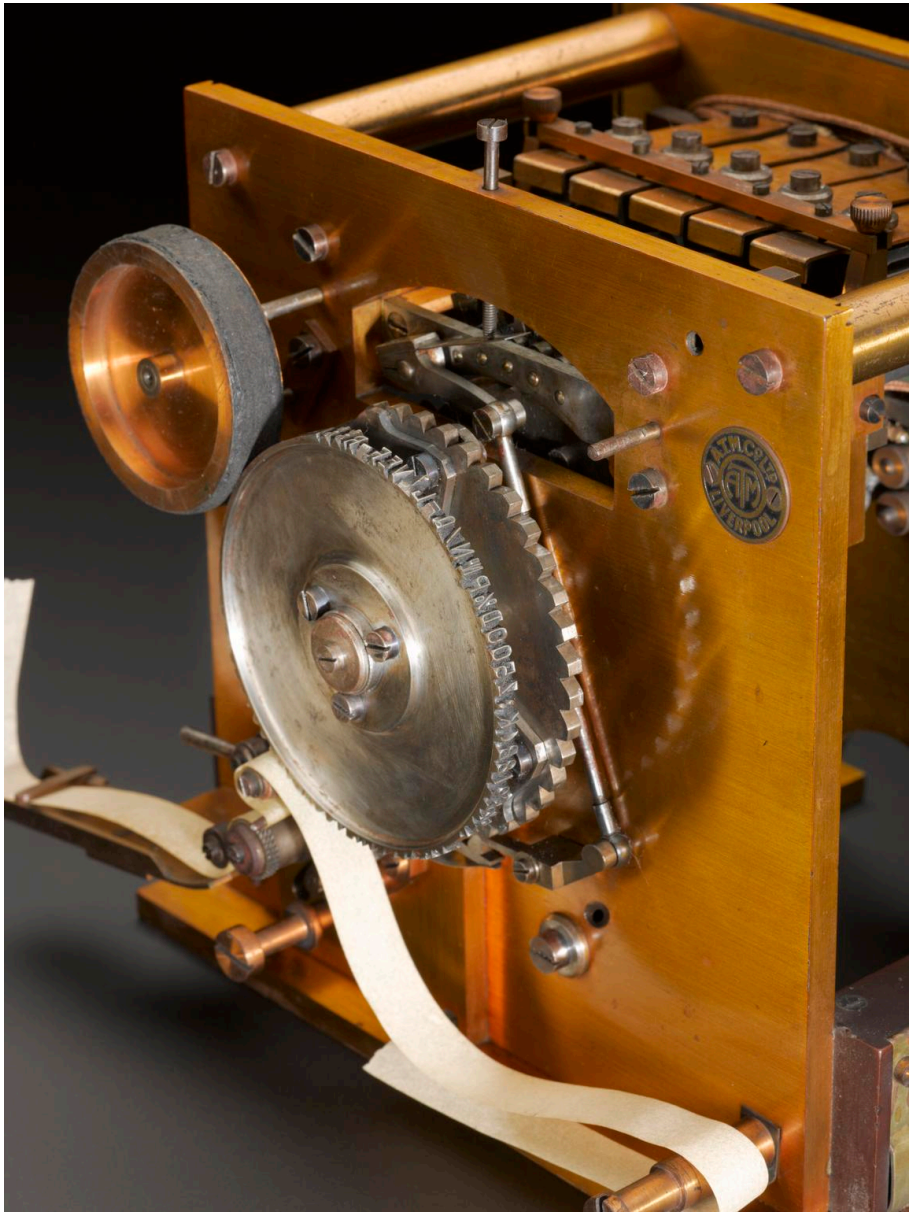


Crédits : Wikipedia

Télégraphe Baudot : émetteur

Crédits : Science Museum Londres

Télégraphe Baudot : récepteur



Crédits : Science Museum Londres

Code Baudot

Caractéristiques et remarques

- (Premier) code binaire de longueur fixe.
- Cinq bits, donc 32 caractères différents. Comment coder 26 lettres, 10 chiffres et quelques symboles ?
- Les codes 11111 et 11011 permettent de basculer d'un alphabet à l'autre.
- On a déjà vu le même principe appliqué au sémaphore ⇑.

Code ASCII

Caractéristiques

- American Standard Code for Information Interchange.
- Norme de codage des caractères alphabétiques.
- Publiée en 1963.
- Permet un codage uniforme et transportable de fichiers textes.
- Codage de 128 caractères imprimables ou non (caractères de contrôle).
- Chaque caractère est codé sur un octet (8 bits) : le bit de poids fort n'est pas utilisé.

Code ASCII

Hex	Value	Hex	Value	Hex	Value	Hex	Value	Hex	Value	Hex	Value	Hex	Value	Hex	Value
00	NUL	10	DLE	20	SP	30	0	40	@	50	P	60	`	70	p
01	SOH	11	DC1	21	!	31	1	41	A	51	Q	61	a	71	q
02	STX	12	DC2	22	"	32	2	42	B	52	R	62	b	72	r
03	ETX	13	DC3	23	#	33	3	43	C	53	S	63	c	73	s
04	EOT	14	DC4	24	$	34	4	44	D	54	T	64	d	74	t
05	ENQ	15	NAK	25	%	35	5	45	E	55	U	65	e	75	u
06	ACK	16	SYN	26	&	36	6	46	F	56	V	66	f	76	v
07	BEL	17	ETB	27	'	37	7	47	G	57	W	67	g	77	w
08	BS	18	CAN	28	(	38	8	48	H	58	X	68	h	78	x
09	HT	19	EM	29	)	39	9	49	I	59	Y	69	i	79	y
0A	LF	1A	SUB	2A	*	3A	:	4A	J	5A	Z	6A	j	7A	z
0B	VT	1B	ESC	2B	+	3B	;	4B	K	5B	[	6B	k	7B	{
0C	FF	1C	FS	2C	,	3C	<	4C	L	5C	\	6C	l	7C	|
0D	CR	1D	GS	2D	-	3D	=	4D	M	5D	]	6D	m	7D	}
0E	SO	1E	RS	2E	.	3E	>	4E	N	5E	^	6E	n	7E	~
0F	SI	1F	US	2F	/	3F	?	4F	O	5F	_	6F	o	7F	DEL

Code ASCII : examen de la table

- Les codes sont donnés en hexadécimal : ils vont de $\overline{00}_{16}$ (0) à $\overline{7\mathrm{F}}_{16}$ (127)
- Le premier caractère imprimable est le 32ème, soit $\overline{20}_{16}$ en hexadécimal : c'est l'espace.
- Les chiffres se trouvent dans des emplacements successifs à partir du 48ème rang ($\overline{30}_{16}$). Ils sont en ordre croissant.
- Les lettres minuscules commencent en 97ème position ($\overline{61}_{16}$)
- Les lettres majuscules commencent en 65ème position ($\overline{41}_{16}$)
- Pas de caractères accentués.

Code ASCII : limitations

- Le besoin de standardisation fait qu'on a ignoré les caractères accentués.
- D'un autre côté, le nombre de caractères spéciaux à l'ensemble des langues est plus grand que 256.
- Si on intègre tous les caractères accentués de toutes les langues (partageant l'alphabet latin), on dépasse les 256 caractères.
- Il ne serait plus possible de coder les caractères sur un octet.

Code ASCII : extensions

- Plusieurs extensions de la norme ASCII ont été définies, ajoutant des caractères aux positions non utilisées en ASCII.
- La norme qui concerne l'Europe Occidentale est la norme ISO8859-1 (ou latin-1).

ISO8859-1

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
20		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
30	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
40	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
50	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
60	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
70	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	
80			‚	*ƒ*	„	…	†	‡	ˆ	‰	Š	‹	Œ			
90		‘	’	“	”	•	–	—	˜	™	š	›	œ			Ÿ
A0		¡	¢	£	¤	¥	¦	§	¨	©	ª	«	¬	-	®	¯
B0	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
C0	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D0	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
E0	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
F0	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

ASCII et Python

Trouver un caractère à partir de son code

```
>>> chr(65)
'A'
```

Connaître le code ASCII d'un caractère

```
>>> ord('A')
65
```

Transformer un caractère chiffre en entier

```
>>> ord('7') - ord('0')
7
```

Unicode et UTF-8

Caractéristiques

- Unicode est un projet de codage de tous les alphabets (et d'autres symboles).
- Première publication : 1991.
- Principe : à chaque caractère est associé une valeur, nommée *point de code*
- Créé pour dépasser les limitations des ASCII étendus.
- UTF-8 est une des normes permettant de représenter les *points de code*.
- En UTF-8, un *point de code* (ou caractère) est codé sur 1,2,3 ou 4 octets.
- Il permet de coder 1.112.064 caractères différents.
- UTF-8 est compatible avec les codes ASCII standard et ISO8859-1.

Unicode : codage des caractères

Octets	Premier	Dernier	Octet 1	Octet 2	Octet 3	Octet 4
1	U+0000	U+7F	0xxxxxxx			
2	U+0080	U+07FF	110xxxxx	10xxxxxx		
3	U+0800	U+FFFF	1110xxxx	10xxxxxx	10xxxxxx	
4	U+10000	U+10FFFF	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

Exemples

Codage sur un octet

- La lettre **H** a pour point de code $\overline{0048}_{16}$
- Soit en binaire : 0100 1000
- Codé en UTF-8 sur 1 octet :

$$\overbrace{\mathbf{0}100}^{4}\ \overbrace{1000}^{8}$$

- Les caractères ASCII purs conservent leur code et leur écriture sur un octet.

Exemples

ASCII étendu, codage sur deux octets

- La lettre ASCII étendu **ã** a pour point de code $\overline{00\mathrm{E}3}_{16}$
- Soit en binaire : 1110 0011
- Codé en UTF-8 sur 2 octets :

$$\overbrace{\mathbf{110}0}^{C}\ \overbrace{0011}^{3}\ \overbrace{\mathbf{10}10}^{A}\ \overbrace{0011}^{3}$$

- Cette lettre n'est pas codée sur un octet, sinon il y aurait confusion possible avec la convention de codage en UTF-8

Exemples

Codage sur deux octets

- Le caractère arabe بْ a pour point de code $\overline{\texttt{06FA}}_{16}$
- Soit en binaire : 0110 1111 1010
- Codé en UTF-8 sur 2 octets :

$$\overbrace{\mathbf{1101}}^{D}\ \overbrace{1011}^{B}\ \overbrace{\mathbf{10}11}^{B}\ \overbrace{1010}^{A}$$

Exemples

Codage sur trois octets

- Le caractère gurmukhi ੜ a comme point de code $\overline{\texttt{0A23}}_{16}$
- Soit en binaire : 0000 1010 0010 0011
- Codé en UTF-8 sur 3 octets :

$$\overbrace{\mathbf{1110}}^{E}\ \overbrace{0000}^{0}\ \overbrace{\mathbf{10}10}^{A}\ \overbrace{1000}^{8}\ \overbrace{\mathbf{10}10}^{A}\ \overbrace{0011}^{3}$$

Exemples

Codage sur quatre octets

- L'idéogramme 龕 a comme point de code $\overline{28403}_{16}$
- Soit en binaire : 0 0010 1000 0100 0000 0011
- Codé en UTF-8 sur 4 octets :

$$\overbrace{\mathbf{1111}}^{F}\ \overbrace{\mathbf{0}000}^{0}\ \overbrace{\mathbf{10}10}^{A}\ \overbrace{1000}^{8}\ \overbrace{\mathbf{10}01}^{9}\ \overbrace{0000}^{0}\ \overbrace{\mathbf{10}00}^{8}\ \overbrace{0011}^{3}$$

Unicode, UTF-8, et les logiciels courants

Petit à petit

- Tous les logiciels ne sont pas encore compatibles avec UTF-8 / Unicode.
- L'accès aux caractères Unicode n'est pas le même selon les logiciels.
- Un logiciel donné peut n'implémenter qu'une partie des points de code.

Unicode, UTF-8 : Python

```
import unicodedata
DATA=[ "\U00028403","\U00028405","\U0001d11e","\U0001f600"]
fichier=open("test.txt",'w',encoding='utf-8')
for code in DATA :
 fichier.write(code)
 print( "--->", code," ",unicodedata.name(code))
 fichier.write("\n")

fichier.close()
print("\n lecture")

with open('test.txt', encoding='utf-8') as f:
  i=0
  for line in f:
    #print(repr(line))
    print(i," ",line,end="")
    i=i+1
```

Python : exécution du code

```
---> 鱷    CJK UNIFIED IDEOGRAPH-28403
---> 殺    CJK UNIFIED IDEOGRAPH-28405
---> 𝄞    MUSICAL SYMBOL G CLEF
---> 😀    GRINNING FACE

  lecture
0    鱷
1    殺
2    𝄞
3    😀
>>>
```

Python : Texworks

```
\begin{frame}[fragile]\frametitle{Unicode, UTF-8, et les logiciels courants : Python}
\begin{center}
\begin{block}{Editeur de texte Texworks}
\small
\begin{verbatim}
---> 龕  CJK UNIFIED IDEOGRAPH-28403
---> 戟  CJK UNIFIED IDEOGRAPH-28405
---> 𝄞  MUSICAL SYMBOL G CLEF
---> 😀  GRINNING FACE

 lecture
0  龕
1  戟
2  𝄞
3  😀

\end{verbatim}
\end{block}
\end{center}
   \end{frame}
```

Unicode, UTF-8, et les logiciels courants

Compilation du transparent (slide)

```
--->     CJK UNIFIED IDEOGRAPH-28403
--->     CJK UNIFIED IDEOGRAPH-28405
--->     MUSICAL SYMBOL G CLEF
--->     GRINNING FACE

 lecture
0
1
2
3
```