

Biblio-prospective sur la transcription automatique (style transkribus)

Transkribus est une IA qui permet de numériser et transcrire des documents historiques automatiquement.

Voici un exemple d'utilisation de cette application sur un texte allemand, qui est vraisemblablement bien transcrit :

The screenshot shows the Transkribus software interface. On the left is a scanned page of handwritten German text. The right side displays the transcribed text with numbered annotations. The annotations correspond to the following steps:

- 1 Germteig.
- 2 In laue Milch, Germ hinein, und etwas
- 3 Mehl, von den 50 dkg Mehl verschrudeln,
- 4 u. am Herd lau machen u. aufgehen lassen.
- 5 50 dkg Mehl 1-2 dkg Germ
- 6 ½ l laue Milch, salzen
- 7 Verfeinerung 2 Eier 4 dkg Butter od. Fett
- 8 Vanille, Zitronengeschmack.
- 9 Für Milchbrot, Kipferl, Gugelhupf, Strudel.
- 10 Mürber Teig.
- 11 25 dkg Mehl ½0 l Wasser, 1 Ei, Zucker, salzen,
- 12 10 dkg Butter od. Fett, Verfeinerung
- 13 2 Dotter, (mehr Butter—15 dkg) statt
- 14 Wasser 1/0 l Rahm, Teig machen, an
- 15 kühlem Orte ½ Stunde rasten lassen.
- 16 brei.
- 17 2 l Milch =20 dkg Grieß, Zitronengeschmack,
- 18 für.
- 19 salzen, Zuckern, gestürzten Grieß — 24 dkg Grieß.

Puis, sur le texte de l'enfant prodigue du canton de Caraman avec une aide automatique de type "français" :

The screenshot shows the Transkribus software interface. On the left is a scanned page of handwritten Breton text. The right side displays the transcribed text with numbered annotations. The annotations correspond to the following steps:

- 1 Cantor de Ocruman
- 2 Commung de Caraman.
- 3 296
- 4 enjan proudione.
- 5 1.
- 6 Uen omé nalo qué dus fils. Dé pus journe
- 7 diqué à soun pairé: "Es lems que siosqui mour mestl,
- 8 etqu'avry dargent. Calqué pousqui m'en ana et qui
- 9 bejoy de pais. Partial. fas boste bé, et dounamme eo
- 10 aue mes dibut: O, mour fil diauét le payré, coumme
- 11 bouldras. Es un maissant et séras purnt n Apreps
- 12 dubeisaudl un tiroué, ppatalgét soun bie et en fasquet
- 13 dos poussions éoalos.
- 14 2. aove dé jaun apiifs, le mairent til s'en
- 15 anquel del bilatgé en fesan lé fier, et san diré adious
- 16 à diqus. bzarest pla di bosguéés, dé landos, dé
- 17 Librées, à benguil dins unr grando bils, our

Ou encore... :

The screenshot shows the Transkribus software interface. On the left, there is a handwritten document page with various numbers (1-20) written above the text. The main text area contains the following transcription:

1. a do samo mans
2. eman as
3. Dora man.
4. fi
5.
6. e
7. enicui
8. novel que
9. Alo onb nsala q ava fala pat ens
10. Cq laval iculic dele na que lorqua anotn meilal
11. qu cioiq a cualal Calqui lovi quamli en eina el que
12. beo asials dfalaa borsit Qeb dounaman eo-
13. qui mo dilus Qonova bil digidel cusil e ciorfa
14. e mlbras deu valobanh e srras dnsi oha
15. dubrba qu ddel dar foreusta cufalqrs obat Qebnaiquil
16. Sor o sussou quie son
17. Aqais Al ocní c pbl E mda nos fal
18. cone al lasil ania Ooreb ven du tibln
19. tqor be cullaLa olorquula I anordi-
20. aqq sorquil dia dorro qe reinao bibar ov

On remarque cependant que la traduction n'est pas exactement juste dans notre cas du texte de l'enfant prodigue. On en déduit donc qu'il est nécessaire de réaliser une relecture afin de corriger le texte transcrit, ou bien de créer/utiliser une aide "occitane" (qui est d'ailleurs proposée par le site "Transkribus").

Nous allons donc voir comment procéder avec un style Python à la place, et si la traduction devient plus "accurate".