

Philologie  
numérique :  
constituer un  
corpus

Jean-Baptiste  
Camps

Un outil "tout  
en un" :  
Transkribus

Pas-à-pas avec  
ScanTailor et  
OCRopy

Traitement des  
images

Analyse de la mise en  
page

Reconnaissance des  
écritures manuscrites

# Philologie numérique : constituer un corpus

## *Atelier Humanités numériques*

Jean-Baptiste Camps

École nationale des chartes | Paris, Sciences & Lettres

Casa de Velázquez  
Madrid, 9 octobre 2018

# Acquisition du texte

Philologie  
numérique :  
constituer un  
corpus

Jean-Baptiste  
Camps

Un outil "tout  
en un" :  
Transkribus

Pas-à-pas avec  
ScanTailor et  
OCRopy

Traitement des  
images

Analyse de la mise en  
page

Reconnaissance des  
écritures manuscrites

Dans la constitution d'un corpus de textes, la première phase est bien sûr l'acquisition du contenu des textes envisagés.

**Transcription** des témoins, selon les critères scientifiques du projet (transcriptions allographétiques, graphématiques, normalisées ; édition critique ; etc.). Méthode souvent la plus sûre, mais aussi la plus lente ;

**"Transcription" assistée par ordinateur** en utilisant un algorithme permettant la reconnaissance optique de caractères (*optical character recognition* ou OCR) imprimés, ou la reconnaissance des écritures manuscrites (*handwritten text recognition* ou HTR).

**Téléchargement** de textes depuis des corpus en lignes, des sites d'édition électronique, des bases de données d'éditeurs, etc.

# Reconnaissance des écritures manuscrites

Philologie  
numérique :  
constituer un  
corpus

Jean-Baptiste  
Camps

Un outil "tout  
en un" :  
Transkribus

Pas-à-pas avec  
ScanTailor et  
OCRopy

Traitement des  
images

Analyse de la mise en  
page

Reconnaissance des  
écritures manuscrites

## Reconnaissance optique des caractères (imprimés) *Optical character recognition* (OCR)

- "problème résolu" de l'informatique ;
- aisé d'obtenir des taux d'erreur caractère (CER)  $< 2\%$  ;
- outils libres : Tesseract 4, ... ;
- existence de modèles génériques (par langue).

## Reconnaissance des écritures manuscrites

### *Handwritten text recognition* (HTR)

- très peu fonctionnel jusqu'à ces dernières années ;
- nouveaux développements : IA (réseaux de neurone récurrents LSTM...);
- outils libres : OCRopy, ...
- modèles spécifiques à entraîner (pour chaque main, écriture,...).

# Les étapes

Philologie  
numérique :  
constituer un  
corpus

Jean-Baptiste  
Camps

Un outil "tout  
en un" :  
Transkribus

Pas-à-pas avec  
ScanTailor et  
OCRopy

Traitement des  
images

Analyse de la mise en  
page

Reconnaissance des  
écritures manuscrites

- ① traitement des images ;
- ② analyse de la mise en page et identification des lignes ;
- ③ reconnaissance des caractères ;
- ④ d'éventuels post-traitements, visant à améliorer les résultats.

# Plan

Philologie  
numérique :  
constituer un  
corpus

Jean-Baptiste  
Camps

Un outil “tout  
en un” :  
Transkribus

Pas-à-pas avec  
ScanTailor et  
OCRopy

Traitement des  
images

Analyse de la mise en  
page

Reconnaissance des  
écritures manuscrites

## 1 Un outil “tout en un” : Transkribus

## 2 Pas-à-pas avec ScanTailor et OCRopy

- Traitement des images
- Analyse de la mise en page
- Reconnaissance des écritures manuscrites

# Transkribus

Philologie  
numérique :  
constituer un  
corpus

Jean-Baptiste  
Camps

Un outil "tout  
en un" :  
Transkribus

Pas-à-pas avec  
ScanTailor et  
OCRopy

Traitement des  
images

Analyse de la mise en  
page

Reconnaissance des  
écritures manuscrites



- développé par un consortium de recherche européen (projet READ ; Univ. Innsbruck et al.) ;
- financé par la Commission Européenne (Horizon 2020) ;
- permet de charger des images, analyser la mise en page, segmenter...
- opérations réalisées sur les **serveurs de Transkribus.**

# Installer Transkribus

Philologie  
numérique :  
constituer un  
corpus

Jean-Baptiste  
Camps

Un outil "tout  
en un" :  
Transkribus

Pas-à-pas avec  
ScanTailor et  
OCRopy

Traitement des  
images

Analyse de la mise en  
page

Reconnaissance des  
écritures manuscrites

- Se rendre sur :  
`https://transkribus.eu/Transkribus/`;
- se connecter (ou créer un compte) et télécharger le logiciel ;
- extraire l'archive ;
- lancer le logiciel (`Transkribus.sh` ou `Transkribus.exe`) ;
- tutoriel : `https://transkribus.eu/wiki/images/7/77/How\_to\_use\_TRANSKRIBUS\_-\_10\_steps.pdf` ;
- wiki : `https://transkribus.eu/wiki/index.php/Main\_Page`.

# T.P. Transkribus

Philologie  
numérique :  
constituer un  
corpus

Jean-Baptiste  
Camps

Un outil "tout  
en un" :  
Transkribus

Pas-à-pas avec  
ScanTailor et  
OCRopy

Traitement des  
images

Analyse de la mise en  
page

Reconnaissance des  
écritures manuscrites

- ❶ se connecter ;
  - ❷ importer un document (choisir un ou deux fol. du dossier digby\_23) ;
  - ❸ lancer l'analyse de la mise en page ;
  - ❹ corriger la segmentation
    - régions de texte,
    - zone de ligne,
    - ligne de référence,
    - mots,... ;si besoin en
    - étendant le rectangle ;
    - déplaçant les points.
- N.B. : il est possible de choisir les types de zone que l'on veut afficher ou non ;
- ❺ commencer à transcrire ;
  - ❻ ajouter quelques balises et compléter quelques métadonnées ;
  - ❼ sauvegarder la transcription ;
  - ❽ consulter la source ;
  - ❾ exporter le document en TEI (attention au "tag abuse" de l'export par défaut).



# Plan

Philologie  
numérique :  
constituer un  
corpus

Jean-Baptiste  
Camps

Un outil “tout  
en un” :  
Transkribus

Pas-à-pas avec  
ScanTailor et  
OCRopy

Traitement des  
images

Analyse de la mise en  
page

Reconnaissance des  
écritures manuscrites

## 1 Un outil “tout en un” : Transkribus

## 2 Pas-à-pas avec ScanTailor et OCRopy

- Traitement des images
- Analyse de la mise en page
- Reconnaissance des écritures manuscrites

Dans une démarche de reconnaissance des écritures, la qualité des images et de leurs traitements est cruciale.

Besoins :

- ① images en 300 DPI ;
- ② redressées, débruitées ;
- ③ binarisées.

Outils : logiciels de traitement d'image, par ex. ScanTailor

# T.P. ScanTailor

Philologie  
numérique :  
constituer un  
corpus

Jean-Baptiste  
Camps

Un outil "tout  
en un" :  
Transkribus

Pas-à-pas avec  
ScanTailor et  
OCRopy

Traitement des  
images

Analyse de la mise en  
page

Reconnaissance des  
écritures manuscrites

- ➊ Démarrer un nouveau projet ;
- ➋ charger les images du dossier digby\_23 ;
- ➌ suivre les différentes étapes dans le logiciel ;
- ➍ exporter en tiff binarisé 300 DPI.

# Analyser la mise en page

Philologie  
numérique :  
constituer un  
corpus

Jean-Baptiste  
Camps

Un outil "tout  
en un" :  
Transkribus

Pas-à-pas avec  
ScanTailor et  
OCRopy

Traitement des  
images

Analyse de la mise en  
page

Reconnaissance des  
écritures manuscrites

## Identifier

- zones de texte ;
- décoration ;
- colonnes ;
- lignes ;
- mots ;
- lettres.

## Approches

- Sans apprentissage, par ex.
  - OCRopy 1 ;
  - ORIFLAMMS (IRHT) ;
- fondée sur des méthodes d'apprentissage (IA), par ex.
  - OCRopy 2 ?

# Installer OCRopy

Philologie  
numérique :  
constituer un  
corpus

Jean-Baptiste  
Camps

Un outil "tout  
en un" :  
Transkribus

Pas-à-pas avec  
ScanTailor et  
OCRopy

Traitement des  
images

Analyse de la mise en  
page

Reconnaissance des  
écritures manuscrites

```
$ git clone https://github.com/tmbdev/ocropy.git
$ cd ocropy
$ virtualenv ocropus_venv/
$ source ocropus_venv/bin/activate
$ pip install -r requirements.txt
$ python setup.py install
```

# Analyser la mise en page avec OCRopy

Philologie  
numérique :  
constituer un  
corpus

Jean-Baptiste  
Camps

Un outil "tout  
en un" :  
Transkribus

Pas-à-pas avec  
ScanTailor et  
OCRopy

Traitement des  
images

Analyse de la mise en  
page

Reconnaissance des  
écritures manuscrites

Depuis la racine du dossier :

*#Binarisation*

```
$ ./ocropy/ocropus-nlbin tif/* -o book
```

*#Segmentation en lignes*

```
$ ./ocropy/ocropus-gpageseg -n book/*.bin.png^^I
```

Vérifier la qualité du résultat.

# Différentes solutions techniques

Philologie  
numérique :  
constituer un  
corpus

Jean-Baptiste  
Camps

Un outil "tout  
en un" :  
Transkribus

Pas-à-pas avec  
ScanTailor et  
OCRopy

Traitement des  
images

Analyse de la mise en  
page

Reconnaissance des  
écritures manuscrites

- approches segmentées ou non segmentées ;
- mesures de distance ; méthodes statistiques (chaînes de Markov) ou d'intelligence artificielle (réseaux de neurones convolutifs ou récurrents, LSTM 1D, LSTM 2D, etc.) ;
- outils directement opérationnels ou nécessitant un entraînement.

## Ocropy et CLSTM

Outils développés par Thomas M. Breuel  
(<https://github.com/tmbdev/ocropy> ;  
<http://github.com/tmbdev/clstm>).

- approche non segmentées ;
- réseaux de neurones récurrents (LSTM) ;
- *open source* et nécessitant l'entraînement d'un modèle.

# Ocropy : un apprentissage guidé

Prétraitement des images

Analyse mise en page

transcription

Vérité de terrain

Entraînement

relecture

Test

Sortie

Philologie  
numérique :  
constituer un  
corpus

Jean-Baptiste  
Camps

Un outil "tout  
en un" :  
Transkribus

Pas-à-pas avec  
ScanTailor et  
OCROPy

Traitement des  
images

Analyse de la mise en  
page

Reconnaissance des  
écritures manuscrites



# Ocropy : un apprentissage guidé

Philologie  
numérique :  
constituer un  
corpus

Reconnaissance des écritures manuscrites

## Déclaration de caractères

```

1 # -*- encoding: utf-8 -*-
2
3 import re
4
5 # common character sets
6
7 digits = u"0123456789"
8 letters = u"ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz"
9 symbols = ur'!"#$%&'()*+,-./:;<=>?@[\\]^_`{|}~'"""
10 ascii = digits+letters+symbols
11
12 xsymbols = u'""€£¢¥«»<>÷©®†‡•••¶§±ìíîï"'"
13 german = u"ÄäÖöÜüß"
14 french = u"ÀàÂâÃãÄäÊêËëÊêÎîïÔôÙùÛüÜÛÿ"
15 turkish = u"ĞğŞşİı"
16 greek = u"ΑαΒβΓγΔδΕεΖζΗηΘθΙιΚκΛλΜμΝνΞξΟοΠπΡρΣσςΤτΥυΦφΧχΨψΩω"
17
18 gbank =
19     u"  1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 771 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800 801 802 803 804 805 806 807 808 809 810 811 812 813 814 815 816 817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 9
```

NB : avec CLSTM, cette étape n'est plus nécessaire.

Copier le fichier `chars.py` qui se trouve à la racine du dossier à la place de celui qui se trouve dans `ocrolib`.

# Ocropy : un apprentissage guidé

Philologie  
numérique :  
constituer un  
corpus

Jean-Baptiste  
Camps

Un outil "tout  
en un" :  
Transkribus

Pas-à-pas avec  
ScanTailor et  
OCRopy

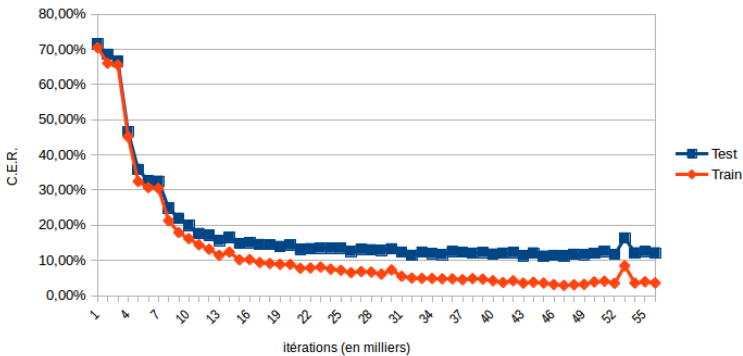
Traitement des  
images

Analyse de la mise en  
page

Reconnaissance des  
écritures manuscrites

## Entraînement sur un ms. (ici Roland d'Oxford)

Modèle Digby n° 3



# Ocropy : un apprentissage guidé

Philologie  
numérique :  
constituer un  
corpus

Jean-Baptiste  
Camps

Un outil "tout  
en un" :  
Transkribus

Pas-à-pas avec  
ScanTailor et  
OCRopy

Traitement des  
images

Analyse de la mise en  
page

Reconnaissance des  
écritures manuscrites

## Test

```
19 Doctrinal-00010000.pyrnn.gz
20 errors          313
21 missing         84
22 total          14404
23 err             2.173 %
24 errnomiss       1.590 %
25 Doctrinal-00012000.pyrnn.gz
26 errors          285
27 missing         84
28 total          14404
29 err             1.979 %
30 errnomiss       1.395 %
```

```
120 errors ~      80
121 missing      0
122 total        5741
123 err          1.393 %
124 errnomiss    1.393 %
125 4
126 4      r      l
127 4
128 3      e      -
129 3      e      c
130 3
131 3      t      f
132 2      t      c
133 2
134 2      c      t
135 0.013934854555
```

# Ocropy : un apprentissage guidé

Philologie  
numérique :  
constituer un  
corpus

Jean-Baptiste  
Camps

Un outil "tout  
en un" :  
Transkribus

Pas-à-pas avec  
ScanTailor et  
OCRopy

Traitement des  
images

Analyse de la mise en  
page

Reconnaissance des  
écritures manuscrites

## Résultats

- pour un imprimé ancien, des taux d'erreur de l'ordre de 1% sont atteignables...
- pour un ms., des taux inférieurs à 10% sont atteignables avec seulement 400 lignes d'entraînement environ (cas du Digby 23, jusqu'à 7% de CER).

# Confusions fréquentes (ms.)

Philologie  
numérique :  
constituer un  
corpus

Jean-Baptiste  
Camps

Un outil "tout  
en un" :  
Transkribus

Pas-à-pas avec  
ScanTailor et  
OCRopy

Traitement des  
images

Analyse de la mise en  
page

Reconnaissance des  
écritures manuscrites

digby4/test3-138000.clstm

errors 193

missing 0

total 14767

err 1.307 %

errnomiss 1.307 %

25 s S

17 \_ .

14 \_

# T.P. OCRopy : Préparation des données

Philologie  
numérique :  
constituer un  
corpus

Jean-Baptiste  
Camps

Un outil "tout  
en un" :  
Transkribus

Pas-à-pas avec  
ScanTailor et  
OCRopy

Traitement des  
images

Analyse de la mise en  
page

Reconnaissance des  
écritures manuscrites

## 1. Essayer d'appliquer un modèle déjà entraîné

*#Lancer la reconnaissance*

```
$ ./ocropy/ocropus-rpred -n -m  
./ocropy/models/digby23-00106000.pyrnn.gz  
book/*/*.bin.png
```

*#Extraire le résultat*

```
$ ./ocropy/ocropus-gtedit html -H35  
book/*/*.bin.png -o gt.html
```

Pas terrible... Mais comment améliorer le modèle ?

## 2. Corriger / transcrire

# T.P. OCRopy : Entraînement

## 3. Extraire et entraîner

### *#Extraire*

```
$ ./ocropy/ocropus-gtedit extract gt.html
```

### *#Normalisation des caractères*

```
$ for f in book/*/*.gt.txt; do uconv -f utf8 -t utf8  
-x nfc -o "${f/gt.txt/gtneu.txt}" "$f"; done  
$for f in book/*/*.gtneu.txt; do mv "$f"  
"${f/gtneu.txt/gt.txt}"; done
```

Puis placer 90% des lignes corrigées dans un dossier train et 10% dans un dossier test.

On peut ensuite lancer un entraînement (de zéro ou à partir du modèle précédent, aux choix).

### *#Lancer l'entraînement*

```
$ ./ocropy/ocropus-rtrain -o digby  
-d 1 train/*/*.bin.png
```

# T.P. OCRopy : Test des résultats

Philologie  
numérique :  
constituer un  
corpus

Jean-Baptiste  
Camps

Un outil "tout  
en un" :  
Transkribus

Pas-à-pas avec  
ScanTailor et  
OCRopy

Traitement des  
images

Analyse de la mise en  
page

Reconnaissance des  
écritures manuscrites

## 4. Tester les résultats

Une fois que l'entraînement a atteint un niveau satisfaisant, on peut tester la qualité du résultat,

```
# Calcul des erreurs de différents modèles, comparati
for i in *.pyrnn.gz; do
echo "$i" >> modeltest
./ocropy/ocropus-rpred -n -m $i test/*/*.bin.png
./ocropy/ocropus-errs test/*/*.gt.txt 2>>modeltest
done
# Confusions de caractères
./ocropy/ocropus-econf test/*/*.gt.txt
```