Institut National des Langues et Civilisations Orientales

Département Textes, Informatique, Multilinguisme

Titre du mémoire

MASTER

TRAITEMENT AUTOMATIQUE DES LANGUES

Parcours :

Ingénierie Multilingue / Traductique et Gestion de l'Information

par

Prénom NOM

Directeur de mémoire :

Zellig Harris

Encadrant :

Xavier Niel

Année universitaire 2015/2016

Contents

List of Figures

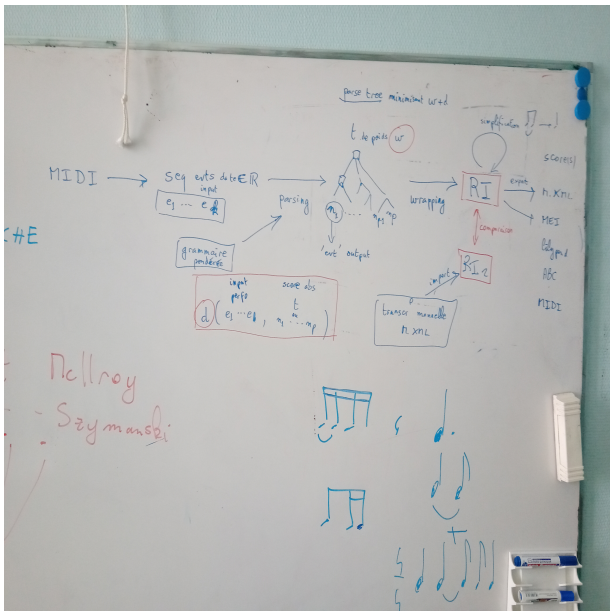
List of Tables

Simple Single Page Abstract template

Arthur Author[1], Cecilia CoAuthor[2]

[1]) First affiliation
arthur.author@correspondence.email.com
[2]) Second affiliation

MÉMOIRE



En entrée : midi (séquence d'événements datés (piano roll) accompagné d'une grammaire pondérée)
⇒ parsing
⇒ global parsing tree
⇒ RI (Représentation Intermédiaire) arbres locaux par intruments
⇒ Sortie (xml, mei, lilypond,. . .)
Minimiser la distance entre le midi et la représentation en arbre.

Le but du stage est d'améliorer qparse, un outil de transcription et d'écriture automatique de la batterie (entre autre)

Le sujet de ce mémoire est de proposer une tâche de reconnaissance du regroupement des notes par les ligatures dans l'écriture de la batterie.

Pour cela, nous utiliserons la logique des systèmes (selon la définition agostinienne).
⇒ Motif répétitif de plusieurs instruments coordonnées accompagnés d'un texte varié joué par un autre instrument de la batterie.

Nous partirons de propositions génériques de systèmes (environs trois systèmes dans

différents style de batterie) que nous tenterons de détecter dans le jeu de données groove.

Nous travaillerons aussi sur la détection de répétitions sur plusieurs mesures afin de pouvoir corriger des erreurs sur une des mesures qui aurait dû être identique au autres mais qui présente des différences.

INTRODUCTION

Présentation générale

étonnant non ?

Ce document donne un modèle possible pour rédiger un mémoire sous LaTeX. Il donne quelques informations sur les commandes les plus utiles. Ce modèle est inspiré de celui utilisé pour les rapports de M2R à l'université Paris Sud, complété par des fonctionnalités pratiques (mini table des matières au début de chaque chapitre, liens cliquables et colorés pour les références, etc.). Il est évident que vous n'êtes pas obligés de rédiger votre mémoire sous LaTeX. Si vous préférez le faire sous un autre éditeur de texte, c'est tout à fait possible.

Ouvrir le fichier "modele.tex" avec un éditeur de texte, et effectuer les modifications nécessaires (cf. lignes qui commencent par TODO). Les commandes pour compiler les fichiers sont dans "lanceur.sh" (voir section A.1).

Plan de lecture

Un mémoire de recherche ou un article scientifique se composent généralement [1] des chapitres suivants :

- Introduction : présentation générale du contexte et de la problématique traitée, plan suivi dans le mémoire ;
- État de l'art (chapitre 1) : les articles qui traitent du même sujet que vous, présentés en un tout cohérent *(extraire de chaque article lu les points essentiels et présenter dans ce chapitre le résultat de ces lectures en regroupant les articles par point essentiel)* ;
- Corpus (chapitre 3) : le corpus utilisé *(caractéristiques, pré-traitements appliqués)* ;
- Méthodes (chapitre 2) : les méthodes appliquées, avec le détail des expériences réalisées (différentes configurations) ;
- Résultats (chapitre 4) : les résultats obtenus sur chacune des expériences ;
- Discussion (chapitre 5) : la discussion des résultats obtenus (quelle expérience a produit les meilleurs résultats, de manière globale, dans le détail des catégories) avec, si possible, une analyse des erreurs pour comprendre les possibilités d'amélioration ;
- Conclusion : la conclusion globale du mémoire.

1. Il est accepté que vous, ou votre directeur de mémoire, estimiez qu'une autre organisation s'impose pour votre problématique de recherche.

En règle générale, l'introduction et la conclusion sont les deux sections de contenu à ne pas être numérotées. Idéalement, chaque chapitre commence par une introduction rapide et se termine par une conclusion rapide pour aider le lecteur à mémoriser et comprendre ce qui a été fait.

Part I

Contexte général

CHAPTER 1

ÉTAT DE L'ART

Sommaire

1.1 Introduction

Dans ce chapitre, nous présentons...

1.2 Contenu

Une section dans ce chapitre avec un appel cliquable de référence bibliographique [Grouin and Névéol, 2014].

1.3 Conclusion

Conclusion de ce chapitre.

CHAPTER 2

MÉTHODES

Sommaire

2.1 Introduction

Dans ce chapitre...

2.2 Contenu

Une section dans ce chapitre...

2.3 Conclusion

Conclusion de ce chapitre.

Part II

Expérimentations

CHAPTER 3

CORPUS

Sommaire

3.1 Introduction

Dans ce chapitre...

3.2 Contenu

Une section dans ce chapitre...

3.3 Conclusion

Conclusion de ce chapitre.

CHAPTER 4

RÉSULTATS

Sommaire

4.1 Introduction

Dans ce chapitre...

4.2 Contenu

Une section dans ce chapitre...

4.3 Conclusion

Conclusion de ce chapitre.

CHAPTER 5

DISCUSSION

Sommaire

5.1 Introduction

Dans ce chapitre...

5.2 Contenu

Une section dans ce chapitre...

5.3 Conclusion

Conclusion de ce chapitre.

Conclusion générale

Dans ce mémoire, nous avons traité de la problématique...

BIBLIOGRAPHIE

[Bossy et al., 2012] Bossy, R., Nédellec, C., and Jourde, J. (2012). *Bacteria Biotope (BB) task at BioNLP Shared Task 2013. Task proposal*. INRA, Jouy-en-Josas, France. – Cité page .

[Bretonnel-Cohen and Hunter, 2008] Bretonnel-Cohen, K. and Hunter, L. (2008). Getting started in text mining. *PLoS Comput Biol*, 4(1):e20. – Cité page .

[Brown et al., 1992] Brown, P. F., Della Pietra, V. J., de Souza, P. V., Lai, J. C., and Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–79. – Cité page .

[Ehrmann, 2008] Ehrmann, M. (2008). *Les entités nommées de la linguistique au TAL : statut théorique et méthodes de désambiguïsation*. PhD thesis, Université Paris VII - Denis Diderot. – Cité page .

[Grouin and Névéol, 2014] Grouin, C. and Névéol, A. (2014). De-identification of clinical notes in French: towards a protocol for reference corpus development. *J Biomed Inform*, 50:151–61. – Cité page .

[Lafferty et al., 2001] Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proc of ICML*, pages 282–9, Williamstown, MA. – Cité page .

[Lavergne et al., 2010] Lavergne, T., Cappé, O., and Yvon, F. (2010). Practical very large scale CRFs. In *Proc of ACL*, pages 504–13, Uppsala, Sweden. – Cité page .

[Sekine and Ranchhod, 2009] Sekine, S. and Ranchhod, E., editors (2009). *Named Entities*. John Benjamins Publishing. – Cité page .

ANNEXE A

DOCUMENTATION

A.1 Compilation

La compilation d'un fichier LaTeX nommé « modele.tex » se fait au moyen des étapes suivantes (en lignes de commande) :

- `pdflatex modele` (première compilation pour l'appel des différentes fonctionnalités contenues dans le mémoire) ;
- `bibtex modele` (préparation de la bibliographie) ;
- `makeindex modele` (préparation de l'index) ;
- `pdflatex modele` (à faire 2 fois, produit le PDF complet).

A.2 Les images

Le schéma A.1 présente le schéma d'annotation en entités nommées du domaine biomédical que nous avons utilisé pour annoter nos corpus de données.

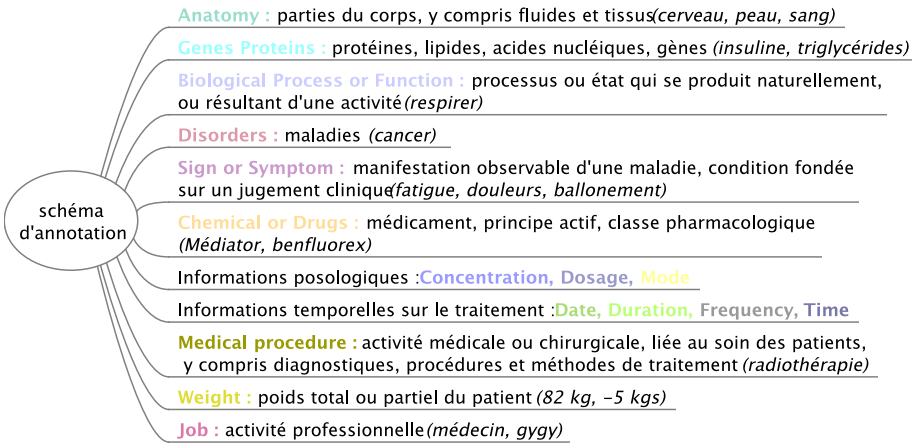


FIGURE A.1 – Schéma d'annotation défini pour les entités nommées biomédicales

L'intégration d'une image (format PDF, PNG) se fait au moyen de la commande `\includegraphics{fichier.pdf}` avec d'éventuelles options entre crochets pour spécifier la taille de l'image (height, width) par rapport à la page.

A.3 Les tableaux

Pour réaliser un tableau en LaTeX, la syntaxe ressemble à (voir tableau A.1) :

Expérience	Rappel	Précision	F-mesure
Baseline	0,372	0,500	0,427
Lexique	0,907	0,903	0,905
Apprentissage	0,880	0,942	0,910

TABLE A.1 – Résultats généraux pour chaque expérience

Dans la commande `\begin{tabular}{|l|c|c|r|}` on définit le nombre de colonnes (ici 4), la manière dont le texte est mis en forme dans chaque colonne (l=left, c=center, r=right), et le séparateur de colonne (ici une ligne verticale). La commande `\hline` permet de tracer une ligne horizontale. La commande `\caption{Titre ou légende}` permet de définir la légende d'un tableau. Et la commande `\label{nom}` permet de nommer le tableau pour le désigner dans le corps du texte avec la commande `\ref{nom}` (e.g., tableau A.1).

Les commandes `\multirow{n}{*}{Texte}` et `\multicolumn{n}{c}{Texte}` permettent de fusionner plusieurs lignes et plusieurs colonnes, avec n le nombre de lignes ou colonnes fusionnées. La commande `\cline{2-4}` permet de dessiner une ligne horizontale de la colonne 2 à la colonne 4 (voir tableau A.2).

Expérience	Mesures		
	Rappel	Précision	F-mesure
Baseline	0,372	0,500	0,427
Lexique	0,907	0,903	0,905
Apprentissage	0,880	0,942	**0,910**

TABLE A.2 – Fusion de lignes et de colonnes dans un tableau

A.4 Mise en forme

Il est possible de mettre du texte en *emphase* `\emph{texte}`, en *version penchée* `\textsl{texte}`, en **gras** `\textbf{texte}`, en version Sans Serif (similaire à Arial et Helvetica) `\textsf{texte}` ou encore en PETITES CAPITALES `\textsc{Texte}`. Il n'est pas recommandé d'utiliser le souligné au vu de l'effet produit. Pour ajouter une note de bas de page[1], on utilise la commande `\footnote{Contenu}`. L'espace insécable (c.-à-d. qui ne peut pas être coupée si elle est en fin de ligne), est représentée par le tilde. On l'utilise généralement avant un appel de référence, ou avant la référence d'un tableau ou d'une figure `Tableau~\ref{clé}`.

Il est possible de ne pas numéroter les titres de section, sous-section, etc., en utilisant la commande `\section{Titre}`. Dans ce cas, il est nécessaire d'ajuster les mini tables des matières qui figurent au début de chaque chapitre au moyen de la commande `\adjustmtc` en intégrant cette commande autant de fois que nécessaire avant la commande `\minitoc`

1. Les notes de bas de page sont numérotées automatiquement en fonction de leur utilisation tout au long du texte. Avec ce modèle, la numérotation est remise à 1 à chaque début de chapitre.

Pour forcer un saut de page, on utilise la commande `\newpage` et pour un saut de ligne `\\`

A.5 Formules mathématiques

Le rappel (formule A.1) mesure le nombre d'éléments correctement étiquetés par le système (vrais positifs) rapporté au nombre d'éléments étiquetés dans la référence (vrais positifs et faux négatifs).

$$\text{Rappel} = \frac{\text{vrais positifs}}{\text{vrais positifs + faux négatifs}} \tag{A.1}$$

La précision (formule A.2) mesure le nombre d'éléments correctement étiquetés par le système (vrais positifs) rapporté au nombre total d'éléments étiquetés par le système (vrais et faux positifs).

$$\text{Précision} = \frac{\text{vrais positifs}}{\text{vrais positifs + faux positifs}} \tag{A.2}$$

La F-mesure (formule A.3) est la moyenne harmonique pondérée du rappel et de la précision. La valeur accordée à β permet de pondérer le rappel ou la précision, ou d'équilibrer les deux mesures (avec $\beta = 1$).

$$\text{F-mesure} = \frac{(1+\beta^2) \times \text{ précision } \times \text{ rappel}}{\beta^2 \times \text{ précision } + \text{ rappel}} \tag{A.3}$$

Les plus motivés d'entre vous pourront également réaliser des figures (arbre syntaxique, histogramme, diagramme circulaire, etc.) directement sous LaTeX en utilisant le package Tikz [2] une fois qu'ils auront terminé la rédaction de leur mémoire pour ne pas perdre de temps dès le départ...

A.6 Gestion de la bibliographie

La bibliographie figure dans un fichier nommé « biblio.bib ». Ce fichier peut être édité dans un éditeur de texte classique *(emacs, vi, notepad)*, ou par le biais d'un outil de gestion de bibliographie tel que JabRef, un programme Java qui permet de gérer efficacement les fichiers *.bib et de remplir les différents champs nécessaires à chaque entrée.

A.6.1 Appels de référence

On appelle les références au moyen de la commande `\cite{clé}` avec une clé de citation qui est définie dans la bibliographie. On intègre généralement une référence bibliographique après avoir introduit le concept. Par exemple : nos expériences en approche statistique reposent sur le formalisme des CRF [Lafferty et al., 2001] implémenté dans l'outil WAPITI [Lavergne et al., 2010]. Nous avons suivi le protocole défini par [Grouin and Névéol, 2014, p. 3] pour constituer les annotations de corpus.

Il est possible de citer plusieurs articles en même temps en séparant chaque clé par une virgule. Par exemple : notre travail repose sur la détection d'entités nommées [Ehrmann, 2008, Sekine and Ranchhod, 2009] du domaine biomédical, en particulier la détection des mentions de bactéries et de biotopes [Bossy et al., 2012].

2. Voir http://bertrandmasson.free.fr/index.php?categorie6/latex-pgf-tikz pour de nombreux exemples utiles.

A.6.2 Format

LaTeX gère automatiquement le format de présentation des références, selon que l'article cité a été rédigé par un auteur [Ehrmann, 2008] (auquel cas sont mentionnés le nom et l'année), deux auteurs [Bretonnel-Cohen and Hunter, 2008] (les noms des deux auteurs et l'année), ou plus de deux auteurs [Brown et al., 1992] (le nom du premier auteur, la mention *et al.* [3], et l'année).

3. Locution latine signifiant « et d'autres ».

ANNEXE B

PRINCIPES À SUIVRE

B.1 Le sujet de votre mémoire

Vous avez acquis, au cours de l'année 2015-2016, des compétences d'ingénieur-linguiste ; vous savez donc analyser un problème, proposer une méthodologie permettant d'arriver à une solution et montrer les limites de cette dernière. C'est cette démarche qui constituera le fil directeur de votre mémoire.

Ce travail devra être original et personnel. Le cadre de votre travail est naturellement la linguistique et, étant donné le diplôme que vous préparez, la linguistique appliquée, plutôt que théorique. Ceci ne veut néanmoins pas dire que vous ne devrez pas situer votre démarche à l'intérieur d'un cadre théorique, au contraire. On souhaite cependant que ce cadre serve d'appui à la création ou à la transformation d'outils, à la mise au point de méthodologies vous permettant de proposer un résultat.

Cela revient à dire que votre mémoire constitue une tentative de problématiser une approche méthodologique, de proposer une piste nouvelle, de comparer des méthodes, des outils, etc. Il contiendra en tout cas un état de l'art et s'appuiera sur une bibliographie précise et récente. L'état de l'art ne doit pas être déconnecté de la question traitée : on ne vous demande pas de « faire un état de l'art pour faire un état de l'art » mais, au contraire, de montrer comment se situe votre travail par rapport à cet état de l'art. Si votre sujet s'y prête, et afin d'en faciliter la réalisation, vous pouvez segmenter votre état de l'art en plusieurs parties ciblées à placer en tête des chapitres correspondant plutôt que d'écrire un chapitre consacré qui risque d'être généraliste et donc insuffisamment précis.

Vous devrez avoir choisi un sujet de mémoire à la mi-mai ou, à tout le moins, avoir réfléchi à des pistes sérieuses. Vous devrez vous assurer auprès d'un intervenant du TIM/ER-TIM que vous ne faites pas fausse route et que votre mémoire ne sera pas hors-sujet. Il s'agit d'éviter que vous ne traitiez un sujet dont les exigences techniques pourraient s'avérer supérieures à ce que vous croyez connaître. Le(s) stage(s) de fin d'études que vous devez entreprendre peu(t/vent) vous aider à affiner votre choix de sujet, mais vous devez garder à l'esprit que votre mémoire ne doit pas se confondre avec une description de votre stage. Notez bien que les rapports de stage ne sont pas pris en compte dans l'évaluation de votre Master.

Pour vous aider, vous pouvez consulter les meilleurs mémoires des années précédentes (et dont les résumés sont en ligne sur le site `www.er-tim.fr`). Évidemment, vous consulterez également les articles scientifiques liés à votre problématique : outre les connaissances que vous pourrez ainsi acquérir, cela vous permettra aussi de vous familiariser avec ce genre bien spécifique. Si vous ne trouviez pas de sujet vous permettant de mettre en pratique les connaissances acquises au cours de cette année, en

fonction de vos goûts et attentes personnels ou professionnels, nous vous en proposerions un (consultez-nous, donc).

B.2 L'encadrement du mémoire

Vous avez toute latitude pour choisir, selon affinités, la/les personne(s) qui va/vont diriger vos recherches. Mais un/des intervenant(s) du TIM/ER-TIM figurera/ont nécessairement dans votre jury lors de la soutenance. Il faut donc nécessairement avoir pris contact avec ces personnes et s'assurer de leur collaboration. Si vous envisagez de faire une thèse ensuite, il est recommandé de solliciter un enseignant assimilé professeur ou habilité à diriger des recherches ou de mettre en place un co-encadrement en ce sens.

En règle générale, le TIM/ER-TIM souhaite, autant que faire se peut, que les personnes qui vous ont encadré lors de votre stage et qui ont pu vous conseiller pour la rédaction de votre mémoire, soient présentes lors de la soutenance. Elles apportent un complément d'information interne sur le stage et les conditions de réalisation du mémoire, éclairage qui peut être tout à fait pertinent.

Si vous rencontrez des problèmes et souhaitez poser des questions, il est impératif, dans un premier temps, de les formuler par courrier électronique plutôt que de venir immédiatement au TIM/ER-TIM, riche en compétences mais pauvre en personnel. Par ailleurs, vous ne devez pas envoyer par courrier électronique des centaines de pages à fin de re-lecture : lorsqu'une pré-version de votre travail vous semblera digne de relecture, déposez-la au TIM/ER-TIM, ou postez-la.

B.3 L'évaluation du mémoire

L'évaluation du mémoire est fonction de la qualité de votre travail écrit et de votre capacité à répondre aux questions, remarques, critiques qui peuvent vous être adressées pendant la soutenance. La qualité du travail écrit dépend de plusieurs critères, dont voici une liste non-exhaustive :

- votre mémoire forme-t-il un ensemble cohérent qui doit son unité à la volonté de répondre à une problématique bien définie ?
- votre mémoire est-il réutilisable par une personne souhaitant faire un bilan de la problématique soulevée, tant du point de vue fond que forme (clarté de la bibliographie, description en annexe des outils utilisés avec liens aux sources, disponibilités des sources sur le CD-ROM d'accompagnement de votre mémoire, index permettant une consultation rapide, table des matières, pagination, etc.) ?
- votre mémoire répond-t-il vraiment à l'objectif fixé au départ ? le titre de votre mémoire correspond-il vraiment au contenu ? les mots-clés qui seront mis en ligne sont-ils pertinents ?
- votre mémoire met-il en valeur un angle de vue original sur un savoir-faire classique ?
- votre mémoire parvient-il à mettre la théorie à l'épreuve ? Êtes-vous capable de fournir des résultats, des exemples, un bilan d'expérience, des critères d'évaluation, une évaluation ?

- la bibliographie doit être totalement normalisée, de façon à permettre une consultation aisée, les annexes contiendront un descriptif pratique et les références des outils utilisés, un échantillon des corpus utilisés et des programmes que vous avez écrits et, de manière générale, tout ce qui peut illustrer le travail réalisé. Attention, pour des raisons de place, vous ne devez évidemment pas présenter tous vos corpus et tous vos programmes en annexe, mais un simple échantillon. En revanche, corpus [1] et programmes figureront impérativement et exhaustivement sur le CD fourni.

La qualité de votre prestation orale est importante. Vous devrez vous assurer, en particulier, que :

- vous savez vous affranchir du plan de votre mémoire mais vous devez néanmoins faire un bref résumé de la problématique car tous les membres du jury n'auront pas lu votre mémoire
- vous donnez des exemples concrets des questions qui se sont posées et des solutions apportées, de façon à montrer que vous ne traitez pas le sujet de façon purement théorique
- vous savez situer la problématique de votre mémoire par rapport aux travaux les plus connus et les plus récents sur la question
- vous savez faire le lien entre les connaissances acquises au cours de l'année et la mise en pratique de ces connaissances lors de la réalisation du mémoire
- vous savez répondre aux questions ou critiques qui vous sont soumises

B.4 La démarche à suivre pour soutenir

Trois semaines avant la date de soutenance, vous devez envoyer une version présentable de votre mémoire à votre encadrant et à l'équipe de formation, pour déterminer si le mémoire est soutenable. Vous devez remettre une version papier définitive de vos mémoires au moins 15 jours avant la soutenance.

- La soutenance pour la première session est fixée entre le 20 et le 24 juin 2016 (à préciser) pour ceux d'entre vous qui candidateraient à un contrat doctoral INaLCO (voir la procédure sur le site www.inalco.fr, le comité de sélection ayant lieu le 1 juillet 2016).
- Pour la deuxième session (inscription en doctorat à l'INaLCO selon la procédure normale), la soutenance est fixée le 30 septembre 2016.
- Pour la dernière session, la date de soutenance est fixée le 18 novembre 2016.

Vous devez déposer votre travail au moins deux semaines avant d'espérer soutenir. Il faut en effet qu'il soit lu, puis, si nécessaire, amendé et corrigé – voire rejeté et réécrit – de façon que la soutenance ne verse pas dans la critique systématique.

Au plus tard la veille de votre soutenance, vous aurez envoyé à crim@inalco.fr et à sophie.urbaniak@inalco.fr un résumé de votre mémoire de 10 lignes

1. Vérifiez toutefois que vous avez le droit de reproduire tout ou partie du corpus sur lequel vous aurez travaillé, en particulier pour les corpus de documents cliniques.

maximum ainsi que 5 mots-clés permettant de situer votre travail. Attention, ces informations sont destinées à être consultées et doivent donc être le reflet fidèle de votre travail final.

Une fois votre travail accepté, nous vous proposerons un ordre de passage pour la soutenance. Vous devrez fournir 3 exemplaires/support-papier et 3 exemplaires/support-électronique de votre mémoire (ces exemplaires sont destinés aux membres du jury et aux futurs étudiants). Sur le 4ème de couverture vous agraferez une enveloppe format 21-27 qui contiendra le CD correspondant à votre travail. Ce CD contiendra, outre la version électronique de votre mémoire, toutes les annexes ne pouvant figurer dans le mémoire pour des raisons de place : corpus, code source des outils utilisés, polices de caractères utilisées, code des programmes que vous avez élaborés.