NLP Extraction & Tagging Test Pack

Document multilingue (français, anglais, arabe) conçu pour valider : extraction PDF, segmentation layout, détection de langue, tokenisation, POS tagging (NN/VB/JJ...), lemmatisation, NER et détection de bruit.

Mode d’emploi : exécute ton pipeline sur ce PDF. Compare la sortie (sentences_layout / chunks / tokens) avec les attentes ci-dessous. Le header/footer est volontairement répétitif (avec chiffres) pour tester la détection de bruit et le masquage de digits.

- Easy : phrases simples + dates (2026-02-12, 12/02/2026), emails, URLs, nombres, ponctuation standard.
- Medium : listes numérotées, sections (1.1 / a. / (iv)), camelCase, ALLCAPS, mots collés chiffres/lettres, hyphenation sur saut de ligne.
- Complex : micro-table header multi-colonnes, tableau de lignes (facture), colonnes alignées, séparateurs, blocs adresse, multi-colonnes.
- Noise/Edge : répétitions (header/footer), variations de chiffres, longues séquences de ponctuation, mélange scripts (AR/FR/EN).

Quick sanity strings (should be easy to spot)

Emails: support+qa@example.co.uk, facturation@entreprise.dz, test.user+tag@sub.domain.org

URLs: https://example.com/path?x=1&y;=2#frag — http://intranet.local/docs/v1/index.html

IDs: customerID42, invoiceNumber2026A-0007, HTTPStatus200OK, sha1=da39a3ee5e6b4b0d3255bfef95601890afd80709

English — Complex sentences, contractions, and tricky tokens

Although the committee, which had been debating for months, finally approved the proposal, the CEO insisted that we shouldn't've shipped anything until the QA lead had signed off.

John's colleague said: "If the Joneses' router isn't working, they won't be able to access the U.S. lab's Ph.D. dataset."

Edge cases: 20-year-old, state-of-the-art, non-functional, re-enter, O'Reilly (do not split like a contraction).

Dates & formats: 2026-02-12, 12/02/2026, Feb 12, 2026, 12 Feb 2026. Numbers: 1,234.56 and 1234,56 (comma decimal).

Contact: support+qa@example.co.uk; phone +44 20 7946 0958; ticket #A-00981/2026; ref: INV-EN-0003.

Address block: 221B Baker Street, London NW1 6XE, United Kingdom. Ship-to: 500 Fifth Ave, New York, NY 10110, USA.

Punctuation torture: Wait... what?! This is fine!!! (maybe); brackets (a) [b] {c}; quotes “smart” and 'plain'.

Glued tokens: abc123def, x=1&y;=2, path/to/file_v2.0.tar.gz, user_name@host, 192.168.1.42:8080.

Numbered / alpha / roman list (layout segmentation)

1. First item: Make sure tokenization keeps "shouldn't've" as a chain or decomposes deterministically.

1.1 Sub-item: Multi.level.numbering should start new sections.

a) Alpha item: The label-only line below should be merged with its content.

(iv) Roman item: Check that (iv) is recognized as a section start.

2) Second item: A long hyphenated word split across lines: internationalisation and multilingualization.

Français — Sections, élisions, tirets, adresses

Article 1 — Objet : Ce document sert à tester l'extraction et l’étiquetage grammatical. S’il n’est pas robuste, tu verras beaucoup de NN par défaut.

1.1 Détection d’élisions : d'Air France-KLM, l'aéroport, qu'il n'est pas, jusqu'à, lorsqu'il arrive, puisqu'on l'a dit.

a) Tokens piégeux : aujourd'hui (ne pas découper), presqu'île, quelqu'un, e-mail, co-fondateur, va-nu-pieds.

(iv) Ponctuation : « guillemets », points-virgules ; tirets — et parenthèses (test).

Coordonnées : facturation@entreprise.dz, tel +213 (0)21 12 34 56, site https://exemple.dz/portail?ref=AB12.

Adresse : 12, Rue Didouche Mourad, BP 123, 16000 Alger, Algérie. Livraison : Zone Industrielle Oued Smar — Hangar 7A.

Montants : Total HT = 1 234,56 DZD ; TVA (19%) = 234,57 ; Total TTC = 1 469,13.

Phrase complexe : Si, malgré l'avis du comité, le directeur — qui était pourtant d'accord — change d’idée, alors il faudra re-documenter l’API, re-tester et re-déployer.

Bloc multi-colonnes (clé / valeur) — observe les grands espaces à l’extraction

Client: Société Exemple SARL	Commande: PO-7781
N° Client: CL-00042	Référence: REF/2026/AL-19
Email: facturation@entreprise.dz	Téléphone: +213 21 12 34 56
Date: 15/02/2026	Statut: PAYÉ

Tables & aligned columns — Invoice-style (mixed FR/EN)

Below are tables designed to trigger table detection, micro-table detection, and wrapped description lines. Separators like '----' are included on purpose.

INVOICE	DATE	PO
INV-2026A-0007	2026-02-15	PO-7781
Currency: DZD	VAT: 19%	Terms: Net 30

Qté	Désignation / Description	P. unitaire	TVA	Montant
2	Service de traitement NLP (pack "pro")	45000,00	19%	107100,00
1	Audit + rapport détaillé — includes: tokenisation, POS, lemma, NER, noise	65000,00	19%	77350,00
3	Support (email: support+qa@example.co.uk) / hotline +213 21 12 34 56	12000,00	19%	42840,00
1	Livraison express (221B Baker Street -> Alger) / tracking: 192.168.10.42:8080	8000,00	0%	8000,00
	Note: internationalisation + multilingualization appear with manual line breaks.			
	Extra: path/to/file_v2.0.tar.gz ; sha1=da39a3ee... ; HTTPStatus200OK.			

Total HT	180000,00
TVA (19%)	34200,00
Total TTC	214200,00

العربية — نص عربي + أرقام + عناوين + علامات ترقيم

هذا الملف مُصمّم لاختبار استخراج النص العربي، واكتشاف اللغة، والتجزئة، والإسناد النحوي، والتعرّف على الكيانات.

سؤال: هل تعمل الخوارزمية بشكل صحيح؟ إذا لم تعمل، ستلاحظ ضجيجاً في النتائج، أو أخطاءً في التقسيم.

بيانات الاتصال: بريد إلكتروني test.user+tag@sub.domain.org ، هاتف: +213 21 12 34 56

التاريخ: 15/02/2026 ، رقم الفاتورة: INV-AR-0005 ، المبلغ: 1,234.56

العنوان: ١٢ شارع ديدوش مراد، الجزائر 16000، الجزائر.

ملاحظة: الرموز العربية مثل (، ؛ ؟) يجب أن تُحفظ كما هي.

كتلة أعمدة بمحاذاة (تُشبه جدولاً)

الكمية	الوصف	السعر	الضريبة	الإجمالي
2	خدمة تحليل نصوص	45000.00	19%	107100.00
	سطر تابع للوصف بدون أرقام كثيرة			
1	شحن سريع — Baker Street -> الجزائر	8000.00	0%	8000.00

Mixed-script stress line

AR/FR/EN in one line: الجزائر Algiers — Rue Didouche Mourad — customerID42 — بريد: facturation@entreprise.dz — URL: https://example.com/x?y=2