

Université de Montréal

**Étude comparative du vocabulaire de description
de la danse dans les archives et du vocabulaire de
représentation de la danse dans la littérature**

par

Ève Paquette-Bigras

École de bibliothéconomie et des sciences de l'information

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures et postdoctorales
en vue de l'obtention du grade de maître
en sciences de l'information

Mai, 2013

© Ève Paquette-Bigras, 2013

Université de Montréal
Faculté des études supérieures et postdoctorales

Ce mémoire intitulé :

Étude comparative du vocabulaire de description de la danse dans les archives
et du vocabulaire de représentation de la danse dans la littérature

présenté par :

Ève Paquette-Bigras

a été évalué par un jury composé des personnes suivantes :

Michèle Hudon, président-rapporteur

Dominic Forest, directeur de recherche

James Turner, co-directeur

Yvon Lemay, membre du jury

Résumé

Notre recherche s'insère dans la mouvance des humanités numériques; nous y faisons dialoguer les arts et les sciences de l'information. Depuis quelques décennies, la danse est un sujet d'études et de recherche à part entière. Il devient donc nécessaire de mieux décrire la danse dans les archives, sachant que la description en amont influe grandement sur l'accès en aval. Les méthodes d'extraction automatique de connaissances nous semblent offrir de nouvelles possibilités.

L'objectif de notre recherche est de contribuer au développement d'outils de gestion de l'information dans les archives de la danse en comparant un vocabulaire de description de la danse dans les archives et un vocabulaire de représentation de la danse dans la littérature, recueilli grâce à des méthodes d'extraction automatique de connaissances, pour en distinguer une possible complémentarité, particulièrement en ce qui a trait au vocabulaire de l'expérience esthétique.

D'abord, nous analysons un vocabulaire de description de la danse dans les archives. Nous décrivons certains outils de description des archives de la danse et nous analysons le thésaurus de descripteurs Collier. Nous constatons que le vocabulaire de description de la danse dans les archives ne semble pas prendre en compte l'expérience esthétique. Ensuite, nous analysons un vocabulaire de représentation de la danse dans la littérature. Un vocabulaire structuré de l'expérience esthétique de la danse moderne est ainsi extrait d'un corpus de textes de l'écrivain français Stéphane Mallarmé et analysé. Puis nous comparons les deux vocabulaires afin d'en distinguer la complémentarité quant à la description de l'expérience esthétique.

Nous formulons une première suggestion d'amélioration de certains thésaurus employés dans les archives de la danse : un thésaurus au vocabulaire

essentiellement factuel, comme le thésaurus de descripteurs Collier, peut être enrichi de termes à propos de l'expérience esthétique. Le vocabulaire de représentation de la danse dans la littérature est jusqu'à un certain point complémentaire au vocabulaire de description de l'expérience esthétique de la danse dans les archives.

Nous menons ainsi une première expérimentation qui justifie en partie la pertinence de certaines méthodes d'extraction de connaissances dans le développement et la maintenance de ressources documentaires pour le domaine des arts d'interprétation tels que la danse.

Mots-clés : Archives; Description des documents; Arts; Danse; Humanités numériques; Extraction automatique de connaissances; Fouille de textes; Stéphane Mallarmé.

Abstract

This research falls within the field of digital humanities; arts and information science engage in dialogue. In the last few decades, dance has become a distinct research subject. Dance description in archives needs to be improved, because the quality of the description impacts access to the documentation. Knowledge extraction seems to offer new opportunities in this regard.

The goal of this research is to contribute to the development of information management tools by comparing a vocabulary for describing dance in archives with a vocabulary for representing dance in literature obtained through knowledge extraction. We look for possible complementarity, particularly in regard to the aesthetic experience.

First, some tools for describing dance in archives are described, and the *Collier Descriptor Thesaurus* is analyzed. We observe that this vocabulary for describing dance in archives does not take into account aesthetic experience. Second, a vocabulary for representing dance in literature is analyzed. More specifically, a structured vocabulary of aesthetic experience of modern dance is drawn from a corpus of texts from the French writer Stéphane Mallarmé, and the vocabulary obtained is analyzed. Finally, the two vocabularies are compared to consider their complementarity.

We conclude that some vocabularies for describing dance in archives, consisting mainly of factual terms, such as the *Collier Descriptor Thesaurus*, can be enriched with terms related to aesthetic experience. The vocabulary for representing dance in literature complements to a certain extent the vocabulary for describing dance in archives. Thus this initial experiment supports the relevance of knowledge extraction in information resources maintenance and development for performing arts such as dance.

Keywords : Archives; Document description; Arts; Dance; Digital humanities; Knowledge extraction; Text mining; Stéphane Mallarmé.

Abstrakt

Diese Arbeit beschäftigt sich mit dem Fachgebiet der Digital Humanities und verbindet dabei Kunst mit informationswissenschaftlichen Methoden. In den letzten Jahrzehnten ist Tanz ein eigenständiges Forschungsgebiet geworden. Da sich die Qualität der Beschreibung direkt auf den Zugang zu Dokumenten im Archiv auswirkt, bedarf die Beschreibung von Tanz in Archiven Verbesserung.

Ziel der Forschung ist es zur Entwicklung von Informationsverwaltungs-Tools beizutragen, indem das Vokabular der Beschreibung von Tanz im Archiv mit Vokabular aus der Literatur, extrahiert aus textuellen Datenbanken, verglichen wird. Dabei liegt der Fokus auf der Komplementarität beider Quellen, besonders in Bezug auf die Beschreibung von ästhetischen Erfahrungen.

Zunächst werden Tools für die Beschreibung von Tanz in Archiven beschrieben und der Collier Descriptor Thesaurus analysiert. Dabei zeigt sich, dass das Vokabular der Tanz-Beschreibung im Archiv ästhetische Erfahrung generell nicht berücksichtigt. Daraufhin wird das Vokabular der Tanz-Darstellung in der Literatur am Beispiel der Text-Sammlung des französischen Dichters Stéphane Mallarmé analysiert. Im Anschluss werden die zwei Wortschätze verglichen, um die Komplementarität beider Quellen zu beschreiben.

Die Arbeit kommt zu dem Schluss, dass das Vokabular der Tanz-Beschreibung im Archiv hauptsächlich aus sachbezogenen Begriffen besteht (z.B. der Collier Descriptor Thesaurus), welche um Begriffe zur ästhetischen Erfahrung ergänzt werden können. Die Begriffe für die Tanz-Beschreibung in der Literatur komplementieren bis zu einem gewissen Grad das Vokabular der Tanz-Beschreibung im Archiv. Demzufolge bildet diese Arbeit eine Grundlage

für weitere Forschung im Bereich der Wissensextraktion in textuellen Datenbanken im Fachgebiet darstellender Künste wie Tanz.

Stichwörter : Archiv; Dokumentbeschreibung; Kunst; Tanz; Digital Humanities; Wissensextraktion in textuellen Datenbanken; Stéphane Mallarmé.

Table des matières

Liste des tableaux	xii
Liste des figures	xiii
Liste des sigles et des acronymes	xiv
Remerciements	xvi
Introduction.....	1
1. La danse comme sujet.....	1
2. Objectifs.....	4
2.1. Objectif général	4
2.2. Objectifs spécifiques.....	5
3. Questions de recherche	5
4. Plan du mémoire	6
5. Domaines de recherche.....	7
Chapitre 1. Revue de la littérature	9
1.1. Introduction.....	9
1.2. Description de la danse.....	9
1.2.1. Qu'est-ce qu'un langage documentaire?	12
1.2.2. Qu'est-ce qu'un thésaurus?.....	14
1.2.3. Qu'est-ce qu'une liste de vedettes-matière?	15
1.2.4. Thésaurus de descripteurs Collier.....	17
1.2.5. Vedettes-matière du <i>Dictionary catalog of the Dance Collection</i> ...	20
1.3. L'expérience esthétique comme point d'accès	23
1.4. Extraction de termes et classification	25
1.4.1. Qu'est-ce que l'extraction de termes?	25
1.4.2. Qu'est-ce que la classification?	26
1.5. Conclusion	31
Chapitre 2. Méthodologie	33
2.1. Introduction.....	33

2.2. Vocabulaire de description de la danse dans les archives : le thésaurus de descripteurs Collier.....	33
2.3. Vocabulaire de représentation de la danse dans la littérature : quelques textes choisis de Stéphane Mallarmé.....	34
2.3.1. Constitution d'un corpus de textes.....	37
2.3.2. Collecte des données.....	40
2.4. Conclusion.....	45
Chapitre 3. Résultats et discussion.....	47
3.1. Introduction.....	47
3.2. Objectifs de recherche.....	48
3.2.1. Premier objectif de recherche : analyse d'un vocabulaire de description de la danse dans les archives.....	48
3.2.2. Deuxième objectif de recherche : analyse d'un vocabulaire de représentation de la danse dans la littérature.....	52
3.2.3. Troisième objectif de recherche : comparaison entre le vocabulaire de description et le vocabulaire de représentation.....	68
3.3. Limites de la recherche.....	69
3.3.1. Limites relatives aux archives de la danse.....	69
3.3.2. Limites relatives à l'extraction automatique de connaissances.....	70
3.4. Questions de recherche.....	71
3.4.1. Première question de recherche.....	71
3.4.2. Seconde question de recherche.....	72
3.5. Conclusion.....	73
Conclusion.....	75
1. Résumé de la recherche.....	75
2. Contribution au domaine des sciences de l'information.....	79
2.1. Contribution théorique.....	79
2.2. Contribution méthodologique.....	79
2.3. Contribution pratique.....	79

3. Recherches futures.....	80
Bibliographie	81
Annexe 1. Extrait du thésaurus Collier : le terme <i>ballet</i>	xvii
Annexe 2. Catégorie <i>interprétation</i>	xviii
Annexe 3. Catégorie <i>autres arts</i>	xix
Annexe 4. Catégorie <i>études de la danse</i>	xxi
Annexe 5. Catégorie <i>administration/gestion</i>	xxii
Annexe 6. Catégorie <i>formation</i>	xxiii
Annexe 7. Catégorie <i>matériel</i>	xxiv
Annexe 8. Catégorie <i>diffusion</i>	xxv
Annexe 9. Catégorie <i>personne(s)</i>	xxvi
Annexe 10. Catégorie <i>santé</i>	xxvii
Annexe 11. Catégorie <i>sport</i>	xxviii
Annexe 12. Extrait du dendrogramme de classification du corpus	xxix

Liste des tableaux

Tableau 1. Caractéristiques des listes de vedettes-matière et des thésaurus.....	16
Tableau 2. Caractéristiques de différents algorithmes de classification.....	28
Tableau 3. Vingt termes extraits d'une classe de taille restreinte.....	53
Tableau 4. Dix-sept termes au poids informatif le plus élevé	54
Tableau 5. Neuf thèmes de l'expérience esthétique	56
Tableau 6. Quatre termes au poids informatif le plus élevé de la classe de 73 documents.....	60
Tableau 7. Vingt-quatre termes au poids informatif le plus élevé de la sous-classe de quarante documents.....	62
Tableau 8. Sept termes au poids informatif le plus élevé de la sous-classe de onze documents	63
Tableau 9. Neuf termes au poids informatif le plus élevé de la sous-classe de cinq documents	63
Tableau 10. Quinze termes au poids informatif le plus élevé de la sous-classe de trois documents	64

Liste des figures

Figure 1. Extrait du thésaurus Collier : le terme <i>ballet</i>	19
Figure 2. Extrait du <i>Dictionary catalog of the Dance Collection</i> : la chorégraphie <i>Après-midi d'un faune</i> de Serge Lifar.....	22
Figure 3. Classification hiérarchique agglomérative	30
Figure 4. Classification hiérarchique divisive	30
Figure 5. Schéma de la méthode de collecte du vocabulaire de représentation de la danse dans la littérature.....	36
Figure 6. Équation de la fonction <i>tf.idf</i>	43
Figure 7. Pseudo-algorithme de classification hiérarchique agglomérative	44
Figure 8. Extrait du thésaurus Collier : le descripteur <i>administration</i>	50
Figure 9. Extrait du thésaurus Collier : le descripteur <i>gestion</i>	50
Figure 10. Nuage des quatre termes au poids informatif le plus élevé de la classe de 73 documents.....	60
Figure 11. Vocabulaire structuré de l'expérience esthétique de la danse moderne extrait d'un corpus de textes de Stéphane Mallarmé.....	66

Liste des sigles et des acronymes

BIDC	Base de données intégrée de danse canadienne
BnF	Bibliothèque nationale de France
BtB	Bureau de la traduction
CIDOC	Comité international pour la documentation
CRM	Conceptual Reference Model
DCD	Dance Collection Danse
DHC	Dance Heritage Coalition
FRBR	Functional Requirements for Bibliographic Records
ISO	International Organization for Standardization
LC	Library of Congress
LCSH	Library of Congress Subject Headings
NYPL	New York Public Library
OQLF	Office québécois de la langue française
RDDA	Règles pour la description des documents d'archives
RAMEAU	Répertoire d'autorité-matière encyclopédique et alphabétique unifié

*Tanzen-können mit den Füßen,
mit den Begriffen, mit den Worten*

*Savoir danser avec les pieds,
avec les idées, avec les mots*

FRIEDRICH NIETZSCHE (1889)

Remerciements

Puisque tout mémoire de maîtrise est une oeuvre collective, je tiens à remercier sincèrement...

le directeur de ce mémoire, Dominic Forest;

le codirecteur, James Turner;

mes collègues : Marcela Baiocchi, Guylaine Blais, Dany Bouchard, Stéphanie Cadieux, Aline Crédeville, Laure Guitard, Anne Klein et Philippe Mongeon;

Martin Bélanger, responsable de formation professionnelle;

Alain Tremblay, technicien à la gestion des dossiers étudiants;

Marie-Josée Lecours et Chloé Bélanger St-Germain, respectivement bibliothécaire et technicienne en documentation à la bibliothèque Vincent-Warren de l'École supérieure de ballet du Québec;

Stefanie Haustein, postdoctorante à l'Université de Montréal, et David Weigl, doctorant à l'Université McGill;

Audrey Laplante, professeure à l'École de bibliothéconomie et des sciences de l'information;

les bibliothécaires de la Bibliothèque des lettres et sciences humaines;

les enseignants et les professeurs qui m'ont encouragée au cours de mes études;

mes amis;

ma mère, mon père, ma soeur et mon frère.

Introduction

1. La danse comme sujet

Si, déjà à la fin du 19^e siècle, le philosophe allemand Friedrich Nietzsche écrivait dans *Ainsi parlait Zarathoustra* « que chaque jour où l'on n'a pas dansé une fois au moins soit perdu pour nous¹ » et l'auteur français Stéphane Mallarmé louait dans *Autre étude de danse* les éblouissantes performances de Loïe Fuller, une des pionnières de la danse moderne, depuis un peu plus de trois décennies, la danse, comme pratique artistique, comme sujet d'études et de recherche, comme expression culturelle, est valorisée comme rarement auparavant (Valaskakis-Tembeck, Odom et Fisher-Stitt 1997; Desalme 2007 : 13). La documentation de la danse connaît parallèlement une croissance considérable (Couch 1994 : 42; Desalme 2007 : 13; Chaffee 2011 : 125) et la discipline a maintenant besoin plus que jamais de la contribution des sciences de l'information.

Une réticence quant à la documentation de la danse existe encore toutefois de la part de certains artistes et théoriciens de l'art. Selon Peggy Phelan (1993 : 146), la performance n'existe qu'au présent, car la « [p]erformance [...] becomes itself through disappearance ». Ainsi, tenter de sauvegarder, de documenter la performance en trahirait l'essence. Ce discours n'est pas sans trouver un certain écho dans le milieu de la danse. La critique de danse américaine Marcia B. Siegel (1972 : 1) écrit que la « [d]ance exists at a perpetual vanishing point. At the moment of its creation it is gone. » Philippe Le Moal (1998 : 39) parle de rhétorique de l'éphémère. Michelle Potter (1992 :

¹ « Und verloren sei uns der Tag, wo nicht Ein Mal getanzt wurde! » traduit de l'allemand par Henri Albert dans les *Œuvres complètes de Frédéric Nietzsche*, Paris, Mercure de France, 1903.

115) souligne quant à elle « the widely held idea that dance, that most ephemeral of arts, does not express itself in words, and hence has no written documentation of value ». La danse possède plutôt une tradition de transmission s'appuyant sur la mémoire du corps (Hansen Kopp 1995). Le maître forme ses élèves; il transmet la danse en façonnant le corps de ses disciples à grand renfort d'exercices inlassablement répétés.

Certaines incompréhensions quant à la performance artistique existent aussi du côté des sciences de l'information. Dans cette perspective, David Miller et Patrick Le Boeuf (2005 : 154) vont même jusqu'à affirmer que « [f]or your work to be recognized by catalogers as a topic, you have to produce a text, a score, a movie, a sound recording, a picture, or a sculpture... but not a live performance ». La performance, sa description, sa classification, son indexation, n'est pas traitée de manière satisfaisante; la spécificité de la danse est occultée. Les deux auteurs soutiennent que le catalogage faillit, rien de moins, à prendre en compte et à conceptualiser la performance (Miller et Le Boeuf 2005 : 153). Dans les archives, les pratiques de documentation de la performance sont relativement éclatées et peu explicitées. L'expérience et la mémoire de l'archiviste font souvent foi de tout (Evens et Hautekeete 2011: 162).

Pourtant nombreux sont les artistes contemporains qui puisent dans les archives un matériau qu'ils intègrent à leurs pratiques artistiques (Poinsot 2004; Lemay 2009; Bégoc, Boulouch et Zabunyan 2010). En Grande-Bretagne, le projet de recherche *Performing the Archive: the future of the past*, une collaboration de l'Université de Bristol, de l'Université d'Exeter et de la Galerie Arnolfini, explore entre autres l'utilisation des documents d'archives et leur influence sur les pratiques artistiques et le discours des artistes (Clarke et Warren 2009). Dans le domaine de la danse, les exemples d'utilisation d'archives sont nombreux. Pensons à *zukunft_erinnern*, une chorégraphie des

artistes allemands Dieter Baumann et Jutta Hell, où un hommage est rendu à Gerhard Bohner, leur mentor, mort prématurément. Dieter Baumann y danse un solo en duo avec des extraits vidéos – présence fantomatique et numérique – de Gerhard Bohner (Scholz 2005).

Les artistes s'inspirent des archives; les archivistes peuvent s'inspirer de l'art. Comme certains auteurs le soulignent, « many performers use archives as sources of inspiration for new works, performance records should arguably be more creative and experimental in nature » (Jones, Abbott et Ross 2009 : 166). Un extrait du rapport *Securing our dance heritage* du Council on Library and Information Resources se lit comme suit « [p]reserving the remnants of an art form and cultural activity dependent on time and space requires no less creativity than that needed to create the dance itself » (Johnson et Fuller Snyder 1999). Ainsi, les archives de la danse doivent saisir l'essence de la danse, « capture the art form's essence and spirit » (Chaffee 2011 : 125). Gabriele Brandstetter argue même la nécessité de surmonter le paradoxe apparent entre performance et préservation pour alors « binding meaningfully together the 'art of production' and the practice of preservation » (Brandstetter 2006 : 17). L'art qu'est la danse et la documentation de la danse ne sont pas antinomiques, mais plutôt hautement conciliables, voire intrinsèquement liés.

Dans la *Review of recent approaches to growing dance legacy* présenté au Conseil des arts du Canada (Rowat 2005 : 5), on plaide d'ailleurs ceci :

Your future is now.

Dance is an ephemeral art form presented in the moment. Looking ahead to how your work will survive is an act of self-respect and empowerment.

Everyone who has produced an artistic legacy whether dancers (interpretive artists) or

choreographers (creative artists) – deserves more than an ephemeral existence.

Allow your work to survive.

La danse est un art particulier, un art d'interprétation, un *performing art*². Le moment de la performance est singulier et ne dure que quelques instants. Se produit alors un amalgame unique de contingences humaines, matérielles et temporelles. De là jaillit l'art; de là jaillit l'expérience esthétique. Les représentations de *Torn Roots, Broken Branches*, une chorégraphie de Margie Gillis, ne sont plus depuis longtemps déjà, pourtant reste un sillage. Si la performance reste insaisissable, les traces qu'elle laisse peuvent être documentées, méritent d'être documentées, doivent être documentées.

2. Objectifs

2.1. Objectif général

Puisque la danse est maintenant un sujet d'études et de recherche à part entière, puisqu'il est nécessaire de mieux décrire la performance en danse dans les archives, puisque le monde de l'art, de la danse, et le monde des sciences de l'information, de la documentation de la danse, ont tout intérêt à s'inspirer un de l'autre, l'objectif général de notre recherche est de contribuer au développement d'outils de gestion de l'information en comparant un

² Nous avons choisi de privilégier l'expression *arts d'interprétation* pour traduire *performing arts* plutôt que les synonymes *arts du spectacle* ou *arts de la scène*, car elle nous semblait plus juste dans le cas de la danse (OQLF 2012; BtB 2012). Notez que le terme *art performance*, qui existe bel et bien en français, n'est pas la traduction de *performing art*. Il désigne plutôt un mode d'expression artistique particulier : la performance.

vocabulaire de description³ de la danse dans les archives et un vocabulaire de représentation de la danse dans la littérature pour en distinguer une possible complémentarité, tout particulièrement en ce qui a trait au vocabulaire de l'expérience esthétique.

2.2. Objectifs spécifiques

De l'objectif général découlent plusieurs objectifs spécifiques :

1. Analyser un vocabulaire de description de la danse dans les archives.
2. Analyser un vocabulaire de représentation de la danse dans la littérature recueilli grâce à des méthodes d'extraction automatique de connaissances.
3. Comparer les deux vocabulaires afin d'en distinguer la complémentarité quant à la description de l'expérience esthétique.

3. Questions de recherche

Deux questions sous-tendent notre recherche :

1. Est-ce que le vocabulaire de description de la danse dans les archives prend en compte l'expérience esthétique?
2. Est-ce que le vocabulaire de représentation de la danse dans la littérature, recueilli grâce à des méthodes d'extraction automatique de connaissances, peut enrichir le vocabulaire de description de l'expérience esthétique de la danse dans les archives?

Dans les cinq prochains chapitres, nous explorerons ces questions.

³ Le terme correspond ici à l'indexation thématique qui accompagne généralement la description des documents d'archives selon les *Règles pour la description des documents d'archives* (RDDA).

4. Plan du mémoire

Dans le premier chapitre, nous présenterons la revue de la littérature. Nous effectuerons d'abord un survol de la description de la danse dans les services documentaires. Puis nous définirons quelques concepts. Nous présenterons ensuite deux cas exemplaires de vocabulaire de description de la danse. Le premier cas est celui du thésaurus de descripteurs Collier, développé par Clifford Collier, qui est inclus dans la Base de données intégrée de danse canadienne, un outil de gestion des collections distribué par la Dance Collection Danse, un centre d'archives et de recherche de la danse torontois (Adams, Bowring et Collier 2008). Le second cas est celui des vedettes-matière du *Dictionary catalog of the Dance Collection* de la Jerome Robbins Dance Division de la New York Public Library for the Performing Arts (NYPL 1974). Suite à la présentation des deux cas de vocabulaire de description de la danse, nous discuterons de la pertinence de l'expérience esthétique comme point d'accès aux archives de la danse. Finalement, nous parlerons d'extraction automatique de connaissances.

Dans le deuxième chapitre, nous traiterons de méthodologie de recherche. D'abord, nous justifierons le choix du thésaurus de descripteurs Collier comme vocabulaire de description de la danse dans les archives et nous présenterons la démarche pour en extraire une liste de termes. Nous expliquerons ensuite comment nous avons recueilli un vocabulaire structuré de l'expérience esthétique de la danse à partir d'un corpus de textes littéraires grâce à des méthodes d'extraction de termes et de classification automatique. Le corpus littéraire utilisé lors de notre expérimentation est constitué de deux recueils de textes de l'auteur français Stéphane Mallarmé, soit *Divagations* et *Poésies*. Nous justifierons ce choix.

Dans le troisième chapitre, nous présenterons et nous analyserons les résultats obtenus. Nous pourrions ainsi atteindre les objectifs de notre recherche et répondre aux questions de recherche. D'abord, nous présenterons et nous analyserons les descripteurs qui composent le thésaurus Collier pour atteindre le premier des objectifs de notre recherche, soit analyser un vocabulaire de description de la danse dans les archives. Ensuite, nous présenterons et nous analyserons le vocabulaire structuré de l'expérience esthétique de la danse moderne extrait d'un corpus de textes de l'auteur Stéphane Mallarmé pour atteindre le deuxième objectif de notre recherche, soit analyser un vocabulaire de représentation de la danse dans la littérature recueilli grâce à des méthodes d'extraction automatique de connaissances. Puis nous comparerons les descripteurs du thésaurus Collier et le vocabulaire structuré de l'expérience esthétique de la danse moderne en portant une attention particulière à la question de l'expérience esthétique pour atteindre le troisième objectif de notre recherche, soit comparer les deux vocabulaires afin d'en distinguer la complémentarité quant à la description de l'expérience esthétique. Nous soulèverons finalement les limites de la recherche. Nous serons alors en mesure de répondre aux deux questions qui sous-tendent notre recherche.

Dans la conclusion, nous reviendrons sur les grandes lignes de notre recherche. Nous mettrons ensuite en exergue les contributions autant théoriques, méthodologiques que pratiques de notre recherche au domaine des sciences de l'information. Finalement, nous évoquerons des recherches futures qui peuvent découler de nos travaux.

5. Domaines de recherche

Notre recherche se situe au confluent du domaine des arts et de celui des sciences de l'information. La danse et la littérature en composent la part liée aux arts. L'indexation dans les archives et l'extraction automatique de

connaissances en composent la part liée aux sciences de l'information. En se basant sur les onze approches d'analyse du domaine des sciences de l'information telles que définies par Birger Hjørland (2002), notre recherche s'inscrit dans les « special classifications and thesauri ». Elle s'insère aussi parfaitement dans la mouvance des humanités numériques.

Les humanités numériques (*digital humanities*) sont ce champ de recherche qui a émergé de l'amalgame des technologies de l'information aux méthodologies des lettres et sciences humaines. À l'origine des humanités numériques se trouve ainsi l'informatique appliquée aux humanités (*humanities computing*), discipline apparue au milieu du 20^e siècle avec la réalisation assistée par ordinateur d'une concordance de l'oeuvre de saint Thomas d'Aquin (Busa 1974-1980; Busa 2004; Hockey 2004; Dalbello 2011 : 481). Si l'utilisation des outils informatiques concerne d'abord principalement la compilation de dictionnaires, elle s'étend rapidement à l'analyse stylistique et à l'étude de la paternité littéraire (Hockey 2004), puis naissent alors les humanités numériques, regroupant des sujets de recherche aussi divers que l'établissement de normes de métadonnées ou l'étude des nanotechnologies dans une perspective artistique (Svensson 2009). Les humanités numériques sont encore maintenant un champ aux contours en constante redéfinition. Comme le souligne Patrik Svensson (2012), « the terrain of digital humanities is not stable nor fixed ».

Chapitre 1. Revue de la littérature

1.1. Introduction

Dans ce chapitre, nous effectuerons d'abord un survol de la description de la danse dans les services documentaires. Puis nous définirons les concepts de langage documentaire, de thésaurus et de liste de vedettes-matière. Nous présenterons ensuite deux cas exemplaires de vocabulaire de description de la danse. Le premier cas est celui du thésaurus de descripteurs Collier qui est inclus dans la Base de données intégrée de danse canadienne, un outil de gestion des collections distribué par la Dance Collection Danse, un centre d'archives et de recherche de la danse torontois (Adams, Bowring et Collier 2008). Le second cas est celui des vedettes-matière du *Dictionary catalog of the Dance Collection* de la Jerome Robbins Dance Division de la New York Public Library for the Performing Arts (NYPL 1974). Suite à cela, nous discuterons de la pertinence de l'expérience esthétique comme point d'accès aux archives de la danse. Finalement, nous parlerons d'extraction de connaissances, en particulier d'extraction de termes et de classification automatique.

1.2. Description de la danse

Comme nous l'avons déjà mentionné, les pratiques de documentation de la performance sont relativement éclatées et peu explicitées (Evens et Hautekeete 2011: 162). La littérature à ce sujet est parcellaire; le territoire reste à explorer. C'est pourquoi nous n'avons pas la prétention de tracer dans les pages qui suivent un portrait d'une impossible exhaustivité. Nous mettrons plutôt en exergue des cas exemplaires de langages documentaires de la danse fréquemment cités dans les écrits. Nous discuterons ainsi du thésaurus de

descripteurs Collier et nous discuterons ensuite des vedettes-matière du *Dictionary catalog of the Dance Collection*.

Le manque de documentation à ce sujet ne nous permettra toutefois pas de traiter de l'application aux arts d'interprétation des listes de vedettes-matière comme le *Répertoire d'autorité-matière encyclopédique et alphabétique unifié*. La Bibliothèque nationale de France, qui gère le RAMEAU en collaboration avec le ministère de l'Enseignement supérieur et de la Recherche, abrite pourtant un département des Arts du spectacle; cela sans compter les collections sur la danse du département de la Bibliothèque-musée de l'Opéra (BnF 2012). Nous ne traiterons pas non plus spécifiquement des *Library of Congress Subject Headings*. Toutefois, la Library of Congress, qui gère les LCSH, est membre de la Dance Heritage Coalition, tout comme la Jerome Robbins Dance Division, et une certaine collaboration existe entre les deux organismes, puisque la Dance Division via la DHC participe au Name Authority Cooperative Program du Program for Cooperative Cataloging de la Library of Congress depuis 1997 (Desalme 2007 : 20). Ainsi, les LCSH seront mentionnées lorsque nous discuterons du *Dictionary catalog of the Dance Collection*.

Soulignons également que, puisque nos objectifs de recherche portent sur le vocabulaire de description et le vocabulaire de représentation de la danse, sur des langages documentaires, sur des terminologies, et non pas sur des ontologies de la danse, modèles conceptuels du domaine, nous ne traiterons pas d'ontologie dans la présente revue de la littérature. Nous ne traiterons pas non plus de l'application de normes et de formats d'encodage aux arts d'interprétation bien que plusieurs travaux de recherche se penchent sur la question. Soulignons à ce propos les efforts de Patrick Le Boeuf pour adapter le modèle conceptuel d'information bibliographique FRBR (*Functional Requirements for Bibliographic Records*) aux arts d'interprétation – malgré des résultats mitigés selon l'aveu même de l'auteur – ou pour arrimer le modèle

FRBR et le modèle conceptuel d'information muséographique du Comité international pour la documentation du Conseil international des musées, le CIDOC CRM (*Conceptual Reference Model*), dont est issue la norme ISO 21127:2006 pour l'échange d'informations sur le patrimoine culturel, dans le but d'appliquer ce modèle hybride aux arts d'interprétation (Le Boeuf 2002; Le Boeuf et Miller 2005; Le Boeuf 2006; Bekiari, Doerr et Le Boeuf 2010). Soulignons également les initiatives de la Visual Resources Association que sont le VRA Core, un schéma d'encodage du patrimoine culturel, et la publication de l'ouvrage *Cataloging cultural objects : a guide to describing cultural works and their images* (Baca et coll. 2006).

Puisque la danse est liée au patrimoine culturel immatériel, nous excluons d'office de la revue de la littérature les langages et autres outils documentaires qui ont pour objet le patrimoine culturel matériel, entre autres les arts visuels. Nous ne traiterons pas, par exemple, du *Art & Architecture Thesaurus* du Getty Vocabulary Program, pourtant un modèle dans le domaine des thésaurus (Aitchison et Dextre Clarke 2004 : 5; Getty Research Institute 2012).

Finalement, même si nous utilisons le terme *vocabulaire* dans l'énoncé de nos objectifs de recherche, nous ne traiterons pas des travaux de recherche qui concernent le vocabulaire du corps, terme employé par certains théoriciens de la danse. Bergesen et Jones (1992 : 169) s'expriment ainsi à ce propos :

The expressive arts obviously constitute a means of communication, and as such are clearly susceptible to linguistic analysis. [...] Modern dance is described in terms of its movement vocabulary and there are syntactical rules for combining these body gestures into a larger structure. Choreography is dance syntax for bodily vocabulary.

S'il n'existe pas en danse de notation universelle, pas de partition comme en musique (Potter 2001 : 11), certains standards de notation du mouvement du corps sont d'utilisation relativement répandue : la notation Eskol-Wachmann, la notation Feuillet, la notation Benesh et la Labanotation de Rudolf Laban (Desalme 2007 : 15). Certains chercheurs envisagent ainsi la possibilité d'utiliser certaines de ces notations dans la description de la documentation de la danse (El Raheb et Ioannidis 2011).

Avant de présenter les deux cas exemplaires de vocabulaire de description de la danse, le thésaurus de descripteurs Collier et les vedettes-matière du *Dictionary catalog of the Dance Collection*, définissons dans un premier temps les concepts suivants : langage documentaire, thésaurus et liste de vedettes-matière.

1.2.1. Qu'est-ce qu'un langage documentaire?

Un langage documentaire est, entre autres, un « [l]angage contrôlé et normalisé, utilisé dans un système documentaire pour l'indexation et la recherche » (Dégez et Ménillet 2001 : 28). Un langage documentaire est « un sous-ensemble structuré de termes extraits de la langue naturelle dont on a contrôlé la forme et la signification » (Hudon et coll. 2009 : 85). Plus encore, selon Michèle Hudon (2009 : 260), un langage documentaire est un :

Langage artificiel, contrôlé et normalisé, constitué de représentations de concepts et de relations entre ces concepts, utilisé pour l'indexation et la recherche documentaire. Le langage documentaire permet la représentation précise et univoque des concepts constitutifs du sujet d'un document ou d'une question soumise au système d'information.

Un langage documentaire se compose ainsi d'une structure relationnelle et d'un lexique. C'est un vocabulaire contrôlé – par opposition au vocabulaire

libre. Selon Birger Hjørland (2008 : 89), « [t]he principle of controlled vocabulary is essentially a way of avoiding synonyms and homonyms as indexing terms by using standardized vocabulary ». Un langage documentaire est ainsi biunivoque, c'est-à-dire qu'un terme y représente un et un seul concept, un concept y est représenté par un et un seul terme.

Les langages documentaires se divisent, selon certains auteurs, en deux principales catégories : les langages postcoordonnés, comme le thésaurus de descripteurs, et les langages précoordonnés, comme la liste de vedettes-matière (Hudon et coll. 2009 : 86). Lorsque le langage documentaire est postcoordonné, les termes utilisés pour décrire le sujet du document restent indépendants les uns des autres jusqu'au moment de la recherche. Lorsque le langage documentaire est précoordonné, l'ordre des termes utilisés pour décrire le sujet du document est déterminé au moment de l'indexation.

Certains auteurs divisent autrement les langages documentaires. Ils parlent de langages de type classificatoire, comme la classification Dewey, et de langages de type combinatoire – que Georges Van Slype nomme aussi langages d'indexation – comme les listes de vedettes-matière et les thésaurus (Van Slype 1987 : 19; Waller 1999). Plus précisément, selon Georges Van Slype (1987 : 23), les listes d'autorité et les thésaurus de descripteurs sont des langages combinatoires contrôlés lorsqu'ils sont utilisés en contexte d'indexation. Suzanne Waller considère quant à elle que les listes de vedettes-matière « pourraient être combinatoires » (1999 : 196), mais que les « vrais langages à structure combinatoire » (1999 : 203) sont les lexiques et les thésaurus. Elle souligne aussi les limites de l'utilisation des langages de type classificatoire dans le processus d'indexation (Waller 1999 : 195).

Notons au passage que les langages documentaires sont des systèmes d'organisation des connaissances parmi d'autres (Hodge 2000).

1.2.2. Qu'est-ce qu'un thésaurus?

En sciences de l'information, les définitions du thésaurus abondent dans la littérature (Thomas 2004 : 24). Selon Danièle Dégez et Dominique Ménillet (2001 : 44), un thésaurus est d'abord une « [I]liste organisée de termes contrôlés et normalisés (descripteurs et non descripteurs [sic]) servant à l'indexation des documents et des questions dans un système documentaire ». Les deux auteures nomment *thésaurofacettes* les thésaurus qui associent la présentation alphabétique et la présentation à facettes inspirée des travaux de Shiyali Ramamrita Ranganathan (Dégez et Ménillet 2001 : 44). Selon Michèle Hudon (2009 : 263), le thésaurus se définit comme un « [I]angage documentaire de type combinatoire fondé sur une structuration hiérarchisée d'un ou de plusieurs domaines de la connaissance et dans lequel les concepts sont représentés par des termes de la langue naturelle dont on a contrôlé la forme et la signification ». Soulignons, dans la dernière définition, l'idée du thésaurus comme structuration d'un ou des domaines de connaissance, ce qui souligne sa nature de langage spécialisé et non encyclopédique. Selon Suzanne Waller (1999 : 205), le thésaurus est le « langage d'indexation par excellence ».

Nous avons déjà mentionné qu'un langage documentaire se compose d'un lexique, mais aussi d'une structure relationnelle. Plusieurs relations sémantiques peuvent ainsi exister entre les termes d'un thésaurus unilingue : une relation d'équivalence, une relation de hiérarchie, une relation d'association (Waller 1999 : 208-210). La relation d'équivalence est une « relation d'équivalence intralinguistique entre descripteurs et non-descripteurs d'une même langue naturelle » (Hudon 2009 : 25). La relation de hiérarchie est une « relation hiérarchique entre descripteurs d'une même langue naturelle » (Hudon 2009 : 26). La relation d'association est une « relation associative entre descripteurs d'une même langue naturelle » (Hudon 2009 : 26).

Pour répondre à ceux qui sont tentés de douter de la pertinence des thésaurus à l'ère numérique, plusieurs auteurs mentionnent un intérêt renouvelé pour le thésaurus (Roe et Thomas 2004 : 1; Hudon 2009 : 20). Jean Aitchison et Stella Dextre Clarke (2004 : 13) soulignent l'importance des thésaurus dans la recherche d'informations malgré de possibles réticences :

We face a paradox. Ostensibly, the need and the opportunity to apply thesauri to information retrieval are greater than ever before. On the other hand, users resist most efforts to persuade them to apply one.

Ajoutons à cela la nécessité de développement de thésaurus de descripteurs pour l'indexation des documents visuels, les archives vidéo de la danse par exemple, qui ne peuvent, pour des raisons évidentes, entièrement bénéficier des avancées dans le domaine de la recherche plein texte et de l'indexation par extraction automatique.

1.2.3. Qu'est-ce qu'une liste de vedettes-matière?

Une liste de vedettes-matière, est un « [l]angage documentaire composé de vedettes matière [sic] et éventuellement de subdivisions affranchies, et pouvant comporter des relations sémantiques entre vedettes matière [sic] » (Dégez et Ménillet 2001 : 30).

Une vedette-matière est une « [s]uite ordonnée de termes représentant synthétiquement le sujet ou l'un des sujets d'un document » (Dégez et Ménillet 2001 : 44). Une vedette-matière est « formée d'une tête de vedette et d'une ou de plusieurs subdivisions » (Hudon et coll. 2009 : 87). La tête de la vedette-matière représente le sujet; les subdivisions, les aspects par lesquels est abordé le sujet. Les subdivisions peuvent aussi spécifier le lieu, la temporalité ou la forme (Waller 1999 : 198).

Selon Suzanne Waller, la liste de vedettes-matière « pose de vrais problèmes d'indexation » (Waller 1999 : 199). En effet, il faut comprendre qu'un tel système est né dans un environnement non informatisé où la postcoordination était plus difficilement applicable. On peut songer au système de classeurs et de fiches papier, qui ne permettait évidemment pas une recherche plein texte.

Pour plus de clarté, voici un tableau comparatif des deux types de langages documentaires que nous venons de présenter.

Tableau 1. Caractéristiques des listes de vedettes-matière et des thésaurus (tiré de Hudon 2009 : 45)

Liste de vedettes-matière	Thésaurus
Langage documentaire utilisé principalement pour l'indexation et la recherche de <i>documents</i> dans les collections documentaires	Langage documentaire utilisé principalement pour l'indexation et la recherche d' <i>information</i> dans les documents
Outil de type <i>pragmatique</i> dont l'organisation interne reflète celle d'une collection documentaire	Outil de type <i>philosophique</i> dont l'organisation interne reflète celle d'un champ d'activité ou d'une discipline
Décrit souvent l'ensemble des connaissances (Langage documentaire <i>encyclopédique</i>)	Décrit de préférence un seul domaine, ou quelques domaines de la connaissance liés entre eux (Langage documentaire <i>spécialisé</i>)
Unité de base (<i>vedette-matière</i>) représente un <i>sujet</i> , c'est-à-dire une combinaison de concepts, d'aspects, de points de vue, etc.	Unité de base (<i>descripteur</i>) représente un <i>concept</i>

Unité de base (vedette-matière) est pleinement signifiante	Unité de base (descripteur) tire sa signification des rapports qu'elle entretient avec d'autres concepts et/ou termes
Structure hiérarchique lâche	Structure hiérarchique logique et rigoureuse
Pleine efficacité en contexte traditionnel (<i>précoordination</i>)	Pleine efficacité en contexte automatisé (<i>postcoordination</i>)
Créée et structurée par et pour des bibliothécaires et généralement peu exploitée par les usagers et par les moteurs de recherche informatiques	Créé et structuré par des spécialistes pour une grande variété d'utilisateurs d'information et exploité par les moteurs de recherche informatiques

Voyons maintenant les deux cas exemplaires de vocabulaire de description de la danse : le thésaurus de descripteurs Collier et les vedettes-matière du *Dictionary catalog of the Dance Collection*.

1.2.4. Thésaurus de descripteurs Collier

Au Canada, un acteur majeur des archives de la danse est la Dance Collection Danse. L'organisme bénéficie d'une reconnaissance au sein même du milieu de la danse. En conséquence « [m]any dance artists and choreographers across Canada choose DCD as the final destination for their entire collection » (LaFrance 2011: 69). Il importe de souligner que la DCD a émergé du milieu de la danse; « DCD did not grow out of an existing heritage's institution initiative but rather out of the dance community itself » (LaFrance 2011: 70). Fondée à Toronto au milieu des années 1980 par deux danseurs, Lawrence et Miriam Adams, la DCD est à l'origine de plusieurs initiatives visant à promouvoir la préservation de la danse (Bowring 2011 : 276; LaFrance

2011 : 49). Elle regroupe principalement un centre d'archives et de recherche, un musée ainsi qu'une maison d'édition (Dance Collection Danse 2012). Lawrence Adams (2004) a d'ailleurs publié un guide intitulé *Building your legacy : an archiving handbook for dance*, qui s'adresse aux archivistes néophytes, particulièrement aux artistes ou aux compagnies de danse. Le centre d'archives participe de plus à Artefacts Canada (Bowring 2011 : 284; Patrimoine canadien 2012).

La DCD s'implique également dans le projet Grassroots Archiving en collaboration avec la Société canadienne des études en danse (Dance Collection Danse 2012). Le projet Grassroots Archiving soutient le développement d'un réseau communautaire de support à la préservation de la danse à travers le Canada. Les professionnels de la danse formés à la préservation des documents et des artefacts de la danse peuvent alors devenir des références dans leur propre communauté.

La Dance Collection Danse distribue – et utilise – aussi la Base de données intégrée de danse canadienne, développée par Lawrence Adams et Clifford Collier, un outil de gestion des collections qui inclut le thésaurus de descripteurs Collier (Adams, Bowring et Collier 2008). Le thésaurus a été développé par Clifford Collier, un diplômé en bibliothéconomie et en sciences de l'information de l'University of Western Ontario (Adams, Bowring et Collier 2008 : 39). Il est disponible en version originale anglaise ou en version française, traduite par Marie Claire Forté. Il comporte 225 termes, principalement des noms communs, spécifiquement applicables au domaine de la danse. Il présente les descripteurs et les non-descripteurs en ordre alphabétique. Il ne peut toutefois être consulté ni par champs sémantiques, ni selon un index permuté. Les termes du thésaurus Collier sont liés entre eux par des relations d'équivalence intralinguistique (indiquées par le symbole *v/s* pour *voir source*) et par des relations d'association (indiquées par le symbole *v/a*

pour *voir aussi*), mais ne sont pas liés entre eux par des relations de hiérarchie, du moins pas par des relations de hiérarchiques explicites. L'objectif qui sous-tend le thésaurus Collier est « de regrouper des mots et des termes communément utilisés sous un terme repère subjectif qui peut être entré dans une base de données » (Adams, Bowring et Collier 2008 : 39).

Dans la figure 1, nous présentons un extrait du thésaurus Collier. Les relations d'équivalence intralinguistique pour le terme *ballet* incluent des termes synonymes comme *ballet classique* et *ballet moderne*. Une relation d'association existe quant à elle avec le terme *ballet de cour*. Le ballet de cour est une forme de danse précurseure du ballet.

Figure 1. Extrait du thésaurus Collier : le terme *ballet*
(voir aussi l'extrait original en annexe 1)

<p>BALLET</p> <p>Note d'application : utilisé pour les objets se rapportant à la danse théâtrale présentée dans une forme artistique stylisée qui a été développée en Italie et en France au 17e siècle, et qui s'est propagée en Europe, en Amérique et en Asie.</p> <p>Pour les objets se rapportant à la technique du ballet</p> <p>UTILISER : TECHNIQUE + BALLET</p> <p>v/s : ballet classique ballet d'école ballet moderne balletto ballo danse d'école danse opératique danze accademica</p> <p>v/a : BALLET DE COUR</p>
--

1.2.5. Vedettes-matière du *Dictionary catalog of the Dance Collection*

Aux États-Unis, un acteur majeur de la documentation de la danse est la Dance Heritage Coalition, qui regroupe plusieurs prestigieuses institutions dont la Jerome Robbins Dance Division de la New York Public Library for the Performing Arts – un centre de documentation de la danse de réputation internationale (Matluck Brooks 2011 : 447) – et la Library of Congress (DHC 2012). La DHC a été fondée au début des années 1990 dans la foulée du rapport *Images of American dance : documenting and preserving a cultural heritage* financé par le National Endowment for the Arts et la Andrew W. Mellon Foundation (Keens, Hansen Kopp et Levine 1990; Couch 1994 : 42; Procházka 2006 : 16).

Nena Couch (1994 : 53) souligne que les vedettes-matière de la Library of Congress ne répondent pas aux besoins de la documentation de la danse. Selon l'auteure, « [t]he list lacks specificity and is too general for both large and small collections » (Couch 1994: 53). Cela même si les œuvres chorégraphiques sont soumises au dépôt légal aux États-Unis depuis l'adoption du Copyright Act de 1976 (Desalme 2007 : 16). Cependant, depuis 1997, la DHC verse son fichier de notices d'autorité au fichier national de la Library of Congress (Le Boeuf 2000 : 168; Desalme 2007 : 20). Certains ajustements doivent être effectués lors du passage d'un fichier à l'autre (LC 2009 : 16-19; Procházka 2006 : 17) étant donné le caractère encyclopédique de la liste de vedettes-matière de la Library of Congress. Par exemple, la vedette-matière de la chorégraphie *Apollo* de George Balanchine passe de *Apollo (Balanchine)* à *Apollo (Choreographic work : Balanchine)*. Selon Patrick Le Boeuf (2000 : 169), « [c]es notices d'autorité servent notamment à l'indexation matière ». Les

notices d'autorités versées par la DHC proviennent principalement de la Jerome Robbins Dance Division (Desalme 2007 : 20).

Il faut dire que la Dance Division est plus qu'une simple bibliothèque. En effet, selon la toute première conservatrice de la Dance Division, Genevieve Oswald (1968 : 146) :

The collection is much more than a library in the usual sense to the art it serves ; it is an archive aiming at total documentation where not only the history but the continuity of the art of dance is being preserved.

La Dance Division a développé au fil des ans un « extensive authority file including names, subjects, and titles » (Couch 1994 : 54). Dans les années 1960, sous la direction de Genevieve Oswald, « [t]he decision was made to create an integrated catalog where the reader could find, in a single place, all material, whatever its form, on a particular subject » (Lourdou 1970 : 17). Le titre de l'ouvrage qui en a résulté – publié en 1974 et plusieurs fois mis à jour par la suite – est *Dictionary catalog of the Dance Collection : a list of authors, titles, and subjects of multi-media materials in the Dance Collection of the Performing Arts Research Center of the New York Public Library* (NYPL 1974; Procházka 2006 : 15). Il regroupe les documents de la collection de la Jerome Robbins Dance Division classés selon plus de 8 000 vedettes-matière et 45 000 notices d'autorité (Oswald 1968 : 151; Lourdou 1970 : 22). Notons que le catalogue est maintenant disponible pour la consultation en ligne (NYPL 2012).

La tête des vedettes-matière du *Dictionary catalog* peut être un nom (de personne, de compagnie, d'organisation, de conférences ou d'agences gouvernementales), un titre, un lieu géographique ou un sujet. Les titres des oeuvres de ballet et de danse moderne sont suivis du nom du chorégraphe entre parenthèses. Cinq subdivisions de vedette-matière sont possibles : *Works by*,

Works about, Visual works, Music scores, Audio materials (NYPL 1974; Armelagos 1977: 30).

Dans la figure 2, nous présentons un extrait du *Dictionary catalog*. La tête de la vedette-matière est le nom anglophone de la chorégraphie *Après-midi d'un faune* (*Afternoon of a Faun*) auquel est accolé entre parenthèses le nom de du chorégraphe, dans ce cas-ci *Lifar* pour Serge Lifar. La vedette-matière se divise en deux subdivisions : *Works about*, qui présente la référence d'un article de Peter Williams paru dans la revue *Dance and Dancers*, et *Visual Works*, qui présente la référence de deux photographies de la danseuse Nadia Nerina par P. A. Reuter également parues dans *Dance and Dancers*.

Figure 2. Extrait du *Dictionary catalog of the Dance Collection* : la chorégraphie *Après-midi d'un faune* de Serge Lifar

<p>AFTERNOON OF A FAUN (LIFAR) Chor: Serge Lifar; mus: Claude Debussy; scen & cos: after Léon Bakst. First perf: Paris, Théâtre National de l'Opéra, Mar 18, 1935.//Revived: London, Covent Garden, Autumn 1954; Paris Opera Ballet.</p> <p><i>Works about</i></p> <p>Williams, Peter. International exchange, 2: The first company of France in London. (In: [*MGZA Dance and dancers. London. Nov 1954,] p 13-16. illus) Review of the Paris Opera Ballet during its London season. NN 73-7130637</p> <p><i>Visual works</i></p> <p>Lifar, Serge 10. Photograph, no credit given. In L'après-midi d'un faune. NN 68-250532</p> <p style="text-align: right;">[*MGZE]</p> <p>Nerina, Nadia. Two photographs by P. A. Reuter. (In: [*MGZA Dance and dancers. London. Sept 1956,] p 33) Dancing with Alexis Rassine in L'après-midi d'un Faune and in Fireworks. NN 68-206173</p>
--

Selon Nena Couch (1994 : 54), les vedettes-matière peuvent aider à l'accès du matériel non-livre des collections en danse entre autres en favorisant leur inclusion dans le catalogue. D'ailleurs, même les archives de la Dance Division sont en majeure partie indexées dans le catalogue disponible en ligne de la New York Public Library (NYPL 2012). La chose est d'autant plus importante que (Couch 1994: 55) :

For the dance librarian, the challenge may be in the great variety of physical formats and sizes that may exist in the collection – correspondence, photographs, posters, stage and costume designs, choreographic notes or notation, newspaper clippings, scrapbooks, music in manuscript and published form, sound recordings, video recordings and film in a variety of formats, costumes, and set pieces, to mention a few.

La majeure partie de la collection de la Jerome Robbins Dance Division est d'ailleurs effectivement composée de matériel non-livre (Armelagos 1977 : 30). Nena Couch (1994 : 54) souligne aussi l'importance du vocabulaire contrôlé : « [i]f there is no standard language through which a patron may communicate his or her search, the material may be as lost as if the library had never acquired it ». Sans compter que la limitation de la liste de vedettes-matière du *Dictionary catalog of the Dance Collection* au seul domaine de la danse la rend mieux adaptée à la description d'une collection spécialisée (Matluck Brooks 2011 : 472).

1.3. L'expérience esthétique comme point d'accès

Déjà dans l'introduction, nous avons mentionné plusieurs raisons qui justifient le choix de la danse comme sujet. Le monde de la danse, de l'art, et le monde des sciences de l'information ont tout intérêt à s'influencer mutuellement. À propos de la collection de la Jerome Robbins Dance Division,

Adina Armelagos (1977 : 29) affirme d'ailleurs qu'elle « is in continuous use, not only by scholars, but by choreographers and dancers ». La collection « is not just a depository of materials relating to dance but is a participant in the art form » (Armelagos 1977: 30). La collection est utilisée par les artistes; elle est partie intégrante du monde de la danse.

Comme l'a souligné la chorégraphe allemande Pina Bausch, « la danse n'est pas seulement une technique singulière » (Bavelier 2009). La danse n'est pas qu'un ensemble de contraintes matérielles et humaines. La performance n'est pas non plus réductible au texte qui parfois l'accompagne (Le Boeuf 2000). Plus encore, selon le philosophe américain John Dewey, l'objet d'art, « art product », ne doit pas être confondu avec l'oeuvre d'art, « work of art », qui émerge plutôt de l'interaction de l'objet d'art avec et dans l'expérience (Dewey 1934). L'oeuvre d'art est, ainsi présentée, indissociable de l'expérience esthétique. Toutefois, l'expérience esthétique ne semble pas incluse dans la description des archives de la danse. Comme le souligne pourtant Suzanne Waller (1999 : 157), « *[i]l ne faut pas transformer l'indexation en " résumé du pauvre " mais garder à l'esprit que sa raison d'être est de cibler l'essentiel [en italique dans le texte original] ».*

L'expérience esthétique est un point d'accès essentiel aux archives de la danse. Pourtant, dans nos recherches sur les vocabulaires de description de la danse dans les archives, il n'est pas mention de l'expérience esthétique. La théorie en sciences de l'information étant issue de la pratique, on peut supposer que plusieurs obstacles d'ordre pratique se sont dressés devant l'intégration de termes à propos de l'expérience esthétique au vocabulaire de description des archives de la danse. Certains argueront la nature évanescence, à première vue à tout le moins, de la danse et, surtout, de l'expérience esthétique. Trop subjective, trop instable, l'expérience esthétique ne devrait pas être considérée dans la description de la danse. L'impossibilité jusqu'à présent de développer

un vocabulaire de description de l'expérience esthétique qui puisse être utilisé dans la description des archives de la danse ne doit pourtant pas être envisagée comme une raison d'abdiquer, mais plutôt comme un obstacle à surmonter. Les opportunités nouvelles qu'offrent l'extraction de termes et la classification automatique nous permettront peut-être de le surmonter.

1.4. Extraction de termes et classification

L'extraction de termes et la classification, ou *clustering*, vont de pair. Plusieurs méthodes de classification s'appuient sur une entrée de termes initiaux (Ibekwe-SanJuan 2007 : 115). L'extraction de termes concerne principalement l'aspect terminologique; la classification concerne principalement l'aspect sémantique. Puisqu'un langage documentaire se compose d'un lexique et d'une structure relationnelle, les deux opérations sont essentielles dans la création d'un nouveau langage documentaire.

1.4.1. Qu'est-ce que l'extraction de termes?

L'extraction de termes consiste à extraire d'un corpus de textes des termes caractéristiques. Soulignons que la difficulté n'est pas d'extraire les mots, mais bien d'extraire les termes significatifs. Plusieurs techniques sont ainsi utilisées en traitement automatique des langues pour extraire des termes d'un corpus. Elles se divisent en deux principaux groupes : les techniques basées sur une approche linguistique (*linguistic levels*) (Jacquemin 2001 : 15) et les techniques basées sur une approche informatique (*computational techniques*) (Jacquemin 2001 : 26) aussi appelées probabilistes (Bourigault, Jacquemin et L'Homme 2001 : IX). Patrick Drouin (2003 : 99) parle de techniques linguistiques et de techniques statistiques. Mentionnons également l'existence de méthodes hybrides.

Les deux principales applications de l'extraction de termes sont l'indexation automatique et l'acquisition de connaissances terminologiques (Jacquemin 2001 : 9-10). L'indexation automatique est une « activity that exploits terms as means for assigning descriptors to documents in order to facilitate information access later on » (Jacquemin 2001: 12). Des documents textuels sont indexés grâce à l'extraction de termes significatifs qu'ils contiennent. L'acquisition de connaissances terminologiques est une « activit[y] for which the ultimate purpose is the construction of terminological database [sic] » (Jacquemin 2001: 13). Une terminologie est mise au point à partir des termes significatifs extraits d'un corpus. Le champ de l'acquisition de connaissances terminologique se divise lui-même en deux sous-champs : l'enrichissement de thésaurus déjà existant, ce qui implique l'accès à des données initiales, et l'acquisition de termes, où aucune donnée initiale n'est disponible (Jacquemin 2001 : 15).

L'évaluation de l'extraction de termes est similaire à l'évaluation de la recherche d'informations : « [i]n information retrieval, the goal is to retrieve all the documents relevant to a query and only these documents; in controlled indexing, the purpose is to extract all the occurrences of terms and variants and only them » (Jacquemin 2001: 35). Ainsi, on évalue la précision et le rappel (Jacquemin 2001 : 35; Drouin 2003 : 108). Rappelons que meilleure est la précision lorsque le bruit est moindre et meilleur est le rappel lorsque le silence est moindre (Jacquemin 2001 : 35). Le bruit est, dans le cas de l'extraction de termes, les termes non significatifs, ou parasites, malgré tout extraits; le silence est les termes pourtant significatifs qui n'ont pas été extraits.

1.4.2. Qu'est-ce que la classification?

D'abord, précisons que les méthodes d'analyse de données textuelles se répartissent en deux grandes catégories : les méthodes prédictives, supervisées,

de catégorisation (en anglais : *categorization* ou *classification*), qui doivent classer un nouvel objet dans une structure déjà existante, et les méthodes descriptives, de classification (en anglais : *clustering*), non supervisées, qui doivent produire une nouvelle structure à partir d'un ensemble d'objets (Ibekwe-SanJuan 2007 : 59). Comme les méthodes descriptives sont entre autres utiles « dans la découverte thématique ou dans des applications de traitement automatique des langues, pour découvrir automatiquement des classes sémantiques afin de peupler une ontologie ou de construire des ressources dictionnairiques » (Ibekwe-SanJuan 2007 : 61), dans l'analyse thématique de documents textuels non structurés (Forest 2012), nous explorerons certaines méthodes descriptives plus en détail.

Les algorithmes de classification peuvent être divisés en deux grandes catégories : les algorithmes plats, ou algorithmes de partitionnement, et les algorithmes hiérarchiques (Lebart et Salem 1994 : 111; Manning et Schütze 1999 : 498; Feldman et Sanger 2007 : 85; Ibekwe-SanJuan 2007 : 64). Les deux catégories sont parfois combinées dans certaines approches hybrides (Lebart et Salem 1994 : 127-133). Les algorithmes plats produisent un certain nombre de classes, regroupant des éléments similaires, mais les classes ne sont pas liées entre elles, alors que les algorithmes hiérarchiques produisent des classes positionnées dans une hiérarchie, donc liées entre elles (Manning et Schütze 1999 : 498). Exprimé plus formellement, on peut dire que les méthodes de partitionnement « produisent de simples découpages, ou partition de la population étudiée, sans passer par l'intermédiaire d'une hiérarchie. Ces dernières sont mieux adaptées à de très grands ensembles de données (plusieurs milliers d'individus à classer) » (Lebart et Salem 1994 : 111) et que les méthodes de classification hiérarchique « permettent d'obtenir à partir d'un ensemble d'éléments décrits par des variables (ou dont on connaît les distances

deux à deux), une hiérarchie de classes partiellement emboîtées les unes dans les autres » (Lebart et Salem 1994 : 111).

Pour plus de clarté, voici également présenté un tableau comparatif des deux grandes catégories d'algorithmes de classification (tableau 2).

Tableau 2. Caractéristiques de différents algorithmes de classification
(traduit de Manning et Schütze 1999 : 500)

Classification hiérarchique	Classification non hiérarchique
Privilégiée pour l'analyse détaillée des données.	Privilégiée lorsque l'efficacité est une considération ou que l'ensemble de données est très grand.
Fournit plus d'informations que la classification plate.	La méthode des k-means est la plus simple conceptuellement et devrait probablement être utilisée en premier lieu sur un nouveau jeu de données, puisque les résultats alors obtenus sont souvent suffisants.
Pas de meilleur algorithme (chacun des algorithmes décrits est considéré comme optimal pour certaines applications).	La méthode des k-means se base sur un simple espace euclidien de représentation et ne peut donc pas être utilisée pour plusieurs ensembles de données, par exemple, des données nominales comme les couleurs.
Moins efficace que la classification plate (pour n objets, on doit minimalement calculer une matrice $n \times n$ de coefficients de similarité et ensuite assurer la mise à jour de la matrice).	Dans les cas semblables, l'algorithme EM est la méthode de choix. Il permet une définition des classes et une répartition des objets basés sur des modèles probabilistes complexes.

La classification résultante est généralement rigide (*hard*) ou souple (*soft*) (Manning et Schütze 1999 : 499; Feldman et Sanger 2007 : 85; Ibekwe-SanJuan 2007 : 64). Lorsque la classification est rigide, un objet appartient à une seule grappe (Manning et Schütze 1999 : 499; Feldman et Sanger 2007 : 85; Ibekwe-SanJuan 2007 : 64). Lorsque la classification est souple, un objet est en partie assigné à une seule grappe, donc « un élément peut appartenir à des degrés différents à plusieurs classes » (Ibekwe-SanJuan 2007 : 64). La classification souple s'appuie ainsi sur les principes de logique floue. Autrement dit, « [i]n soft clustering, the membership is fuzzy – objects may belong to several clusters with a fractional degree of membership in each » (Feldman et Sanger 2007: 85). Il ne faut pas mélanger le *hard* et le *soft clustering* avec les modèles de *disjunctive clustering* dans lesquels plusieurs classes peuvent être attribuées à un objet sans en spécifier le degré d'appartenance (Manning et Schütze 1999 : 499).

Les algorithmes plats regroupent entre autres les algorithmes *k-means*, un algorithme itératif avec classification rigide, et EM, une famille d'algorithmes probabilistes avec classification souple (Manning et Schütze 1999 : 515-527; Feldman et Sanger 2007 : 85). Les algorithmes plats s'appuient principalement sur des mesures de distances comme le cosinus (Ibekwe-SanJuan 2007 : 68).

Les algorithmes hiérarchiques regroupent plusieurs types d'algorithmes, qui s'appuient principalement sur des mesures de similarité (Ibekwe-SanJuan 2007 : 68). Ceux-ci peuvent produire soit une classification *bottom-up*, ascendante ou agglomérative, soit une classification *top-down*, descendante ou divisive (Lebart et Salem 1994 : 113; Manning et Schütze 1999 : 500-501; Ibekwe-SanJuan 2007 : 65). Avec les algorithmes agglomératifs, à l'origine, chaque objet est seul dans une grappe et les grappes sont par la suite itérativement regroupées (Feldman et Sanger 2007: 85). Avec les algorithmes

divisifs, à l'origine tous les objets sont dans la même grappe, qui est ensuite séparée en grappes plus restreintes, qui seront elles-mêmes séparées et ainsi de suite (Manning et Schütze 1999 : 500-501; Feldman et Sanger 2007 : 85).

Pour plus de clarté, voici présentés deux schémas qui illustrent la classification agglomérative et la classification divisive (voir la figure 3 et la figure 4).

Figure 3. Classification hiérarchique agglomérative
(tiré de Forest 2006 : 56)

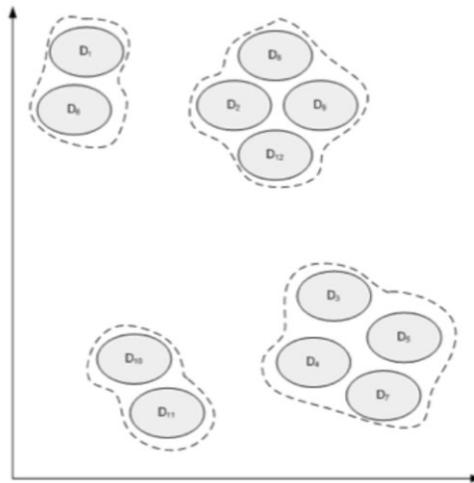
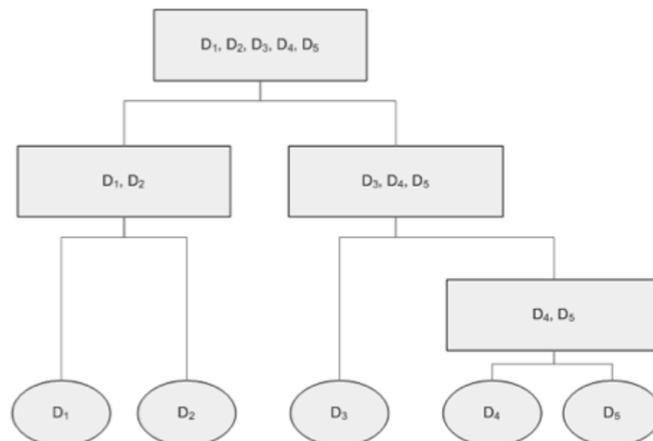


Figure 4. Classification hiérarchique divisive
(tiré de Forest 2006 : 56)



Soulignons que les algorithmes hiérarchiques produisent généralement des classifications rigides (Manning et Schütze 1999 : 499). La représentation de la classification peut s'effectuer de diverses manières, mais « [l]a représentation sous forme de *dendrogramme* constitue sans doute la représentation la plus parlante. » (Lebart et Salem 1994 : 113). Comme le soulignent Manning et Schütze (1999 : 497), « [c]lustering is a particularly important technique for EDA [Exploratory Data Analysis] in Statistical NLP [Natural Language Processing] because there is often no direct pictorial visualization for linguistic objects ». La classification est essentielle à l'analyse de données textuelles en traitement automatique des langues, car elle permet une visualisation des données autrement impossible.

1.5. Conclusion

Dans ce chapitre, nous avons d'abord effectué un survol de la description de la danse dans les services documentaires. Nous avons entre autres présenté deux cas exemplaires de vocabulaire de description de la danse : le thésaurus de descripteurs Collier et les vedettes-matière du *Dictionary catalog of the Dance Collection*. Nous avons ensuite souligné que l'expérience esthétique est une part essentielle de la danse, donc un point d'accès nécessaire aux archives de la danse. Et la description, en amont, influe grandement autant sur la préservation des archives que sur leur accès, en aval. La problématique est alors claire : le vocabulaire de description des archives de la danse peut bénéficier d'être enrichi de termes relatifs à l'expérience esthétique. Le monde des archives, des sciences de l'information, et le monde de la danse, de l'art, pourraient encore mieux se comprendre. Nous avons ainsi discuté des méthodes d'extraction de termes et de classification automatique, qui nous semblent ouvrir de nouvelles avenues de recherche pour permettre l'expérience esthétique comme point d'accès aux archives de la danse.

Dans le prochain chapitre, nous traiterons de la méthodologie employée lors de la réalisation de notre projet de recherche.

Chapitre 2. Méthodologie

2.1. Introduction

Dans ce chapitre, nous traiterons de la méthodologie employée lors de la réalisation de notre projet de recherche. D'abord, nous justifierons le choix du thésaurus de descripteurs Collier comme vocabulaire de description de la danse dans les archives et nous présenterons la démarche pour en extraire une liste de termes. Nous expliquerons ensuite comment nous avons recueilli un vocabulaire structuré de l'expérience esthétique de la danse moderne à partir d'un corpus de textes littéraires grâce à des méthodes de classification automatique et d'extraction de termes. Le corpus littéraire utilisé lors de notre expérimentation est constitué de deux recueils de textes de l'auteur français Stéphane Mallarmé, soit *Divagations* et *Poésies*. Nous justifierons ce choix.

2.2. Vocabulaire de description de la danse dans les archives : le thésaurus de descripteurs Collier

Comme vocabulaire de description de la danse dans les archives, nous avons sélectionné le thésaurus de descripteurs Collier (Adams, Bowring et Collier 2008). Les raisons de ce choix sont doubles. D'abord, comme nous l'avons précédemment mentionné, le thésaurus Collier est un outil en usage dans les archives de la danse au Canada. Ensuite, il est disponible en langue française, contrairement aux vedettes-matière du *Dictionary catalog of the Dance Collection* de la Jerome Robbins Dance Division de la New York Public Library for the Performing Arts (NYPL 1974). Puisque le corpus de textes d'où est extrait le vocabulaire de représentation de l'expérience esthétique de la danse est composé d'œuvres littéraires en langue française, il était préférable de le comparer avec un vocabulaire de description de la danse dans la même langue.

Pour obtenir une version numérique de la liste des descripteurs contenus dans le thésaurus Collier, nous avons numérisé une version papier du thésaurus à l'aide de la version 14.0 du logiciel de reconnaissance optique de caractères OmniPage Pro de la compagnie Nuance Communications. Nous avons corrigé à la main les erreurs de lecture. Nous avons également éliminé du fichier créé tout ce qui n'était pas descripteurs – distingués par l'utilisation d'une police de caractère en gras et d'une casse en majuscule – étant donné que l'objectif de notre recherche consiste à comparer deux vocabulaires, donc deux groupes de termes. Nous avons ainsi obtenu un fichier en format texte contenant une liste alphabétique des descripteurs extraits du thésaurus. Puisque les termes du thésaurus Collier sont liés entre eux par des relations d'équivalence intralinguistique et par des relations d'association, mais ne sont pas liés entre eux par des relations de hiérarchie, du moins pas par des relations de hiérarchiques explicites, nous avons également produit nos propres diagrammes hiérarchiques de descripteurs pour mettre en lumière certaines catégories de termes et certains liens hiérarchiques implicites.

2.3. Vocabulaire de représentation de la danse dans la littérature : quelques textes choisis de Stéphane Mallarmé

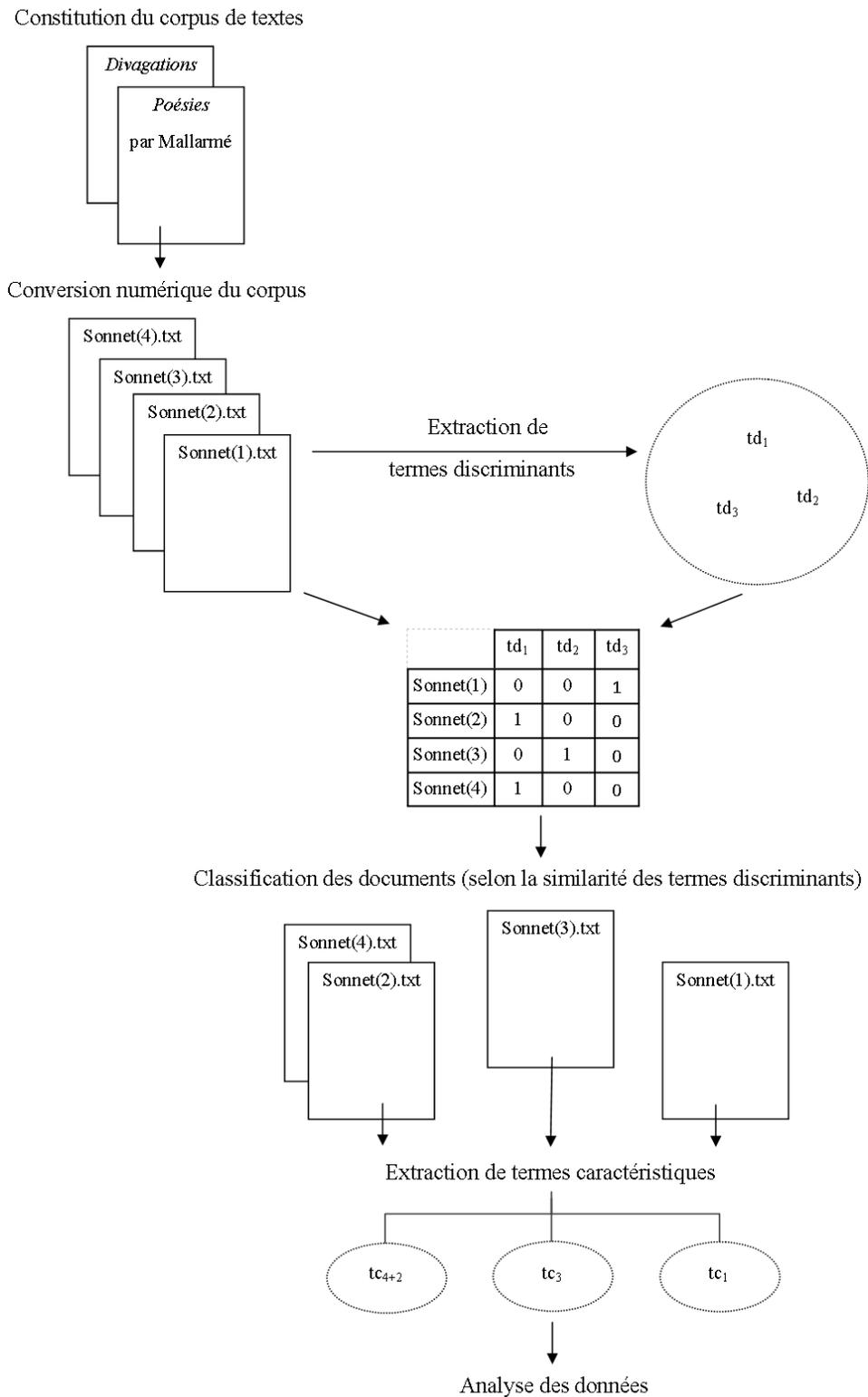
Pour recueillir un vocabulaire de représentation de la danse dans la littérature, nous avons travaillé à partir d'un corpus de textes de l'auteur français Stéphane Mallarmé. Nous avons appliqué sur le corpus des méthodes d'extraction de termes et de classification automatique. Nos recherches se concentrent ainsi plus précisément sur un vocabulaire de représentation de l'expérience esthétique de la danse moderne dans la littérature.

Il est également important de souligner que nous ne définirons pas l'expérience esthétique, car c'est justement ce que notre expérimentation cherche à décrire. Nous souhaitons éviter tout *a priori*.

Voici maintenant les grandes lignes de la méthodologie utilisée pour recueillir le vocabulaire. D'abord, un corpus de textes est sélectionné : les recueils *Poésies* et *Divagations* de Stéphane Mallarmé (voir la section 2.3.1 de ce chapitre). Une version numérique de ce corpus est ensuite constituée. Un certain nombre de fichiers, ou documents, sont créés (voir la section 2.3.2.1). Les fichiers textes sont alors transposés en base de données textuelles, un filtrage du lexique est effectué pour éliminer des termes non significatifs, des mots vides (voir la section 2.3.2.2) suivi d'un filtrage statistique pour optimiser le temps de traitement (voir la section 2.3.2.3). Nous avons ensuite regroupé les documents en classes (*clusters*) selon des termes discriminants (voir la section 2.3.2.4). Nous avons utilisé pour cela une méthode de classification hiérarchique (voir la section 2.3.2.5). Une fois la structure de classification générée, des termes caractéristiques à chacune des classes ont été extraits pour créer un vocabulaire structuré de l'expérience esthétique de la danse moderne (voir la section 2.3.2.6).

Pour plus de clarté, nous présentons un schéma qui résume la méthodologie dont nous venons de présenter les grandes lignes (voir la figure 5). Nous reviendrons ensuite plus en détail sur chacune des étapes de la méthodologie.

Figure 5. Schéma de la méthode de collecte du vocabulaire de représentation de la danse dans la littérature



2.3.1. Constitution d'un corpus de textes

2.3.1.1. Présentation du corpus

Puisque les œuvres complètes de Stéphane Mallarmé ne sont pas disponibles en version numérique, nous avons recensé les textes nécessaires à notre expérimentation publiés en version papier pour ensuite en colliger des versions numériques. Nous avons exclu de la recension : les textes – lettres, textes journalistiques ou autres – non retravaillés par l'auteur à des fins de publication en recueil, les ouvrages pédagogiques, les traductions, les préfaces et toutes les publications posthumes sauf *Poésies*, publié en 1899, quelques mois seulement après la disparition de Stéphane Mallarmé, et dont une première version a été publiée en 1887. La version de 1899 demeure toutefois l'édition de référence (Marchal 2009). Nous avons exclu les autres publications posthumes pour éviter les questionnements et les incertitudes à propos des œuvres inachevées, voire altérées (Marchal 2009). Les deux recueils qui constituent notre corpus sont ainsi *Divagations* (1897) et *Poésies* (1899).

2.3.1.2. Justification

Lors de l'évaluation des corpus de textes possibles, nous avons été confrontés à deux principales difficultés. Premièrement, la danse est souvent considérée comme un art mineur (Potter 2001 : 11). On peut, par exemple, constater à cet égard que la danse est exclue de l'icône système des arts de Georg Wilhelm Friedrich Hegel (Hegel 1835). Nous avons pu remarquer que le corpus de textes liés à la danse est relativement restreint par rapport aux corpus concernant certaines autres formes d'arts. Deuxièmement, la majorité des textes sur la danse que nous avons recensés étaient des textes critiques, qui s'attardent beaucoup aux détails du spectacle, aux contingences humaines, matérielles et temporelles, sans saisir, sans rendre compte de l'expérience esthétique qui jaillit de l'amalgame de ces contingences.

D'ailleurs, selon Stéphane Mallarmé, « [n]ommer un objet, c'est supprimer les trois quarts de la jouissance du poème qui est faite de deviner peu à peu : le suggérer voilà le rêve » (Mallarmé 2003 : 700). À propos de Stéphane Mallarmé, Mary Ann Caws (1998 : 86) écrit : « [w]hat he gives us is everything that comes after him ». Ainsi, pour obtenir un vocabulaire de représentation de l'expérience esthétique de la danse moderne dans la littérature, nous avons puisé aux sources de la modernité littéraire et de l'art moderne : dans l'œuvre mallarméenne.

Précisons que Stéphane Mallarmé est un auteur français de la fin du 19^e siècle. Il est un des poètes maudits de Paul Verlaine (Verlaine 1884). Selon Gérard Delfel (1951), Stéphane Mallarmé est un esthéticien. Plus encore, « son œuvre fut soutenue, enrichie, approfondie, galvanisée et rendue nécessaire par une assidue réflexion sur l'art » (Souriau 1951 : 7). Stéphane Mallarmé crée véritablement une œuvre artistique littéraire à partir de son expérience esthétique. Jean-Paul Sartre écrit ainsi de lui que « puisque l'Art doit être sa propre source, la méthode la moins impure pour s'élever du réel infâme à la Poésie, c'est encore de se faire guider par les œuvres d'autrui » (Sartre et Elkaïm-Sartre 1986 : 126). Pascal Caron (2006 : 48) va même jusqu'à affirmer, parlant des chroniques de Stéphane Mallarmé dans le journal *La Dernière Mode*, que « [p]eu importe le sujet sur lequel on l'interroge, le poète y trouve prétexte à un événement esthétique ». L'œuvre littéraire de Stéphane Mallarmé incarne ainsi une expérience esthétique.

Son œuvre littéraire est également à la base d'une trame de références artistiques qui unit la littérature, la musique, la danse et d'autres formes d'art. Et, si tous les textes de Stéphane Mallarmé ne mentionnent pas directement la danse, son œuvre est malgré tout liée à la danse d'une manière ou d'une autre (Richard 1961; Kristeva 1974 : 537; Block 1977 : 96; Levinson 1983; Caws 1994; Zachmann 2001). Gayle Zachmann (2001 : 188) écrit ainsi : « writers and

critics [...] have highlighted this poet's contributions to the theoretical underpinnings and reading of modern dance and/or the significance of his writings on dance for his own aesthetic ». Stéphane Mallarmé est un passionné de danse (Caws 1994 : 291), « un amoureux du ballet » (Caron 2006 : 48), un « métaphysicien du ballet » (Levinson 1983). Son *Après-midi d'un faune* a d'ailleurs inspiré le compositeur Claude Debussy, qui a composé à la fin du 19^e siècle *Prélude à l'après-midi d'un faune* (Caron 2006 : 257), sur lequel Vaslav Nijinski a chorégraphié *Après-midi d'un faune* en 1912 (NYPL 1974 : 42; Caron 2006 : 17). Au cours du 20^e siècle, de nouvelles chorégraphies inspirées de l'*Après-midi d'un faune* verront le jour; citons plus particulièrement celle de Andreas Pavley et Serge Oukrainski en 1919 (NYPL 1974 : 43), celle de Serge Lifar en 1935 (NYPL 1974 : 42), celle de Jerome Robbins en 1953 (NYPL 1974 : 44) et celle de Marie Chouinard en 1987 (Compagnie Marie Chouinard 2012). L'œuvre de Stéphane Mallarmé est ainsi intrinsèquement liée à la danse. Il affirmait d'ailleurs ceci : « [I]a danse seule, du fait de ses évolutions, avec le mime me paraît nécessiter un espace réel, ou la scène » (Mallarmé 1897 : *Crayonné au théâtre – Le genre ou des modernes*).

En résumé, nous avons creusé au coeur d'un art, la littérature, pour en faire émerger une description de l'expérience de l'art. Une expérience de l'art, unique et personnelle, subjective, celle de Stéphane Mallarmé, mais une expérience de l'art qui a eu une profonde influence sur la suite des choses, sur l'histoire de l'art moderne. En cela, elle possède une certaine universalité. Et la puissance des liens entre l'oeuvre littéraire mallarméenne et la danse permet d'appliquer la description de l'art extrait des textes de Stéphane Mallarmé au cas particulier de la danse.

2.3.2. Collecte des données

Présentons maintenant en détail la méthode de collecte du vocabulaire de représentation de la danse dans la littérature, la méthode de création d'un vocabulaire structuré de l'expérience esthétique de la danse moderne.

2.3.2.1. Conversion numérique des documents

Après avoir défini le corpus de textes avec lequel nous souhaitons travailler, nous en avons colligé une version numérique. Une base de données textuelles du recueil *Divagations* a déjà été créée au début des années 1990 (Matsumuro, Takeuchi et Kaneko 1991), mais les fichiers n'étaient ni compatibles avec les logiciels que nous souhaitons utiliser, ni disponibles. Nous avons alors consulté la page web du Projet Gutenberg, une bibliothèque numérique de livrets (Projet Gutenberg 2012). Ni le recueil *Divagations*, ni le recueil *Poésies* n'y étaient disponibles. Nous avons ensuite consulté Wikisource, une bibliothèque numérique libre (Wikisource 2012). Nous avons alors mis la main sur des versions numériques, en HTML, de l'édition de *Divagations* par Eugène Fasquelle de 1897 et de l'édition de *Poésies* par Émile Deman de 1899. Nous avons comparé les tables des matières des éditions en ligne aux tables des matières des éditions équivalentes incluses dans l'édition de La Pléiade des oeuvres complètes de Stéphane Mallarmé pour nous assurer d'une certaine complétude et qualité des versions numériques.

Nous avons alors copié le contenu textuel disponible en ligne, à l'exclusion des titres et des notes de bas de page, dans plusieurs fichiers de format texte. Chaque fichier représente un texte en prose, un poème ou un dialogue et donc un document. Par exemple, chaque poème contenu dans *Poésies* est enregistré dans son propre fichier. Le corpus comprend ainsi 119 documents et, avant le filtrage du lexique, 11 850 formes, ou mots distincts, sur 70 507 occurrences. Après le filtrage du lexique, nous avons obtenu un corpus

de 7 238 formes (*types*) sur 70 070 occurrences (*tokens*). Il s'agit d'un petit corpus en terme de fouille de textes. À titre de comparaison, le corpus FRANTEXT comprend plus de 4000 textes (FRANTEXT 2012).

Finalement, nous avons transformé l'ensemble des fichiers en base de données textuelles grâce à la version 3.2 du logiciel de fouille de données QDA Miner de la compagnie Provalis Research. Une fois la base de données textuelles créée, il était possible d'effectuer les opérations de filtrage du lexique, de classification automatique et d'extraction de termes grâce à la version 6.1.1 du logiciel d'analyse de contenu et de fouille de textes WordStat de la compagnie Provalis Research.

2.3.2.2. Filtrage du lexique

Avant d'effectuer les opérations d'extraction de termes et de classification automatique, le lexique doit être filtré pour retirer les données qui ne sont pas pertinentes pour l'analyse (Forest 2011). Plusieurs techniques de filtrage existent, dont la suppression de mots fonctionnels, à l'aide d'une liste de mots vides, ou antidictionnaire (*stop words*), la suppression par fréquence, les techniques de lemmatisation ou encore les techniques de racinisation (Forest 2011).

Lors de la réalisation de notre recherche, nous avons appliqué au corpus de textes la technique de filtrage à l'aide d'une liste de mots vides. Nous avons utilisé pour cela l'antidictionnaire de Jean Véronis (Véronis 2012). Nous avons également appliqué la technique de lemmatisation. Nous avons utilisé pour cela la liste de substitution disponible en français incluse à même le logiciel de fouille de textes WordStat.

La technique de filtrage à l'aide d'une liste de mots vides est très simple : une liste de termes non significatifs – par exemple, le déterminant *le* ou les déclinaisons de la conjugaison du verbe *avoir* – permet au logiciel de les

retirer du lexique initial. La technique de la lemmatisation mérite quant à elle quelques explications supplémentaires. Elle consiste à substituer un lemme, forme normalisée d'un mot, à ses différentes flexions. Par exemple, on substitue le singulier au pluriel ou un verbe à l'infinitif à ses déclinaisons (Lebart et Salem 1994 : 36). Étienne Brunet (2000 : 1) en a d'ailleurs discuté les avantages; entre autres, « [q]uand les formes verbales sont rassemblées derrière un même lemme, il est plus aisé de les rechercher dans le texte, et de tirer des conclusions de leurs fréquences cumulées ».

2.3.2.3. Filtrage statistique

Soulignons également que nous avons décidé d'un seuil optimal de sélection basée sur le *tf.idf* de 1000 items pour optimiser l'exécution de l'extraction de termes discriminants et la classification automatique (Forest 2009).

2.3.2.4. Extraction de termes discriminants

Une fois le lexique filtré, nous voici arrivés à l'étape d'extraction de termes discriminants, de termes qui ont un certain poids informatif (Ibekwe-SanJuan 2007 : 46), sur lesquels s'appuie la classification automatique des documents. Les fonctions de pondération du poids informatif prennent généralement en compte « la fréquence d'un terme dans un document donné » et « la distribution d'un terme dans l'ensemble du corpus » (Ibekwe-SanJuan 2007 : 46). Une des fonctions les plus utilisées est la fonction *tf.idf* ou *Term Frequency.Inverse Document Frequency* (Ibekwe-SanJuan : 46) dont voici l'équation (voir la figure 6).

Figure 6. Équation de la fonction *tf.idf*
(adapté de Ibekwe-SanJuan : 46)

$$W_{ij} = tf_{ij} \cdot idf_i$$

où W_{ij} est le poids du terme i dans le document j ;
 tf_{ij} est la fréquence du mot i dans le document j ;
 idf_i est l'inverse du nombre de documents dans lesquels le
 terme i apparaît.

La fonction *tf.idf* calcule le produit de la fonction de *Term Frequency* et de la fonction *Inverse Document Frequency*. L'idée qui sous-tend la fonction *tf.idf* est simple : « les bons candidats termes pour indexer un texte sont ceux qui apparaissent fréquemment dans ce texte [...] mais dans peu d'autres textes » (Ibekwe-SanJuan : 47). Ou, autrement formulé, selon Dominic Forest (2006 : 71) « un terme sera d'autant meilleur pour représenter le contenu d'une classe s'il est à la fois fréquent dans cette classe et rare dans l'ensemble des classes à analyser ». La pondération calculée par la fonction *tf.idf* permet un bon taux de rappel autant qu'une bonne précision. Soulignons qu'en pratique la fonction *tf.idf* doit être adaptée pour résoudre certains biais, par exemple, sa sensibilité aux mots dont le nombre d'occurrences est très élevé (Ibekwe-SanJuan : 47).

2.3.2.5. Classification automatique des documents

La pertinence de la classification automatique est établie à des fins de découverte thématique et de création de terminologies (Ibekwe-SanJuan 2007 : 61; Forest 2012). Ainsi, lors de notre expérimentation, nous avons utilisé une méthode de classification hiérarchique agglomérative dont voici le pseudo-algorithme (voir la figure 7).

Figure 7. Pseudo-algorithme de classification hiérarchique agglomérative
(adapté de Manning et Schütze 1999 : 502)

```

1  Étant donné: un ensemble  $\mathcal{X} = \{x_1, \dots, x_n\}$  d'objets
2                une fonction  $\text{sim}: \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \rightarrow \mathbb{R}$ 
3  pour  $i := 1$  à  $n$  faire
4       $c_i := \{x_i\}$  fin
5   $C := \{c_1, \dots, c_n\}$ 
6   $j := n + 1$ 
7  tant que  $C > 1$ 
8       $(c_{n_1}, c_{n_2}) := \arg \max_{(c_u, c_v) \in C \times C} \text{sim}(c_u, c_v)$ 
9       $c_j = c_{n_1} \cup c_{n_2}$ 
10      $C := C \setminus \{c_{n_1}, c_{n_2}\} \cup \{c_j\}$ 
11      $j := j + 1$ 

```

où c est une classe (*cluster*);

$\mathcal{X} = \{x_1, \dots, x_n\}$ est un ensemble de n objets à être classifié;

$C := \{c_1, \dots, c_n\}$ est un ensemble de classes (*clusters*);

$\mathcal{P}(\mathcal{X})$ est un ensemble de sous-ensembles de \mathcal{X} ;

sim est une fonction de similarité.

Rappelons que la classification est une méthode descriptive, donc non supervisée, qui permet de produire une nouvelle structure à partir d'un ensemble d'objets (Ibekwe-SanJuan 2007 : 59). Les algorithmes de classification peuvent être divisés en deux grandes catégories : les algorithmes de partitionnement et les algorithmes hiérarchiques (Lebart et Salem 1994 : 111; Manning et Schütze 1999 : 498; Feldman et Sanger 2007 : 85; Ibekwe-SanJuan 2007 : 64). Les algorithmes hiérarchiques s'appuient principalement sur des mesures de similarité (Ibekwe-SanJuan 2007 : 68) et ils peuvent produire soit une classification agglomérative, ou ascendante, soit une classification divisive, ou descendante (Lebart et Salem 1994 : 113; Manning et Schütze 1999 : 500-501; Ibekwe-SanJuan 2007 : 65). Lorsqu'une méthode de classification hiérarchique agglomérative est utilisée, comme c'est le cas dans notre expérimentation, chaque objet est d'abord seul dans une classe et les

classes sont par la suite itérativement regroupées (Feldman et Sanger 2007: 85). Autrement dit, chaque texte était seul dans sa classe et les classes de textes étaient par la suite itérativement regroupées jusqu'à créer une certaine structure de classes qui forme la base d'un vocabulaire structuré de l'expérience esthétique de la danse moderne.

2.3.2.6. Extraction de termes caractéristiques

L'étape de l'extraction de termes caractéristiques est relativement simple. Les classes de documents sont déjà formées; les textes, déjà regroupés en classes. Il ne restait plus qu'à extraire de chaque classe les termes qui ont le poids informatif le plus élevé dans les textes qui la composent pour ainsi créer un vocabulaire structuré de l'expérience esthétique de la danse moderne. Toutefois, puisque notre expérimentation est exploratoire et la démarche, encore à perfectionner, nous avons dû réaliser un certain nombre de manipulations humaines suite à l'extraction des termes caractéristiques pour obtenir un vocabulaire structuré.

2.4. Conclusion

Dans ce chapitre, nous avons traité de la méthodologie employée lors de la réalisation de notre projet de recherche.

D'abord, nous avons justifié le choix du thésaurus de descripteurs Collier comme vocabulaire de description de la danse dans les archives : il est en usage dans les archives canadiennes de la danse et il est disponible en langue française. Nous avons alors expliqué la démarche pour en extraire une liste de termes. Ensuite, nous avons expliqué comment nous avons recueilli un vocabulaire structuré de l'expérience esthétique de la danse moderne à partir de la constitution d'un corpus littéraire, suivie de la conversion numérique des documents, du filtrage du lexique, de l'extraction de termes discriminant, de la classification automatique des documents et de l'extraction de termes

caractéristiques. Nous avons choisi le corpus constitué de deux recueils de textes de l'auteur français Stéphane Mallarmé, soit *Divagations* (édition par Eugène Fasquelle datant de 1897) et *Poésies* (édition par Émile Deman datant de 1899), pour diverses raisons : parce que l'œuvre littéraire mallarméenne est aux racines de l'art moderne et de la modernité littéraire; parce que l'œuvre littéraire mallarméenne incarne une expérience esthétique; parce que l'œuvre littéraire mallarméenne est intrinsèquement liée à la danse.

Dans le prochain chapitre, nous exposerons les résultats obtenus grâce à l'application de la méthodologie de recherche.

Chapitre 3. Résultats et discussion

3.1. Introduction

Dans ce chapitre, nous présenterons et nous analyserons les résultats obtenus grâce à l'application de la méthodologie de recherche que nous avons exposée en détail dans le chapitre précédent.

D'abord, nous présenterons et nous analyserons les descripteurs qui composent le thésaurus Collier pour atteindre le premier des objectifs de notre recherche, soit analyser un vocabulaire de description de la danse dans les archives. Ensuite, nous présenterons et nous analyserons le vocabulaire structuré de l'expérience esthétique de la danse moderne extrait d'un corpus de textes de l'auteur Stéphane Mallarmé pour atteindre le deuxième objectif de notre recherche, soit analyser un vocabulaire de représentation de la danse dans la littérature recueilli grâce à des méthodes d'extraction automatique de connaissances. Puis nous comparerons les descripteurs du thésaurus Collier et le vocabulaire structuré de l'expérience esthétique de la danse moderne en portant une attention particulière à la question de l'expérience esthétique pour atteindre le troisième objectif de notre recherche, soit comparer les deux vocabulaires afin d'en distinguer la complémentarité quant à la description de l'expérience esthétique. Nous soulèverons finalement les limites de la recherche. Nous serons alors en mesure de répondre aux deux questions qui sous-tendent notre recherche :

1. Est-ce que le vocabulaire de description de la danse dans les archives prend en compte l'expérience esthétique?
2. Est-ce que le vocabulaire de représentation de la danse dans la littérature, recueilli grâce à des méthodes d'extraction automatique de connaissances,

peut enrichir le vocabulaire de description de l'expérience esthétique de la danse dans les archives?

3.2. Objectifs de recherche

3.2.1. Premier objectif de recherche : analyse d'un vocabulaire de description de la danse dans les archives

Le premier objectif de notre projet de recherche est d'analyser un vocabulaire de description de la danse dans les archives. Pour l'atteindre, nous avons choisi d'étudier les descripteurs qui composent le thésaurus Collier.

Le thésaurus Collier compte 225 descripteurs. Comme nous l'avons précédemment mentionné, puisque les descripteurs du thésaurus Collier sont liés entre eux par des relations d'équivalence intralinguistique et par des relations d'association, mais ne sont pas liés entre eux par des relations de hiérarchie, du moins pas par des relations de hiérarchiques explicites, nous avons produit nos propres diagrammes hiérarchiques de descripteurs pour mettre en lumière certains liens hiérarchiques implicites. Nous avons empiriquement, selon notre connaissance du domaine de la danse, regroupé la majorité des descripteurs sous dix grandes catégories nommées ici suivant l'ordre alphabétique : *administration/gestion, autres arts, interprétation, diffusion, études de la danse, formation, matériel, personne(s), santé, sport*.

La catégorie qui compte le plus grand nombre de descripteurs est la catégorie *interprétation* (voir l'annexe 2). On y retrouve 64 termes regroupés principalement dans cinq sous-catégories : *chorégraphie, contexte, forme de danse, origine et style*. La sous-catégorie la plus peuplée contient les descripteurs des diverses formes de danse dont voici quelques exemples : *ballet, bharatanatyam, claquette, danse de salon, flamenco, jazz, kathak, moderne, Totentanz*.

La catégorie *autres arts* (voir l'annexe 3) regroupe 21 descripteurs, principalement des formes d'art connexes à la danse : le cirque, la comédie musicale, la *commedia dell'arte*, le kabuki, le mime, la moralité (comme genre théâtral), la musique, le noh, l'opérette, le pantomime, la revue, le théâtre et le théâtre d'ombres. Nous avons aussi inclus dans cette catégorie le descripteur *interdisciplinaire* ainsi que les termes *danse dans les arts* et *danse dans la littérature*.

La catégorie *études de la danse* (voir l'annexe 4) regroupe les descripteurs qui touchent à la recherche en danse, aux *dance studies* anglo-saxonnes ou à la *Tanzwissenschaft* allemande. On y retrouve entre autres une sous-catégorie de descripteurs à propos de disciplines de recherche telles *archéologie*, *théorie critique*, *dramaturgie*, *esthétique* (comme discipline philosophique), *histoire*, *politique* et *sociologie* ainsi qu'une sous-catégorie de descripteurs à propos d'angles de recherche ou de sujets de recherche, tels *acceptation*, *censure*, *égalité*, *folklore*, *moralité* (à ne pas confondre avec la moralité théâtrale de la catégorie *autres arts*), *religion*, *symbolisme* ou *valeur*.

La catégorie *administration/gestion* (voir l'annexe 5) concerne autant l'administration d'organisme ou de compagnie que la gestion de studio de danse. Elle regroupe 12 descripteurs. Le nom de la catégorie est duel, puisque autant le descripteur *administration* que le descripteur *gestion* sont présents dans le thésaurus (voir la figure 8 et la figure 9). Le premier concerne la gestion des organisations ou des compagnies de danse; le second concerne la gestion des studios de danse.

Figure 8. Extrait du thésaurus Collier : le descripteur *administration*

<p>ADMINISTRATION</p> <p>Note d'application: utilisé pour les objets se rapportant aux pratiques d'administration, de gestion et de comptabilité d'organisations ou de compagnies de danse.</p> <p>v/s : budget à base zéro comptabilité comptable conseil d'administration finance et comptabilité imposition (de société ou de personne) taxation (de société ou de personne) tenue des comptes</p> <p>Pour les objets se rapportant à l'administration de studios de danse privés</p> <p>UTILISER : GESTION</p>
--

Figure 9. Extrait du thésaurus Collier : le descripteur *gestion*

<p>GESTION</p> <p>Note d'application: utilisé pour les objets se rapportant à la gestion de studios de danse.</p> <p>v/s : budget à base zéro comptabilité comptable finance et comptabilité</p> <p>Pour les objets se rapportant à la gestion de compagnies de danse</p> <p>UTILISER: ADMINISTRATION</p>

La catégorie *formation* (voir l'annexe 6) contient les descripteurs en lien avec l'apprentissage de la danse comme les lieux d'enseignement tels *conservatoire*. La catégorie *matériel* (voir l'annexe 7) concerne autant le matériel nécessaire lors de représentation, par exemple *costume* ou *éclairage*, que le matériel nécessaire lors de la pratique de la danse, par exemple *plancher* ou *vêtements d'exercices*. La catégorie *diffusion* (voir l'annexe 8) concerne quant à elle autant les lieux de diffusion, le mode de diffusion, les médias que la typologie de documents de diffusion.

La catégorie *personne(s)* (voir l'annexe 9) regroupe les termes décrivant les individus ou les groupes qui gravitent dans le monde de la danse comme *agent*, *chorégraphe*, *danseur* ou *répétiteur*. Elle inclut également une sous-catégorie *âge* qui contient les descripteurs suivant : *enfant*, *adolescent*, *adulte*, *ainé*. Soulignons que le terme *enfant* concerne le groupe d'âge de 0 à 11 ans; le terme *adolescent*, le groupe d'âge de 12 à 17 ans; le terme *adulte*, le groupe d'âge de 18 à 64 ans; le terme *ainé*, le groupe d'âge de 65 ans et plus.

La catégorie *santé* (voir l'annexe 10) regroupe seulement trois descripteurs, soit *anorexie*, *blessure* et *nutrition*. La catégorie *sport* (voir l'annexe 11) regroupe quant à elle huit descripteurs, des sports liés d'une manière ou d'une autre à la danse : la callisthénie, la culture physique, la danse aérobique, la gymnastique, le patin, la rythmique, la nage synchronisée et le tai-chi-chuan.

Nous constatons ainsi que le thésaurus Collier est un outil de description essentiellement factuelle des archives de la danse. Les titres de neuf des dix catégories sont d'ailleurs à eux seuls évocateurs : *administration/gestion*, *autres arts*, *diffusion*, *études de la danse*, *formation*, *matériel*, *personne*, *santé*, *sport*. Quant à la catégorie *interprétation*, ce sont les titres de ses sous-catégories qui sont évocateurs: *chorégraphie* avec des termes comme *pas et figures*; *contexte* avec des termes comme *opéra ballet*; *forme de danse* avec des termes comme

claquette; *origine* avec des termes comme *zigane*; *style* avec des termes comme *précision*.

3.2.2. Deuxième objectif de recherche : analyse d'un vocabulaire de représentation de la danse dans la littérature

Le deuxième objectif de notre projet de recherche est d'analyser un vocabulaire de représentation de la danse dans la littérature recueilli grâce à des méthodes d'extraction automatique de connaissances. Pour l'atteindre, nous avons extrait un vocabulaire structuré de l'expérience esthétique de la danse moderne à partir d'un corpus de textes. Nous avons ensuite étudié le vocabulaire obtenu.

Comme nous l'avons déjà mentionné au chapitre précédent, le corpus de textes se compose des textes contenus dans les recueils *Divagations* (1897) et *Poésies* (1899) de l'auteur français Stéphane Mallarmé. Le corpus comprend ainsi 119 documents et, avant le filtrage du lexique, 11 850 formes sur 70 507 occurrences. Après le filtrage du lexique, nous avons obtenu un corpus de 7 238 formes sur 70 070 occurrences. Ainsi, 437 (soit 70 507 - 70 070) occurrences ont été retirées, car elles se trouvaient dans l'antidictionnaire et 4612 formes (soit 11 850 - 7 238) ont été lemmatisées.

Nous avons alors appliqué la méthode d'extraction de termes discriminants et la méthode de classification hiérarchique agglomérative que nous avons exposées en détail au chapitre précédent. Nous avons ainsi obtenu une classification de sept classes. Une première classe compte 73 documents; une deuxième classe, 34 documents; une troisième classe, 2 documents; une quatrième classe, 3 documents; une cinquième classe, 2 documents; une sixième classe, 4 documents; une septième et dernière classe, un seul document.

Après analyse de la fréquence, du nombre de cas et du *tf.idf* des termes contenus dans chacune des cinq classes de taille plus restreinte, nous avons

choisi d'éliminer ces classes de l'expérimentation. Le trop petit nombre de documents rendait les variations de fréquence et du nombre de cas entre les mots non significatives, donc quasi impossible l'extraction de termes caractéristiques. Nous pouvons l'observer dans le tableau 3, le poids informatif des termes contenus dans la classe incluant les documents *Hommage, Salut, La femme de l'ouvrier et Le crieur d'imprimés* – la plus grande des classes de taille restreinte – est pratiquement constant, aucun des termes ne peut donc être considéré comme caractéristique.

Tableau 3. Vingt termes extraits d'une classe de taille restreinte

	Fréquence	Nombre de cas	<i>tf.idf</i>
COUPE	3	2	0,9
CRIER	2	2	0,6
IMPORTER	2	2	0,6
POUPE	2	2	0,6
SALUT	2	2	0,6
SOUCI	2	2	0,6
AMI	1	1	0,6
ANNONCE	1	1	0,6
BARRE	1	1	0,6
BAS	1	1	0,6
BELLE	1	1	0,6
BLANC	1	1	0,6
CAP	1	1	0,6
CARAVELLE	1	1	0,6
CARRIER	1	1	0,6
CHANT	1	1	0,6
CHEMIN	1	1	0,6
COMPLIMENTER	1	1	0,6

CRAINDRE	1	1	0,6
DOUBLE	1	1	0,6

Nous avons par la suite porté toute notre attention sur les deux classes de plus grande taille. De plus, lorsque nous regroupons les documents contenus dans les deux classes de plus grande taille, créant ainsi un nouveau regroupement de 107 documents, le terme ayant le poids informatif le plus élevé est *art*.

De l'analyse du poids informatif des termes contenus dans la classe comptant 34 documents, nous avons d'abord pu constater qu'il n'était pas justifié de rescinder la classe en sous-classes, puisque nous obtenions alors une grande sous-classe et quelques sous-classes de taille restreinte composées de termes dont la variation du poids informatif n'était pas significative. Nous avons ensuite listé les 17 termes de la classe de 34 documents au poids informatif le plus élevé, au poids informatif égal ou supérieur à 5,0 : *vieux, savoir, nu, voix, pur, blanc, aller, enfant, fleur, rire, rêve, rose, sein, œil, grand, jour, seul*.

Tableau 4. Dix-sept termes au poids informatif le plus élevé

	Fréquence	Nombre de cas	<i>tf.idf</i>
VIEUX	8	5	6,7
SAVOIR	12	10	6,4
NU	12	10	6,4
VOIX	7	5	5,8
PUR	9	8	5,7
BLANC	11	11	5,4
ALLER	5	3	5,3
ENFANT	7	6	5,3
FLEUR	7	6	5,3

RIRE	7	6	5,3
RÊVE	7	6	5,3
ROSE	9	9	5,2
SEIN	6	5	5
OEIL	6	5	5
GRAND	6	5	5
JOUR	6	5	5
SEUL	8	8	5

Nous avons alors éliminé de la liste certains de ces termes : *savoir* et *aller*. Le terme *savoir* était ambigu. S'agit-il ici du verbe ou du nom commun? La fonction de visualisation des mots-clé en contexte (*key word in context*) nous apprend qu'il s'agit en fait de diverses déclinaisons de la conjugaison du verbe *savoir* qui ont été lemmatisés lors du prétraitement (*sait, sais, sache, savent, saura*). Nous avons ainsi jugé qu'il s'agissait d'un mot vide. Nous avons éliminé le terme *aller* pour une raison identique.

Nous avons observé une homonymie entre *nue*, adjectif décrivant la nudité, et *nue*, nom féminin signifiant *nuages*. Nous avons conservé le terme *nu(e)*, puisque *nu* comme dans *nudité* est le sens majoritaire et le *tf.idf* ajusté reste au-dessus de 5,0. Citons en exemple la nudité de la nixe – une nymphe des eaux – dans le dernier tercet de l'un des sonnets les plus célèbres de Stéphane Mallarmé (1899), le sonnet en *-yx* :

Ses purs ongles très haut dédiant leur onyx,
L'Angoisse, ce minuit, soutient, lampadophore,
Main rêve vespéral brûlé par le Phénix
Que ne recueille pas de cinéraire amphore
Sur les crédences, au salon vide : nul ptyx,
Aboli bibelot d'inanité sonore,
(Car le Maître est allé puiser des pleurs au Styx
Avec ce seul objet dont le Néant s'honore).

Mais proche la croisée au nord vacante, un or
 Agonise selon peut-être le décor
 Des licornes ruant du feu contre une nixe,
 Elle, défunte **nue** en le miroir, encor
 Que, dans l'oubli fermé par le cadre, se fixe
 De scintillations sitôt le septuor.

Nous avons également observé une homonymie entre le terme *rose* comme nom de fleur et le terme *rose* comme adjectif de couleur. Encore ici, nous avons conservé *rose*, car son emploi comme adjectif ne concerne qu'une seule utilisation sur un total de neuf utilisations. Le *tf.idf* reste ainsi au-dessus de 5,0.

Selon Birger Hjørland (2008 : 89), le principe à la base d'un vocabulaire contrôlé est d'éviter les synonymes. Ainsi, puisque plusieurs mots nous semblaient être métaphoriquement fortement liés, nous avons regroupé les mots de la liste selon le champ lexical auquel ils se rattachent pour obtenir les principaux thèmes de l'expérience esthétique contenus dans le corpus de textes. Soulignons que nous avons puisé le terme représentant chaque champ lexical à même les mots du corpus; ce sont toutefois des termes au poids informatif moindre. Nous avons ainsi obtenu les neuf champs lexicaux, les neuf thèmes de l'expérience esthétique suivants : *corps*, *idéal*, *ironie*, *nature*, *nudité*, *voix*, *pureté*, *solitude*, *temps*.

Tableau 5. Neuf thèmes de l'expérience esthétique

Termes	Champ lexical
SEIN	corps
ŒIL	
RÊVE	idéal
GRAND	
RIRE	rire

FLEUR	nature
ROSE	
NU	nudité
VOIX	voix
PUR	pureté
BLANC	
SEUL	solitude
VIEUX	temps
ENFANT	
JOUR	

Ainsi, nous avons regroupé *sein* et *œil* sous le champ lexical du corps. Voici un extrait du corpus de textes (Mallarmé 1897 : *Anecdotes ou poèmes – Le phénomène futur*) qui justifie le regroupement :

À la place du vêtement vain, elle a un **corps** ; et les **yeux**, semblables aux pierres rares ! ne valent pas ce regard qui sort de sa chair heureuse : des **seins** levés comme s'ils étaient pleins d'un lait éternel, la pointe vers le ciel, aux jambes lisses qui gardent le sel de la mer première.

Nous avons regroupé *rêve* et *grand* sous le champ lexical de l'idéal. La grandeur et le rêve sont parfois liés dans l'oeuvre de Stéphane Mallarmé comme dans l'extrait suivant où il mentionne les « instincts tout de **grandeur** et de beauté déjà, du **rêve** latent » (Mallarmé 1897 : *Quelques médaillons ou portraits en pied – Beckford*). Et le champ lexical de l'idéal est associé entre autres à la danse : « le corps de ballet, total ne figurera autour de l'étoile (la peut-on mieux nommer !) la danse **idéale** des constellations » (Mallarmé 1897 : *Crayonné au théâtre – Ballets*). Le terme *étoile* en ballet désigne le grade de danseur le plus élevé dans une compagnie.

Nous avons tout simplement associé *rire* au champ lexical du rire. Le rire dans l'œuvre de Stéphane Mallarmé est même parfois directement celui du public : « quelque possibilité de rire, abondamment, avec une salle entière » (Mallarmé 1897 : *Crayonné au théâtre – Planches et feuillets*).

Nous avons regroupé *fleur* et *rose* sous le champ lexical plus large de la nature. Soulignons que le lien entre la nature et l'expérience esthétique est explicite dans l'œuvre de Stéphane Mallarmé. Ainsi, « [l]oin de tout, la Nature, en automne, prépare son Théâtre, sublime et pur » (Mallarmé 1897 : *Crayonné au théâtre – Hamlet*).

Nous avons associé *nu* au champ lexical de la nudité, bien que nous aurions aussi pu l'associer au champ lexical de la pureté. Le thème de la nudité nous semblait toutefois essentiel à la description de la danse moderne. D'ailleurs, voici un extrait de *Autre étude de danse* où Stéphane Mallarmé (1897 : *Crayonné au théâtre – Autre étude de danse*) discute de Loïe Fuller, une des pionnières de la danse moderne, qui se produisait relativement dénudée et entourée de voiles :

Le poète, par une page riche et subtile, a, du coup, restitué à l'antique fonction son caractère, qu'elle s'étoffe ; et, sans retard, invoque la Loïe Fuller, fontaine intarissable d'elle-même – près le développement de qui ou les trames imaginatives versées comme atmosphère, les coryphées du Ballet, court-vêtues à l'excès, manquent d'ambiance sauf l'orchestre et n'était que le costume simplifié, à jamais, pour une spirituelle acrobatie ordonnant de suivre la moindre intention scripturale, existe, mais invisible, dans le mouvement pur et le silence déplacé par la voltige. La presque **nudité**, à part un rayonnement bref de jupe, soit pour amortir la chute ou, à l'inverse, hausser l'enlèvement des pointes, montre, pour tout, les jambes – sous quelque signification autre que personnelle, comme un instrument direct d'idée.

Nous avons naturellement associé *voix* au champ lexical de la voix. À propos de la voix, Stéphane Mallarmé (1897 : *Anecdotes ou poèmes – Un spectacle interrompu*) s’interroge d’ailleurs ainsi : « explique-moi la vertu de cette atmosphère de splendeur, de poussière et de **voix**, où tu m’appris à me mouvoir ».

Nous avons regroupé *pur* et *blanc* sous le champ lexical de la pureté, même si la pureté mallarméenne incarne parfois autant plus la mélancolie que la pureté. Lorsque l’écrivain parle, par exemple, de « blanche agonie » (Mallarmé 1899 – *Sonnet*). Le thème de la pureté nous semblait primordial dans l’expérience esthétique de la danse moderne de Stéphane Mallarmé; l’écrivain (1897 : *Crayonné au théâtre – Autre étude de danse*) parle d’ailleurs de « mouvement pur » en qualifiant Loïe Fuller.

Nous avons associé *seul* au champ lexical de la solitude. Stéphane Mallarmé (1899 : *Hérodias*) fait dire à Hérodias, dans le dialogue entre elle et la nourrice – dont Martha Graham s’est inspirée pour créer la chorégraphie *Herodiade* au milieu des années 1940 (Bannerman 2006) – que « [ô] charme dernier, oui! je le sens, je suis seule ».

Nous avons finalement regroupé *vieux*, *enfant* et *jour* sous le champ lexical du temps. Nous avons associé *jour* à *temps* puisque, par exemple, Stéphane Mallarmé (1899 – *Toast funèbre*) écrit : « parmi l’heure et le rayon du jour ». Nous avons aussi associé *vieux* et *enfant* à *temps* à cause de l’idée de passage du temps que sous-tendent ces deux termes. À propos d’un des personnages du ballet *Les deux pigeons*, Stéphane Mallarmé (1897 : *Crayonné au théâtre – Ballets*) écrit ainsi : « homme déjà et enfant encore ». Il associe aussi *enfant* avec le temps qui passe, avec la résurrection (« Anastase »), l’éternité (« éternels »), la vieillesse (« aïeul ») et la mort (« sépulcre ») comme dans cet extrait de *Prose pour des Esseintes* (Mallarmé 1899) :

L'**enfant** abdique son extase
 Et docte déjà par chemins
 Elle dit le mot : **Anastase** !
 Né pour d'**éternels** parchemins,
 Avant qu'un **sépulcre** ne rie
 Sous aucun climat, son **aïeul**,
 De porter ce nom : Pulchérie !
 Caché par le trop grand glaïeul.

Discutons maintenant de l'analyse du poids informatif des termes contenus dans la classe comptant 73 documents (voir l'annexe 12). En émerge un constat : les quatre termes au poids informatif le plus élevé sont *théâtre, art, danse, scène*.

Tableau 6. Quatre termes au poids informatif le plus élevé de la classe de 73 documents

	Fréquence	Nombre de cas	<i>tf.idf</i>
THÉÂTRE	51	18	31
ART	85	36	26,1
DANSE	27	8	25,9
SCÈNE	44	19	25,7

Figure 10. Nuage des quatre termes au poids informatif le plus élevé de la classe de 73 documents



Nous avons donc eu la chance de vérifier la réalité des liens tissés entre l'œuvre de Stéphane Mallarmé, la danse, en particulier, et les arts d'interprétation, en général. D'ailleurs, à propos du terme *danse*, soulignons que, lorsque le terme *ballet* – un terme utilisé comme quasi-synonyme au terme *danse* par Stéphane Mallarmé –, le terme *ballerine* et le terme *danseuse* sont substitués par *danse* lors du filtrage du lexique, *danse* devient alors le terme ayant le poids informatif le plus élevé. Il est également important de souligner que le courant du *Tanztheater* – *Tanz* signifie *danse* et *Theater* signifie *théâtre* en allemand – est un des grands courants de la danse moderne. Ainsi, puisque la classe comptant 73 documents nous semblait fortement liée à la danse, nous y avons puisé des qualificatifs de l'expérience esthétique de la danse moderne.

Nous avons d'abord scindé la classe en sous-classes en nous basant sur le dendrogramme de classification du corpus de textes. Nous avons ainsi obtenu une classification de 14 classes dont 6 singletons. Nous n'avons alors retenu que les quatre classes qui comportaient plus de trois documents. Nous nous sommes ainsi retrouvés avec une classe de 40 documents, une classe de 11 documents, une classe de 5 documents et une classe de 3 documents. Nous avons extrait les termes de chacune des classes et nous avons sélectionné l'adjectif qualificatif au poids informatif le plus élevé de chacune d'elle (voir les tableaux 7, 8, 9 et 10). Puis nous avons cherché dans l'ensemble du corpus l'antonyme de chacun des qualificatifs trouvés pour ainsi obtenir trois paires – deux des classes partageant le même qualificatif au poids informatif le plus élevé – de qualités de l'expérience esthétique. Le poids informatif beaucoup plus faible des antonymes trouvés n'est pas un obstacle à leur utilisation.

À la paire formée des termes *seul* et *en troupe*, nous avons également ajouté un troisième terme *en couple*, puisque la danse en duo, comme le traditionnel pas de deux au ballet, est un élément qui nous semblait essentiel à la description du domaine. D'ailleurs, le terme *couple* est utilisé entre autres par

Stéphane Mallarmé (1897 : *Crayonné au théâtre – Ballets*) pour décrire un duo de danseurs.

Tableau 7. Vingt-quatre termes au poids informatif le plus élevé de la sous-classe de quarante documents

	Fréquence	Nombre de cas	<i>tf.idf</i>
DANSE	27	8	18,9
THÉÂTRE	47	17	17,5
LIVRE	48	20	14,4
BALLET	16	5	14,4
SCÈNE	41	18	14,2
AN	30	14	13,7
LIVRE	48	21	13,4
MUSIQUE	36	17	13,4
ÉCRIRE	38	18	13,2
VILLIERS	10	2	13
SIÈCLE	30	15	12,8
FRANÇAIS	31	16	12,3
JOURNAL	19	10	11,4
TENNYSON	7	1	11,2
VERLAINE	10	3	11,2
RIMBAUD	7	1	11,2
PARLER	28	16	11,1
POÈME	24	14	10,9
DEVENIR	24	14	10,9
VOLUME	18	10	10,8
PRENDRE	31	18	10,8
PIÈCE	19	11	10,7
BEAU	22	13	10,7

VIEUX	22	13	10,7
-------	----	----	------

Le premier adjectif qualificatif trouvé dans la liste des termes de la sous-classe de quarante documents est *français*. Nous l'avons exclu, car il est utilisé comme adjectif et comme nom. Viennent ainsi en second lieu *beau* et *vieux*. Nous avons choisi de garder seulement l'adjectif *beau*, car en additionnant aux statistiques de *beau(x)*, celles de *belle(s)* et de *beauté*, nous avons obtenu un *tf.idf* de 12,5. On ne peut en faire autant avec le terme *vieux*.

Tableau 8. Sept termes au poids informatif le plus élevé de la sous-classe de onze documents

	Fréquence	Nombre de cas	<i>tf.idf</i>
HAMLET	12	2	4,8
THÉÂTRE	4	1	2,8
SEUL	6	2	2,4
SHAKESPEARE	3	1	2,1
POUVOIR	3	1	2,1
PIÈCE	3	1	2,1
CONTEMPORAIN	3	1	2,1

Le premier adjectif qualificatif trouvé dans la liste des termes de la sous-classe de onze documents est *seul*. Il est intéressant de noter que l'adjectif suivant est *contemporain* comme dans *danse contemporaine*.

Tableau 9. Neuf termes au poids informatif le plus élevé de la sous-classe de cinq documents

	Fréquence	Nombre de cas	<i>tf.idf</i>
VOIR	14	5	4,8
OMBRE	6	2	4,4
IGNORER	4	1	4,2
MIROIR	4	1	4,2

SEUL	13	6	3,4
ROSE	6	3	3,4
JOUR	6	3	3,4
EAU	7	4	3,1
SOLITAIRE	3	1	3,1

Le premier adjectif qualificatif trouvé dans la liste des termes de la sous-classe de cinq documents est encore *seul*. Il est intéressant de noter que l'adjectif suivant est *solitaire*; le choix de *seul* s'impose ainsi d'autant plus.

Tableau 10. Quinze termes au poids informatif le plus élevé de la sous-classe de trois documents

	Fréquence	Nombre de cas	<i>tf.idf</i>
PETIT	5	1	2,4
TÊTE	4	1	1,9
HIVER	3	1	1,4
JOURNÉE	3	1	1,4
CRIME	3	1	1,4
CHARBON	3	1	1,4
PIPE	2	1	1
TRAVAIL	2	1	1
POUSSIÈRE	2	1	1
RUE	2	1	1
SAVOIR	2	1	1
MARIER	2	1	1
ORGUE	2	1	1
LONDRES	2	1	1
LATIN	2	1	1

Le premier adjectif qualificatif trouvé dans la liste des termes de la sous-classe de trois documents est *petit*. Bien qu'il s'agisse ici d'une classe de taille

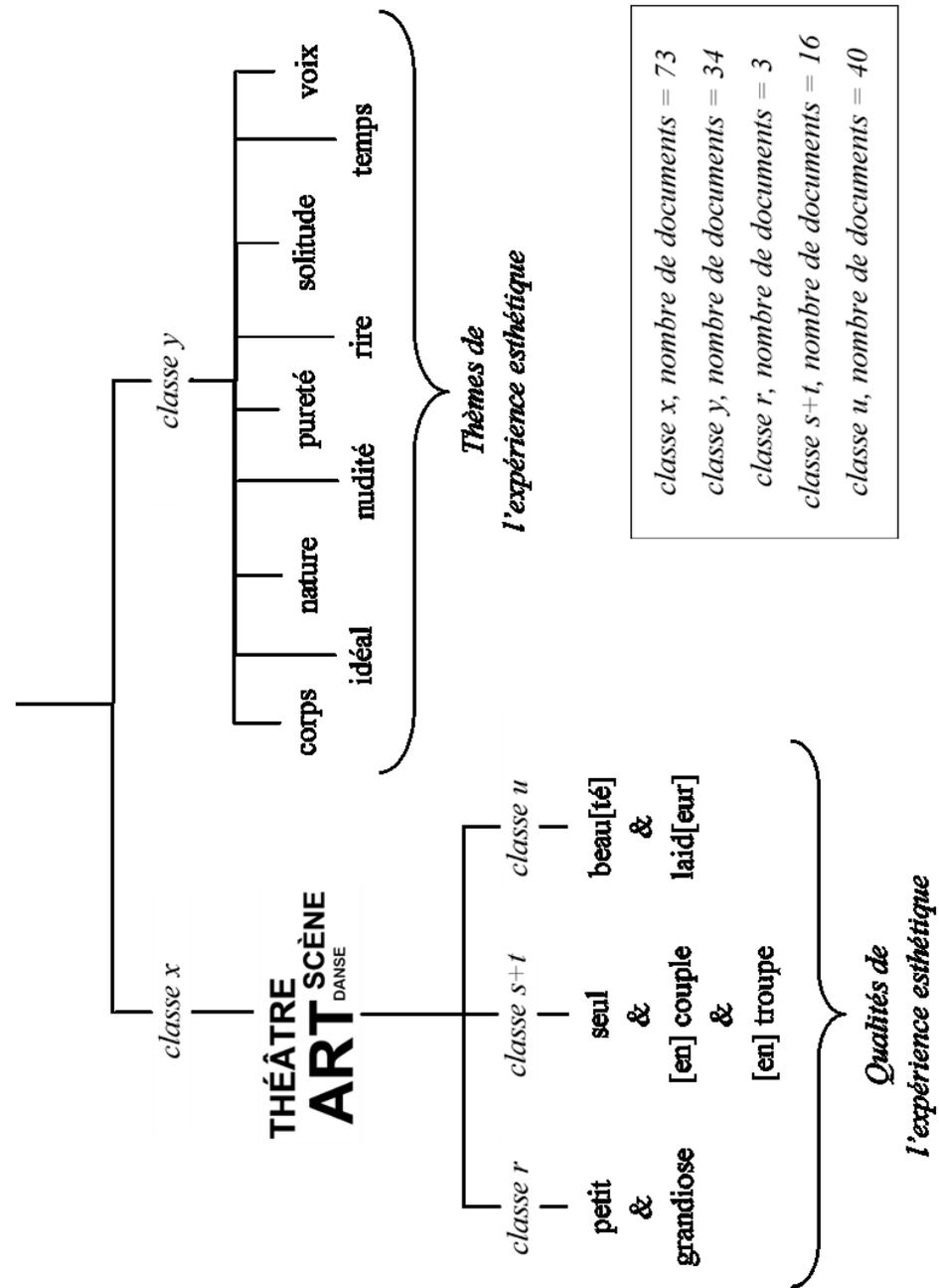
restreinte, nous avons choisi de l'inclure malgré tout dans notre analyse, car le poids informatif de *petit* se démarque. L'adjectif qualificatif suivant est *latin* et il ne se démarque nullement quant à son poids informatif.

Comme nous l'avons déjà mentionné, nous avons ensuite cherché dans l'ensemble du corpus l'antonyme de chacun des qualificatifs trouvés – beau, seul, petit – pour ainsi obtenir deux paires et une triade de qualités de l'expérience esthétique : *petit/grandiose*, *beau[té]/laid[deur]*, *seul/en couple/en troupe*.

Le choix de *grandiose* s'est imposé comme antonyme de *petit*, car il est mieux adapté à la description de spectacles de danse que le terme *grand*, qui était l'antonyme commun de *petit*. Le choix de *en troupe* comme antonyme de *seul* s'est imposé, car le terme *troupe* est utilisé dans un des textes du corpus qui traite du théâtre (Mallarmé 1897 : *Crayonné au théâtre – Hamlet*). Le choix de *laid* comme antonyme de *beau* s'est imposé à cause du contexte d'utilisation du terme *laideur* dans *Morceau pour résumer Vathek* (Mallarmé 1897 : *Volumes sur le divan*) : « les moyens méconnus autrefois de l'art de peindre, tels qu'accumulation d'étrangetés produite simplement pour leur caractère unique ou de laideur ».

À la fin du processus, nous avons ainsi pu identifier et structurer quelques termes principaux décrivant l'expérience esthétique de la danse moderne d'après le corpus constitué d'œuvres de Stéphane Mallarmé et alors esquisser un vocabulaire structuré de l'expérience esthétique de la danse moderne (voir la figure 11).

Figure 11. Vocabulaire structuré de l'expérience esthétique de la danse moderne extrait d'un corpus de textes de Stéphane Mallarmé



Appliquons maintenant le vocabulaire à la description de la captation d'une chorégraphie. Prenons par exemple une captation de la version complète de la reconstitution du *Sacre du printemps*, une chorégraphie de Vaslav Nijinski créée en 1913 sur une musique d'Igor Stravinski, comme celle réalisée avec la troupe de ballet du Théâtre Mariinski par Denis Caïozzi (2008). Nous pouvons la décrire en lui associant les termes *grandiose*, *troupe* et *nature*. Le terme *grandiose* permet de souligner la scénographie élaborée; comparons-la par exemple à une chorégraphie à la scénographie toute en retenue comme le solo *Torn Roots, Broken Branches* de Margie Gillis. Le terme *troupe* indique que les danseurs – jusqu'à 46 danseurs – sont toujours en troupe sur scène, même lors du solo final où les danseurs s'abaissent au pied de la soliste, Alexandra Iosifidi, sans quitter la scène. Le terme *nature* permet de souligner quant à lui l'importance de la nature dans la scénographie; les décors y représentent une pleine nature, les collines de la Russie natale de Vaslav Nijinski peut-être. Igor Stravinski affirme lui-même qu'il a « voulu exprimer dans *Le Sacre du printemps* la sublime montée de la nature qui se renouvelle » (Beltrando-Patier et coll. 2004 : 962).

D'autres exemples sont possibles. Pensons à la chorégraphie contemporaine *Un peu de tendresse, bordel de merde!* de Dave St-Pierre qui présente « 17 danseurs et performeurs », « [d]ix-sept corps, souvent nus (costume préféré du jeune chorégraphe) » et « une maîtresse de cérémonie omnisciente, au cynisme aguicheur » (Doyon 2006). Les termes *troupe*, *nudité*, *rire* et *voix* peuvent bien la décrire.

Les termes qui composent le vocabulaire structuré de l'expérience esthétique de la danse moderne extrait d'un corpus de textes de Stéphane Mallarmé permettent ainsi une certaine description de l'expérience esthétique de la danse en proposant des thèmes ainsi que des adjectifs liés à l'expérience esthétique de la danse. Le vocabulaire comprend des termes comme *petit* et

grandiose pour décrire l'ampleur du spectacle; des termes comme *seul*, *couple*, *troupe* pour décrire le nombre de danseurs présents sur scène; des termes comme *beau* et *laid* pour décrire l'esthétique générale de la performance. Les termes *corps*, *idéal*, *nature*, *nudité*, *pureté*, *rire*, *solitude*, *temps* et *voix* permettent de décrire ce qui se passe sur scène au-delà de la trame narrative. Dans le *Sacre du printemps* de Vaslav Nijinski, comme dans de nombreuses chorégraphies en danse moderne, la trame narrative est de toute façon pratiquement inexistante.

3.2.3. Troisième objectif de recherche : comparaison entre le vocabulaire de description et le vocabulaire de représentation

Notre troisième objectif de recherche est de comparer un vocabulaire de description de la danse dans les archives et un vocabulaire de représentation de la danse dans la littérature afin d'en distinguer la complémentarité quant à la description de l'expérience esthétique, comparons ainsi maintenant les descripteurs du thésaurus Collier et le vocabulaire structuré de l'expérience esthétique de la danse moderne extrait d'un corpus de textes de Stéphane Mallarmé.

Nous venons de le voir, le thésaurus Collier propose des descripteurs plus factuels et le vocabulaire structuré de l'expérience esthétique de la danse moderne extrait d'un corpus de textes de l'auteur Stéphane Mallarmé propose des termes qui décrivent quelque chose de plus fugace. Notons également que les termes des deux vocabulaires ne se recoupent pas, l'ensemble des termes de l'un et de l'autre sont disjoints. Rappelons toutefois qu'un des termes originels extraits du corpus de textes, le terme *enfant*, associé au champ lexical du temps, se retrouve aussi dans la liste des descripteurs du thésaurus Collier. Le sens que prend *enfant* dans les deux cas est cependant différent. Il incarne même à lui seul la différence intrinsèque entre les deux vocabulaires, l'un plus factuel,

l'autre plus fugace. Dans le thésaurus Collier, le terme permet une description factuelle. Le terme *enfant* y est « utilisé pour les objets pour lesquels l'âge est un facteur important, tel que les programmes ou grades de danse entre autres, et qui devrait être noté » et il concerne uniquement les personnes âgées de 0 à 11 ans, alors que l'utilisation de *enfant* dans les textes de Stéphane Mallarmé est hautement métaphorique. Par exemple, à propos du couple de danseurs du ballet *Les deux pigeons*, Stéphane Mallarmé (1897 : *Crayonné au théâtre – Ballets*) écrit : « **[e]nfants**, les voici oiseaux, ou le contraire, d'oiseaux **enfants**, selon qu'on veut comprendre l'échange dont toujours et dès lors, lui et elle, devraient exprimer le double jeu ».

Ainsi, suite à la comparaison du vocabulaire de description de la danse dans les archives qu'est le thésaurus Collier et du vocabulaire de représentation de la danse dans la littérature qu'est le vocabulaire structuré de l'expérience esthétique de la danse moderne extrait d'un corpus de textes de l'auteur Stéphane Mallarmé, nous pouvons constater que le vocabulaire structuré de l'expérience esthétique de la danse moderne d'après Stéphane Mallarmé est jusqu'à un certain point complémentaire à un outil de description des archives de la danse comme le thésaurus de descripteurs Collier.

3.3. Limites de la recherche

3.3.1. Limites relatives aux archives de la danse

Nous l'avons déjà mentionné, les pratiques de documentation de la performance sont relativement éclatées et peu explicitées (Evens et Hautekeete 2011: 162). Une des limites relatives aux archives de la danse est ainsi le manque de documentation.

3.3.2. Limites relatives à l'extraction automatique de connaissances

Trois limites principales quant à l'extraction automatique de connaissances nous sont apparues lors de l'expérimentation : celles touchant la lemmatisation, celles touchant l'opérationnalisation de la démarche et celles touchant la validation des résultats.

Ainsi, nous nous sommes heurtés à certaines limites de la lemmatisation lors du filtrage du lexique. La lemmatisation incluse dans le logiciel WordStat était très partielle. Par exemple, le terme *beau* ne remplaçait pas *belle(s)* et *beauté*. Toutefois, peu importe le logiciel, la lemmatisation n'est jamais une opération simple (Brunet 2000). Ensuite, nous nous sommes heurtés à certaines limites de l'utilisation d'un corpus restreint. Rappelons-le, le corpus de textes ne comprenait que 119 documents et 11 850 mots distincts (avant le prétraitement avec l'antidictionnaire). Il s'agit d'un petit corpus en terme de fouille de textes. Nous nous sommes aussi heurtés à certaines limites à cause de l'unité du vocabulaire extrait du corpus de textes : seulement 11 850 termes uniques sur 70 507 termes au total. L'unité terminologique et thématique dans l'oeuvre de Stéphane Mallarmé semble ainsi très grande; cela est une force – des thèmes récurrents peuvent être dégagés – autant qu'une faiblesse – la classification des documents est plus difficile puisque les termes réellement discriminants peuvent plus difficilement être extraits. Finalement, nous nous sommes heurtés à certaines limites étant donné le caractère disparate de la taille des documents du corpus : certains documents étaient un texte en prose ou un dialogue, généralement plus long, et d'autres étaient un poème, généralement plus court. L'utilisation du *tf.idf* plutôt que de la simple fréquence comme mesure du poids informatif des termes prend alors tout son sens.

Nous nous sommes également heurtés à certaines limites dans l'opérationnalisation de la démarche. Notre recherche est exploratoire, nous

avons ainsi obtenu une preuve du concept, mais les manipulations humaines nécessaires à l'obtention de la preuve sont relativement élevées. Une optimisation de la démarche, qui inclut une diminution des manipulations par automatisation de certaines d'entre elles, est nécessaire avant de l'appliquer à un corpus de textes plus imposant.

Nous nous sommes finalement heurtés à certaines limites quant à la validation de la démarche proposée. Lors de la réalisation de notre projet de recherche, nous n'avons pas appliqué en pratique les résultats obtenus. Nous n'avons pas développé ou enrichi un thésaurus de l'expérience esthétique de la danse, pas plus que nous avons appliqué cet hypothétique thésaurus à la description de documents dans des archives de la danse. Toutefois, tel n'était pas l'objectif de nos travaux de recherche. Il s'agissait plutôt d'effectuer la première expérimentation d'une nouvelle démarche de génération automatique de vocabulaire de description.

3.4. Questions de recherche

3.4.1. Première question de recherche

La première question de recherche s'énonce ainsi : est-ce que le vocabulaire de description de la danse dans les archives prend en compte l'expérience esthétique?

Suite à l'analyse d'un vocabulaire de description de la danse dans les archives, nous pouvons remarquer que le thésaurus de descripteurs Collier est un outil de description essentiellement factuel. Nous pouvons alors affirmer que le cas particulier de vocabulaire de description de la danse dans les archives qu'est le thésaurus de descripteurs Collier ne prend pas en compte l'expérience esthétique et ainsi répondre à la première question de recherche, du moins en

partie, puisque nous n'avons analysé qu'un vocabulaire de description parmi d'autres.

Rappelons toutefois que la documentation restreinte sur le sujet de la description des archives de la danse ne nous a pas permis de généraliser notre affirmation. Un état des lieux à propos de la description de documents dans les archives de la danse reste encore à faire.

3.4.2. Seconde question de recherche

La seconde question de recherche s'énonce ainsi : est-ce que le vocabulaire de représentation de la danse dans la littérature, recueilli grâce à des méthodes d'extraction automatique de connaissances, peut enrichir le vocabulaire de description de l'expérience esthétique de la danse dans les archives?

Suite à l'analyse d'un vocabulaire de représentation de la danse dans la littérature ainsi qu'à la comparaison d'un vocabulaire de description de la danse dans les archives et du vocabulaire de représentation, nous pouvons affirmer que le vocabulaire structuré de l'expérience esthétique de la danse moderne extrait d'un corpus de textes de l'auteur Stéphane Mallarmé permet effectivement une certaine description de l'expérience esthétique de la danse moderne. Il propose des thèmes ainsi que des adjectifs liés à l'expérience esthétique de la danse moderne. Il est alors jusqu'à un certain point complémentaire à un outil de description des archives de la danse comme le thésaurus de descripteurs Collier.

Rappelons toutefois que les limites relatives à l'extraction automatique de connaissances (lemmatisation imparfaite, corpus restreint, unité du vocabulaire et des thèmes du corpus, variations dans la taille des documents du corpus) ont pu introduire certains biais dans les termes qui composent le vocabulaire structuré de l'expérience esthétique de la danse moderne.

Rappelons également que l'absence d'application pratique, d'application à un cas réel de description dans les archives de la danse ne nous permet pas d'affirmer autrement que théoriquement la complémentarité entre les vocabulaires. Nous pouvons toutefois affirmer que de plus amples expérimentations méritent d'être menées.

3.5. Conclusion

Dans ce chapitre, nous avons présenté et analysé les résultats obtenus grâce à l'application de la méthodologie de recherche. Nous avons ainsi atteint les objectifs de notre projet de recherche et répondu, du moins en partie, aux questions de recherche.

Premièrement, nous avons présenté et analysé les descripteurs qui composent le thésaurus Collier pour atteindre le premier des objectifs de notre recherche, soit analyser un vocabulaire de description de la danse dans les archives. Deuxièmement, nous avons présenté et analysé le vocabulaire structuré de l'expérience esthétique de la danse moderne extrait d'un corpus de textes de l'auteur Stéphane Mallarmé pour atteindre le deuxième objectif de notre recherche, soit analyser un vocabulaire de représentation de la danse dans la littérature, dans ce cas-ci un vocabulaire de représentation de l'expérience esthétique de la danse moderne, recueilli grâce à des méthodes d'extraction automatique de connaissances. Troisièmement, nous avons comparé les descripteurs du thésaurus Collier et le vocabulaire structuré de l'expérience esthétique de la danse moderne en portant une attention particulière à la question de l'expérience esthétique pour atteindre le troisième objectif de notre recherche, soit comparer les deux vocabulaires afin d'en distinguer la complémentarité quant à la description de l'expérience esthétique. Finalement, nous avons soulevé les limites de la recherche; les limites relatives aux archives de la danse (documentation restreinte à ce propos) autant que les limites

relatives à l'extraction automatique de connaissances (lemmatisation imparfaite, corpus restreint, unité du vocabulaire et des thèmes du corpus, variations dans la taille des documents du corpus, opérationnalisation de la démarche à perfectionner, validation des résultats à effectuer).

Nous avons alors été en mesure de répondre aux deux questions qui sous-tendent notre recherche. Nous avons pu répondre, du moins en partie, à la première question de recherche, où nous nous interrogeons à savoir si le vocabulaire de description de la danse dans les archives prend en compte l'expérience esthétique. Nous avons conclu que le thésaurus de descripteurs Collier est un outil de description essentiellement factuelle qui ne prend pas en compte l'expérience esthétique. Nous avons pu répondre, du moins en partie, à la seconde question de recherche, où nous nous interrogeons à savoir si le vocabulaire de représentation de la danse dans la littérature, recueilli grâce à des méthodes d'extraction automatique de connaissances, peut enrichir le vocabulaire de description de l'expérience esthétique de la danse dans les archives. Le vocabulaire structuré de l'expérience esthétique de la danse moderne extrait d'un corpus de textes de l'auteur Stéphane Mallarmé permet effectivement une certaine description de l'expérience esthétique de la danse moderne. Il propose des thèmes ainsi que des adjectifs liés à l'expérience esthétique de la danse moderne. Il est alors jusqu'à un certain point complémentaire à un outil de description des archives de la danse comme le thésaurus de descripteurs Collier.

Dans le prochain chapitre, nous concluons en soulignant les grandes lignes du mémoire, en explicitant notre contribution au domaine des sciences de l'information et en évoquant de futures recherches .

Conclusion

Dans ce chapitre, nous reviendrons sur les grandes lignes de notre recherche. Nous mettrons ensuite en exergue les contributions autant théoriques, méthodologiques que pratiques de nos travaux au domaine des sciences de l'information. Finalement, nous évoquerons des recherches futures qui peuvent découler de notre mémoire.

1. Résumé de la recherche

Dans l'introduction, nous avons mentionné que la danse est maintenant un sujet d'études et de recherche à part entière. Nous avons aussi exposé plusieurs raisons qui nous poussent à croire qu'il est nécessaire de mieux décrire la performance en danse dans les archives. Nous avons également souligné que le monde de l'art, de la danse, et le monde des sciences de l'information, de la documentation de la danse, ont tout intérêt à s'inspirer un de l'autre.

L'objectif général de notre recherche était alors de contribuer au développement d'outils de gestion de l'information en comparant le vocabulaire de description de la danse dans les archives et le vocabulaire de représentation de la danse dans la littérature pour en distinguer une possible complémentarité, particulièrement en ce qui a trait au vocabulaire de l'expérience esthétique. Trois objectifs spécifiques découlaient de l'objectif général. Le premier objectif était d'analyser un vocabulaire de description de la danse dans les archives. Le deuxième objectif était d'analyser un vocabulaire de représentation de la danse dans la littérature recueilli grâce à des méthodes d'extraction automatique de connaissances. Le troisième et dernier objectif était de comparer les deux vocabulaires afin d'en distinguer la complémentarité quant à la description de l'expérience esthétique. La réalisation de ces objectifs nous permettait de répondre aux deux questions qui suivent. D'abord, est-ce que le vocabulaire de description de la danse dans les archives prend en compte l'expérience

esthétique? Ensuite, est-ce que le vocabulaire de représentation de la danse dans la littérature, recueilli grâce à des méthodes d'extraction automatique de connaissances, peut enrichir le vocabulaire de description de l'expérience esthétique de la danse dans les archives?

Dans le premier chapitre, celui présentant la revue de la littérature, nous avons effectué un survol de la description de la danse dans les services documentaires. Nous avons entre autres présenté deux cas exemplaires de vocabulaire de description de la danse : le thésaurus de descripteurs Collier et les vedettes-matière du *Dictionary catalog of the Dance Collection*. Nous avons ensuite souligné que l'expérience esthétique est une part essentielle de la danse, donc un point d'accès nécessaire aux archives de la danse. Et la description, en amont, influe grandement autant sur la préservation des archives que sur leur accès, en aval. La problématique était alors claire : le vocabulaire de description des archives de la danse peut bénéficier d'être enrichi de termes relatifs à l'expérience esthétique. Le monde des archives, des sciences de l'information, et le monde de la danse, de l'art, pourraient ainsi encore mieux se comprendre. Nous avons alors discuté des méthodes d'extraction de termes et de classification automatique, qui nous semblaient ouvrir de nouvelles avenues de recherche pour permettre l'expérience esthétique comme point d'accès aux archives de la danse.

Dans le deuxième chapitre, celui présentant la méthodologie, nous avons traité de la méthodologie employée lors de la réalisation de notre projet de recherche. D'abord, nous avons justifié le choix du thésaurus de descripteurs Collier comme vocabulaire de description de la danse dans les archives : il est en usage dans les archives canadiennes de la danse et il est disponible en langue française. Nous avons alors expliqué la démarche pour en extraire une liste de termes. Nous avons ensuite expliqué comment nous avons recueilli un vocabulaire structuré de l'expérience esthétique de la danse moderne à partir de

la constitution d'un corpus littéraire, suivie de la conversion numérique des documents, du filtrage du lexique, de l'extraction de termes discriminant, de la classification automatique des documents et de l'extraction de termes caractéristiques. Nous avons choisi le corpus constitué de deux recueils de textes de l'auteur français Stéphane Mallarmé, soit *Divagations* (édition par Eugène Fasquelle datant de 1897) et *Poésies* (édition par Émile Deman datant de 1899), pour diverses raisons : parce que l'œuvre littéraire mallarméenne est aux racines de l'art moderne et de la modernité littéraire; parce que l'œuvre littéraire mallarméenne incarne une expérience esthétique; parce que l'œuvre littéraire mallarméenne est intrinsèquement liée à la danse.

Dans le troisième chapitre, celui présentant les résultats et la discussion, nous avons présenté et analysé les résultats obtenus grâce à l'application de la méthodologie de recherche. Nous avons ainsi atteint les objectifs de notre recherche et répondu aux questions de recherche. D'abord, nous avons présenté et analysé les descripteurs qui composent le thésaurus Collier pour atteindre le premier des objectifs de notre recherche, soit analyser un vocabulaire de description de la danse dans les archives. Ensuite, nous avons présenté et analysé le vocabulaire structuré de l'expérience esthétique de la danse moderne extrait d'un corpus de textes de l'auteur Stéphane Mallarmé pour atteindre le deuxième objectif de notre recherche, soit analyser un vocabulaire de représentation de la danse dans la littérature recueilli grâce à des méthodes d'extraction automatique de connaissances. Puis nous avons comparé les descripteurs du thésaurus Collier et le vocabulaire structuré de l'expérience esthétique de la danse moderne en portant une attention particulière à la question de l'expérience esthétique pour atteindre le troisième objectif de notre recherche, soit comparer les deux vocabulaires afin d'en distinguer la complémentarité quant à la description de l'expérience esthétique. Finalement, nous avons soulevé les limites de la recherche; les limites relatives aux archives

de la danse autant que les limites relatives à l'extraction automatique de connaissances.

Nous avons ainsi été en mesure d'atteindre les objectifs de notre projet de recherche. Nous avons analysé les descripteurs qui composent le thésaurus Collier. Nous avons analysé un vocabulaire structuré de l'expérience esthétique de la danse moderne extrait à partir d'un corpus de textes de l'auteur Stéphane Mallarmé. Nous avons finalement comparé les deux vocabulaires. Nous avons donc également atteint l'objectif général de notre recherche, qui est de contribuer, du moins l'avons-nous fait en théorie et en partie, au développement d'outils de gestion de l'information en comparant le vocabulaire de description de la danse dans les archives et le vocabulaire de représentation de la danse dans la littérature pour en distinguer une possible complémentarité, particulièrement en ce qui a trait au vocabulaire de l'expérience esthétique.

Nous avons alors été en mesure de répondre aux deux questions qui sous-tendent notre recherche. Nous avons pu répondre, du moins en partie, à la première question de recherche, où nous nous interrogeons à savoir si le vocabulaire de description de la danse dans les archives prend en compte l'expérience esthétique. Nous avons conclu que le thésaurus de descripteurs Collier est un outil de description essentiellement factuelle qui ne prend pas en compte l'expérience esthétique. Nous avons pu répondre, du moins en partie, à la seconde question de recherche, où nous nous interrogeons à savoir si le vocabulaire de représentation de la danse dans la littérature, recueilli grâce à des méthodes d'extraction automatique de connaissances, peut enrichir le vocabulaire de description de l'expérience esthétique de la danse dans les archives. Le vocabulaire structuré de l'expérience esthétique de la danse moderne extrait d'un corpus de textes de l'auteur Stéphane Mallarmé permet effectivement une description de l'expérience esthétique de la danse moderne. Il propose des thèmes ainsi que des adjectifs liés à l'expérience esthétique de la

danse moderne. Il est alors jusqu'à un certain point complémentaire à un outil de description des archives de la danse comme le thésaurus de descripteurs Collier.

2. Contribution au domaine des sciences de l'information

2.1. Contribution théorique

Sur le plan théorique, nous avons effectué un survol de la documentation de la danse, un sujet encore peu exploré. Nous avons présenté certains outils de description de la danse dans les archives, soit les vedettes-matière du *Dictionary catalog of the Dance Collection* et le thésaurus de descripteurs Collier. Nous avons également procédé à une certaine analyse critique de ce dernier. Notre étude exploratoire en appelle d'autres, plus élaborées.

2.2. Contribution méthodologique

Sur le plan méthodologique, nous avons mené une première expérimentation qui justifie en partie la pertinence de certaines méthodes d'extraction de connaissances dans le développement ou dans l'enrichissement de ressources documentaires pour le domaine des arts d'interprétation tels que la danse.

2.3. Contribution pratique

Sur le plan pratique, nous avons formulé une première suggestion d'amélioration de certains thésaurus employés dans les archives de la danse. Un thésaurus au vocabulaire plus factuel, comme le thésaurus Collier, peut être enrichi de termes à propos de l'expérience esthétique comme ceux qui

composent le vocabulaire structuré de l'expérience esthétique de la danse moderne d'après Stéphane Mallarmé.

3. Recherches futures

Les résultats obtenus lors de la réalisation de notre projet de recherche nous amènent à souhaiter une exploration plus en profondeur de la question de la description de l'expérience esthétique de la danse. Si l'art s'inspire des archives, les sciences de l'information doivent s'inspirer de l'art. Rappelons-le, selon Catherine Johnson et Allegra Fuller Snyder (1999), « [p]reserving the remnants of an art form and cultural activity dependent on time and space requires no less creativity than that needed to create the dance itself ». L'extraction de connaissances semble offrir d'innombrables possibilités à ce propos. On peut ainsi espérer créer de nouvelles ontologies et formaliser certains pans de la connaissance humaine que l'on envisageait jusqu'à présent comme insaisissables. Sans oublier que la description des documents est intrinsèquement liée à leur préservation et à leur accès. La mise en valeur du patrimoine culturel qu'est la danse passe ainsi en partie par la description des traces qu'elle laisse dans son sillage.

Bibliographie

- Adams, Lawrence, Amy Bowring et Clifford Collier. 2008. *Base de données intégrée de danse canadienne : guide des normes et thésaurus de descripteurs Collier*. Toronto : Dance Collection Danse.
- Adams, Lawrence. 2004. *Building your legacy : an archiving handbook for dance*. Toronto : Dance Collection Danse.
- Aitchison, Jean et Stella Dextre Clarke. 2004. The thesaurus : a historical viewpoint with a look to the future. In *The thesaurus : review, renaissance, and revision*, sous la dir. de Sandra K. Roe et Alan R. Thomas, 5-21. Binghamton, New York : Haworth Information Press.
- Armelagos, Adina. 1977. « *Dictionary catalog of the Dance Collection* ». *Dance Research Journal* 9(2) : 29-31.
- Baca, Murtha, Patricia Harpring, Elisa Lanzi, Linda MacRae et Ann Whiteside. 2006. *Cataloging cultural objects : a guide to describing cultural works and their images*. Chicago : American Library Association.
- Bannerman, Henrietta. 2006. « A Dance of transition : Martha Graham's Herodiade ». *Dance Research* 24(1) : 1-20.
- Bavelier, Anne. 2009. « Pina Bausch, la danse comme amour pur », *Le Figaro*, 1 juillet 2009.
- Bégoc, Janig, Nathalie Boulouch et Elvan Zabunyan, dir. 2010. *La performance : entre archives et pratiques contemporaines*. Châteaugiron : Archives de la critique d'art; Rennes : Presses universitaires de Rennes.
- Bekiari, Chryssoula, Martin Doerr et Patrick Le Bœuf, dir. 2010. *FRBR, object-oriented definition and mapping to FRBR_{ER} (version 1.0.1)*. International Working Group on FRBR and CIDOC CRM Harmonisation. <http://www.cidoc-crm.org/docs/frbr_oo/frbr_docs/FRBRoo_V1.0.1.pdf>⁴

⁴ Tous les liens vers des références en ligne sont à jour en date du 1^{er} mai 2013.

- Beltrando-Patier, Marie-Claire, dir. 2004. *Histoire de la musique*. Paris : Bordas.
- Bergesen, Albert et Allison Jones. 1992. Decoding the syntax of modern dance. In *Vocabularies of public life : empirical essays in symbolic structure*, sous la dir. de Robert Wuthnow, 169-184. London, New York : Routledge.
- Bibliothèque nationale de France. 2012. *Bibliothèque nationale de France*. <<http://www.bnf.fr>>
- Block, Haskell M. 1977. *Mallarmé and the symbolist drama*. Westport : Greenwood Press.
- Bourigault, Didier, Christian Jacquemin et Marie-Claude L'Homme. 2001. Introduction. In *Recent advances in computational terminology*, sous la dir. de Didier Bourigault, Christian Jacquemin et Marie-Claude L'Homme, viii-xviii. Amsterdam, Philadelphia : John Benjamins Publishing Company.
- Bowring, Amy. 2011. « Dance Collection Danse : Canada's largest archive and research center for theatrical dance history », *Dance Chronicle* 34(2) : 276-289.
- Brandstetter, Gabriele. 2006. « Preserving the performance : scholarship as art? ». *Performance Research* 11(2) : 15-23.
- Brunet, Étienne. 2000. « Qui lemmatise dilemme attise ». *Lexicometrica* 2. <<http://lexicometrica.univ-paris3.fr/article/numero2/brunet2000.html>>
- Bureau de la traduction. 2012. *Termiumplus : la banque de données terminologiques et linguistiques du gouvernement du Canada*. <<http://www.termiumplus.gc.ca/>>
- Busa, Roberto. 1974-1980. *Index Thomisticus : sancti Thomae Aquinatis operum omnium indices et concordantiae, in quibus verborum omnium et singulorum formae et lemmata cum suis frequentiiis et contextibus variis modis referuntur quaeque / consociata plurium opera atque electronico IBM automato usus digessit Robertus Busa SI*. Stuttgart-Bad Cannstatt : Frommann-Holzboog.
- Busa, Roberto. 2004. Foreword. In *A companion to digital humanities*, sous la dir. de Susan Schreibman, Ray Siemens et John Unsworth. Oxford : Blackwell.

- Caïozzi, Denis. 2005. *Stravinsky and the Ballets Russes*. DVD. Paris : Bel Air Media.
- Caron, Pascal. 2006. *Faunes : poésie, corps, danse, de Mallarmé à Nijinski*. Paris : Honoré Champion.
- Caws, Mary Ann. 1994. Dancing with Mallarmé and Seurat (and Loïe Fuller, Hérodiade and La Goulue). In *Artistic relations : literature and the visual arts in nineteenth-century France*, sous la dir. de Peter Collier et Robert Lethbridge, 291-302. New Haven : Yale University Press.
- Caws, Mary Ann. 1998. Mallarmé's Progeny. In *Mallarmé in the twentieth century*, sous la dir. de Robert Greer Cohn et Gerald Gillespie, 86-91. Cranbury, London, Mississauga : Associated University Press.
- Chaffee, Gary. 2011. « Preserving transience : ballet and modern dance archives ». *Libri : International Journal of Libraries & Information Services* 61(2) : 125-130.
- Clarke, Paul et Julian Warren. 2009. « Ephemera : between archival objects and events ». *Journal of the Society of Archivists* 30(1) : 45-66.
- Compagnie Marie Chouinard. 2012. *Faune_CMC*. <<http://www.mariechouinard.com/l-apres-midi-d-un-faune-113.html>>
- Couch, Nena. 1994. Dance collections. In *Managing performing arts collections in academic and public libraries*, sous la dir. de Caroline A. Sheehy, 41-72. Westport: Greenwood Press.
- Dalbello, Marija. 2011. « A genealogy of digital humanities ». *Journal of Documentation* 67(3) : 480 - 506
- Dance Collection Danse. 2012. *Dance Collection Danse : Archives, Publishing, Research and Education*. <<http://www.dcd.ca/>>
- Dance Heritage Coalition. 2012. *Dance Heritage Coalition*. <<http://www.danceheritage.org/>>
- Dégez, Danièle et Dominique Ménillet. 2001. *Thésauroglossaire des langages documentaires : un outil de contrôle sémantique*. Paris : ADBS Éditions.
- Delfel, Guy. 1951. *L'esthétique de Stéphane Mallarmé*. Paris : Flammarion.

- Desalme, Aubierge. 2007. « Sur les pas de la danse : l'exemple des bibliothèques américaines ». *Bulletin des bibliothèques de France* 52(4) : 13-22.
- Dewey, John. 1934. *Art as experience*. New York : Minton, Balch & Company.
- Doyon, Frédérique. 2006. « Petite histoire de censure », *Le Devoir*, 30 octobre 2006.
- Drouin, Patrick. 2003. « Term extraction using non-technical corpora as a point of leverage ». *Terminology* 9(1) : 99-115.
- El Raheb, Katerina et Yannis Ioannidis. 2011. A labanotation based ontology for representing dance movement. In *9th international gesture workshop, May 2011, Athens, Greece*, sous la dir. de E. Efthimiou et G. Kouroupetroglou, 41-45.
- Evens, Tom et Laurence Hautekeete. 2011. « Challenges of digital preservation for cultural heritage institutions ». *Journal of Librarianship and Information Science* 43(3) : 157-165.
- Feldman, Ronen et James Sanger. 2007. *The text mining handbook : advanced approaches in analyzing unstructured data*. New York : Cambridge University Press.
- Forest, Dominic. 2006. *Application de techniques de forage de textes de nature prédictive et exploratoire à des fins de gestion et d'analyse thématique de documents textuels non structurés*. Thèse de doctorat. Montréal : Université du Québec à Montréal.
- Forest, Dominic. 2009. Impacts de la variation du nombre de traits discriminants sur la catégorisation des documents, avec la collaboration de Astrid van Hoeydonck, Danny Létourneau et Martin Bélanger. In *DEFT'09 « DEfi Fouille de Textes », Atelier de clôture, Paris, 22 juin 2009*. <http://deft.limsi.fr/actes/2009/pdf/4_forest.pdf>
- Forest, Dominic. 2011. Notes du cours SCI16136 – Fouille de documents. Inédit. Montréal : Université de Montréal.
- Forest, Dominic. 2012. *Fouille de textes et analyse thématique : techniques prédictive et descriptive pour la gestion et l'analyse de documents textuels non structurés*. Sarrebruck : Éditions universitaires européennes.
- FRANTEXT. 2012. *Base textuelle FRANTEXT*. <www.frantext.fr>

- Getty Research Institute. 2012. *Art & architecture thesaurus online*. <http://www.getty.edu/research/conducting_research/vocabularies/aat/>
- Hansen Kopp, Leslie. 1995. *Dance archives : a practical manual for documenting and preserving the ephemeral art*. Lee, MA : Preserve.
- Hegel, Georg Wilhelm Friedrich. 1835. Troisième partie : système des arts particuliers, introduction et division. In *Esthétique : tome premier*. Trad. de l'allemand par C. Bénard, version numérique par Daniel Banda, 232-242. <http://classiques.uqac.ca/classiques/hegel/esthetique_1/Hegel_Esthetique_tome_I.pdf>
- Hjørland, Birger. 2002. « Domain analysis in information science : eleven approaches ». *Journal of Documentation* no 58(4) : 422-462.
- Hjørland, Birger. 2008. « What is knowledge organization (KO)? ». *Knowledge Organization* 35(2-3): 86-101.
- Hockey, Susan. 2004. The history of humanities computing. In *A companion to digital humanities*, sous la dir. de Susan Schreibman, Ray Siemens et John Unsworth. Oxford : Blackwell.
- Hodge, Gail. 2000. *Systems of knowledge organization for digital libraries : beyond traditional authority files*. Washington : Council on Library and Information Resources.
- Hudon, Michèle, Clément Arsenault, Lyne Da Sylva et Dominic Forest. 2009. Le traitement du document. In *Introduction aux sciences de l'information*, sous la dir. de Jean-Michel Salaün et Clément Arsenault, 53-100. Montréal : Presses de l'Université de Montréal.
- Hudon, Michèle. 2009. *Guide pratique pour l'élaboration d'un thésaurus documentaire*, avec la collaboration de Danièle Dégez et Dominique Ménillet. Montréal : Éditions ASTED.
- Ibekwe-SanJuan, Fidelia. 2007. *Fouille de textes : méthodes, outils et applications*. Paris : Hermès, Lavoisier.
- Jacquemin, Christian. 2001. *Spotting and discovering terms through natural language processing*. Cambridge : MIT Press.
- Johnson, Catherine J. et Allegra Fuller Snyder. 1999. *Securing our dance heritage: issues in the documentation and preservation of dance*. Washington : Council on Library and Information Resources.

- Jones, Sarah, Daisy Abbott et Seamus Ross. 2009. « Redefining the performing arts archive ». *Archival Science* 9(3-4) : 165-171.
- Keens, William, Leslie Hansen Kopp et Mindy Levine. 1990. *Images of American dance : documenting and preserving a cultural heritage*. Washington : National Endowment for the Arts.
- Kristeva, Julia. 1974. *La révolution du langage poétique : l'avant-garde à la fin du XIX^e siècle, Lautréamont et Mallarmé*. Paris : Éditions du Seuil.
- LaFrance, Cheryl. 2011. « Choreographers' archives : three case studies in legacy preservation ». *Dance Chronicle* 34(1) : 48-76.
- Le Boeuf, Patrick et David Miller. 2005. « “Such stuff as dreams are made on” : how does FRBR fit performing arts ». *Cataloging & Classification Quarterly* 39 (3/4) : 151-178.
- Le Boeuf, Patrick. 2002. Le spectacle vivant en tant qu’objet documentaire et le modèle conceptuel de données des FRBR. In *Arts du spectacle : patrimoine et documentation, XXIII^e congrès international, Paris 25-30 septembre 2000*, 162-175. Paris : Société internationale des bibliothèques et musées des arts du spectacle, Bibliothèque nationale de France.
- Le Boeuf, Patrick. 2006. « ...Qui se pavane et s’agite son heure sur scène et qu’ensuite on n’entend plus... » : les éléments à prendre en compte dans un futur modèle conceptuel du spectacle vivant et de l’information relative à ses archives. <http://www.cidoc-crm.org/docs/2006_LeBoeuf_fr.pdf>
- Le Moal, Philippe. 1998. *La danse à l’épreuve de la mémoire*. Paris : Ministère de la culture et de la communication.
- Lebart, Ludovic et André Salem. 1994. *Statistique textuelle*. Paris : Dunod.
- Lemay, Yvon. 2009. « Art et archives : une perspective archivistique ». *Encontros Bibli*, édition spéciale : 64-86. <<http://www.periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2009v14nesp1p64/19834>>
- Levinson, André. 1983. The idea of dance : from Aristotle to Mallarmé. In *What is dance?*, sous la dir. de Roger Copeland et Marshall Cohen, 47-54. New York : Oxford University Press.

- Library of Congress. 2009. Chapter 25 : Uniform titles. In *Library of Congress rule interpretations*. <<http://www.loc.gov/aba/pcc/naco/training/lcri25.pdf>>
- Lourdou, Dorothy. 1970. « The dance collection automated book catalog ». *LARC Reports* 3(3) : 17-38.
- Mallarmé, Stéphane. 1897. *Divagations*. Paris : Bibliothèque Charpentier, Eugène Fasquelle. <<http://fr.wikisource.org/wiki/Divagations>>
- Mallarmé, Stéphane. 1899. *Poésies*. Bruxelles : Edmond Deman. <http://fr.wikisource.org/wiki/Po%C3%A9sies_%28Mallarm%C3%A9_%29/%C3%89dition_1899>
- Mallarmé, Stéphane. 2003. *Oeuvres complètes*. Tome II, sous la dir. de Bertrand Marchal. Paris : Gallimard, Bibliothèque de la Pléiade.
- Manning, Christopher et Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge : MIT Press.
- Marchal, Bertrand. 2009. Mallarmé en « Pléiade » ou le double corp(us) du poète. In *La bibliothèque de la Pléiade : travail éditorial et valeur littéraire*, sous la dir. de Joëlle Gleize et Philippe Roussin, 119-126. Paris : Éditions des archives contemporaines.
- Matluck Brooks, Lynn. 2011. « A bold step forward : Genevieve Oswald and the dance collection of the New York Public Library ». *Dance Chronicle* 34(3): 447-486.
- Matsumuro, Saburo, Nobuo Takeuchi et Masakatsu Kaneko. 1991. *Mallarmé's Text Database*. Tokyo : Taiga Shuppan.
- New York Public Library. 1974. *Dictionary catalog of the Dance Collection : a list of authors, titles, and subjects of multi-media materials in the Dance Collection of the Performing Arts Research Center of the New York Public Library*. New York, Boston : New York Public Library, Astor, Lenox and Tilden Foundations, G. K. Hall & Co.
- New York Public Library. 2012. *Using the dance archives : avenues of access*. <<http://www.nypl.org/locations/tid/55/node/33837>>
- Nietzsche, Friedrich. 1884. *Also sprach Zarathustra*. Chemnitz : Schmeitzner. <<http://www.nietzschesource.org/#eKGWB/Za-III-Tafeln-23>>

- Nietzsche, Friedrich. 1889. *Götzen-Dämmerung oder Wie man mit dem Hammer philosophirt*. Leipzig : Neumann. <<http://www.nietzschesource.org/#eKGWB/GD-Deutsche-7>>
- Nietzsche, Friedrich. 1903. *Ainsi parlait Zarathoustra*. In *Œuvres complètes de Frédéric Nietzsche*, trad. de l'allemand par Henri Albert. Paris : Mercure de France. <http://fr.wikisource.org/wiki/Ainsi_parlait_Zarathoustra/>
- Nietzsche, Friedrich. 2010. *Le crépuscule des idoles ou Comment on philosophe avec un marteau*, trad. de l'allemand par Henri Albert. Paris : L'Herne.
- Office québécois de la langue française. 2012. *Le grand dictionnaire terminologique*. <<http://gdt.oqlf.gouv.qc.ca>>
- Oswald, Genevieve. 1968. « Creating tangible records for an intangible art ». *Special Libraries* 59 : 146-151.
- Patrimoine canadien. 2012. *Artefacts Canada*. <<http://www.pro.rcip-chin.gc.ca/artefact/index-fra.jsp>>
- Phelan, Peggy. 1993. *Unmarked : the politics of performance*. London, New York : Routledge.
- Poinsot, Jean-Marc. 2004. Avant-propos. In *Les artistes contemporains et l'archive*, sous la dir. de Sylvie Mokhtari, 5-8. Rennes : Presses universitaires de Rennes.
- Potter, Michelle. 1992. « Dance archives in Australia: the unique material of the National Library ». *Dance Research* 10(2) : 109-120.
- Potter, Michelle. 2001. « Dancing on the web ». *National Library of Australia News*, April 2001 : 11-14.
- Procházka, David. 2006. « The development of uniform titles for choreographic works ». *Cataloging & Classification Quarterly* 42(1) : 7-20.
- Projet Gutenberg. 2012. <http://www.gutenberg.org/wiki/FR_Principal>
- Richard, Jean-Pierre. 1961. Chapitre VII : l'idée. In *L'univers imaginaire de Mallarmé*, 373-437. Paris : Editions du Seuil.
- Roe, Sandra K. et Alan R. Thomas. 2004. Introduction. In *The thesaurus : review, renaissance, and revision*, sous la dir. de Sandra K. Roe et Alan

- R. Thomas, 23-34. Binghamton, New York : Haworth Information Press.
- Rowat, Theresa. 2005. *Review of recent approaches to growing dance legacy*. Canada Council for the Arts.
- Sartre, Jean-Paul et Arlette Elkaïm-Sartre. 1986. *Mallarmé : la lucidité et sa face d'ombre*. Paris : Gallimard.
- Scholz, Ulrich. 2005. *Choreographic avant-garde 6 : the dance company Rubato*. DVD. Bremen : Deutsches Tanzfilminstitut Bremen.
- Siegel, Marcia B. 1972. *At the vanishing point : a critic looks at dance*. New York : Saturday Review Press.
- Souriau, Étienne. 1951. Préface. In *L'esthétique de Stéphane Mallarmé* de Guy Delfel, 7-14. Paris : Flammarion.
- Svensson, Patrik. 2009. « Envisioning the digital humanities ». *Digital Humanities Quarterly* 6(1).
- Svensson, Patrik. 2012. « Humanities computing as digital humanities ». *Digital Humanities Quarterly* 3(3).
- Thomas, Alan R. 2004. Teach yourself thesaurus : exercices, readings and resources. In *The thesaurus : review, renaissance, and revision*, sous la dir. de Sandra K. Roe et Alan R. Thomas, 23-34. Binghamton, New York: Haworth Information Press.
- Valaskakis-Tembeck, Iro, Selma Odom et Norma Sue Fisher-Stitt. 1997. « Dance research in Canada ». *Dance Research Journal* 29(1) : 107-110.
- Van Slype, Georges. 1987. *Les langages d'indexation : conception, construction et utilisation dans les systèmes documentaires*. Paris : Éditions d'Organisation.
- Verlaine, Paul. 1884. Stéphane Mallarmé. In *Les Poètes maudits : Tristan Corbière, Arthur Rimbaud, Stéphane Mallarmé*, 42-56. Paris : Vanier. <<http://gallica.bnf.fr/ark:/12148/btv1b86184725/f70.image.r=poetes%20maudits.langFR>>
- Véronis, Jean. 2012. *Un antidictionnaire*. <<http://sites.univ-provence.fr/veronis/donnees/index.html>>

Waller, Suzanne. 1999. *L'analyse documentaire : une approche méthodologique*, en collaboration avec Claudine Masse. Paris : ADBS Éditions.

Wikisource. 2012. *Wikisource, la bibliothèque libre*. <<http://fr.wikisource.org/>>

Zachmann, Gayle. 1998. Offensive moves in Mallarmé : dancing with des astres. In *Confrontations : politics and aesthetics in nineteenth-century France*, sous la dir. de Kathryn M. Grossman, Michael E. Lane, Bénédicte Monicat et Willa Z. Silverman, 187-200. Amsterdam, Atlanta : Rodopi.

Annexe 1. Extrait du thésaurus Collier : le terme *ballet*

BALLET

Note d'application : utilisé pour les objets se rapportant à la danse théâtrale présentée dans une forme artistique stylisée qui a été développée en Italie et en France au 17^e siècle, et qui s'est propagée en Europe, en Amérique et en Asie.

Pour les objets se rapportant à la technique du ballet

UTILISER : **TECHNIQUE + BALLET**

v/s : ballet classique

ballet d'école

ballet moderne

balletto

ballo

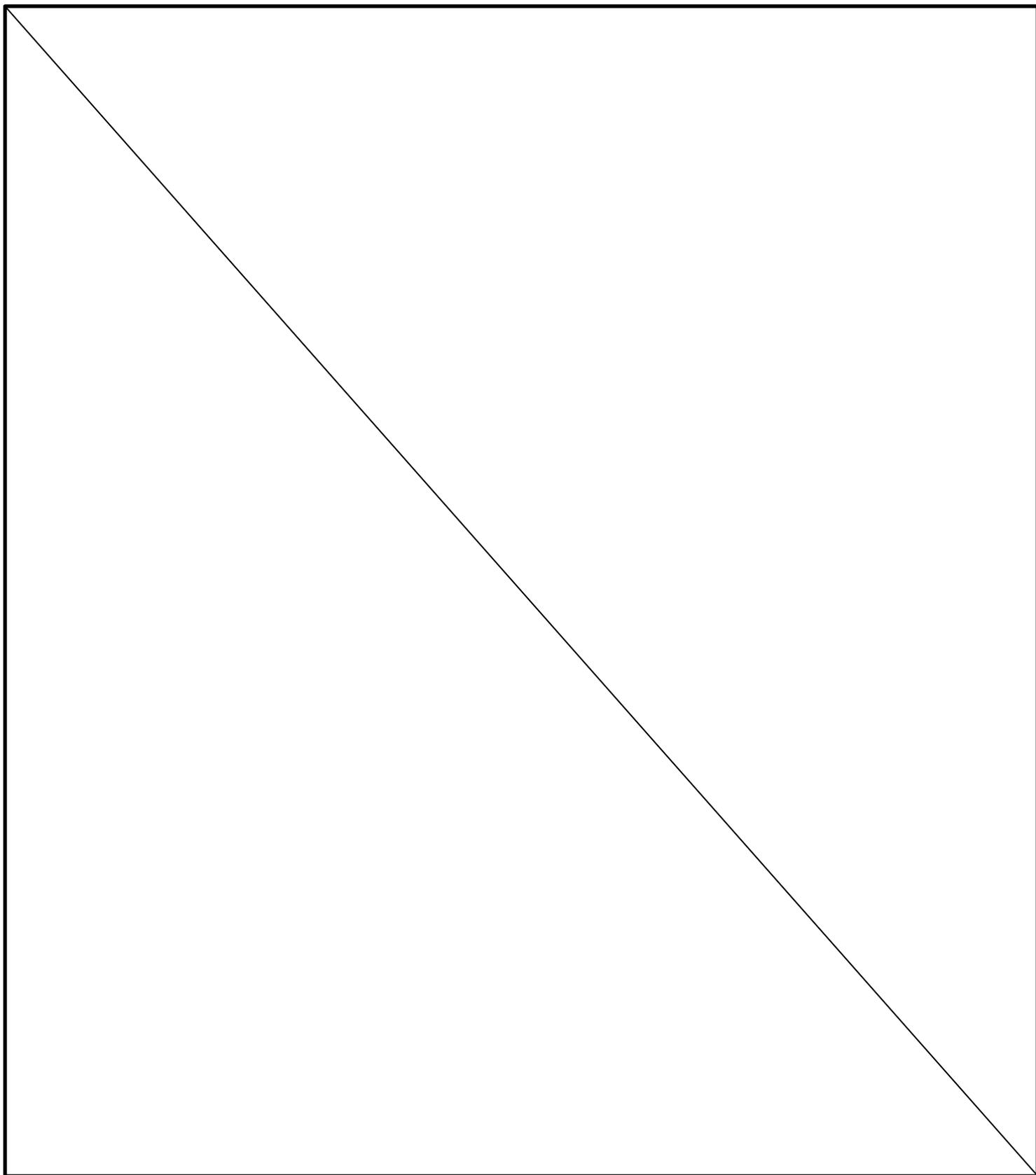
danse d'école

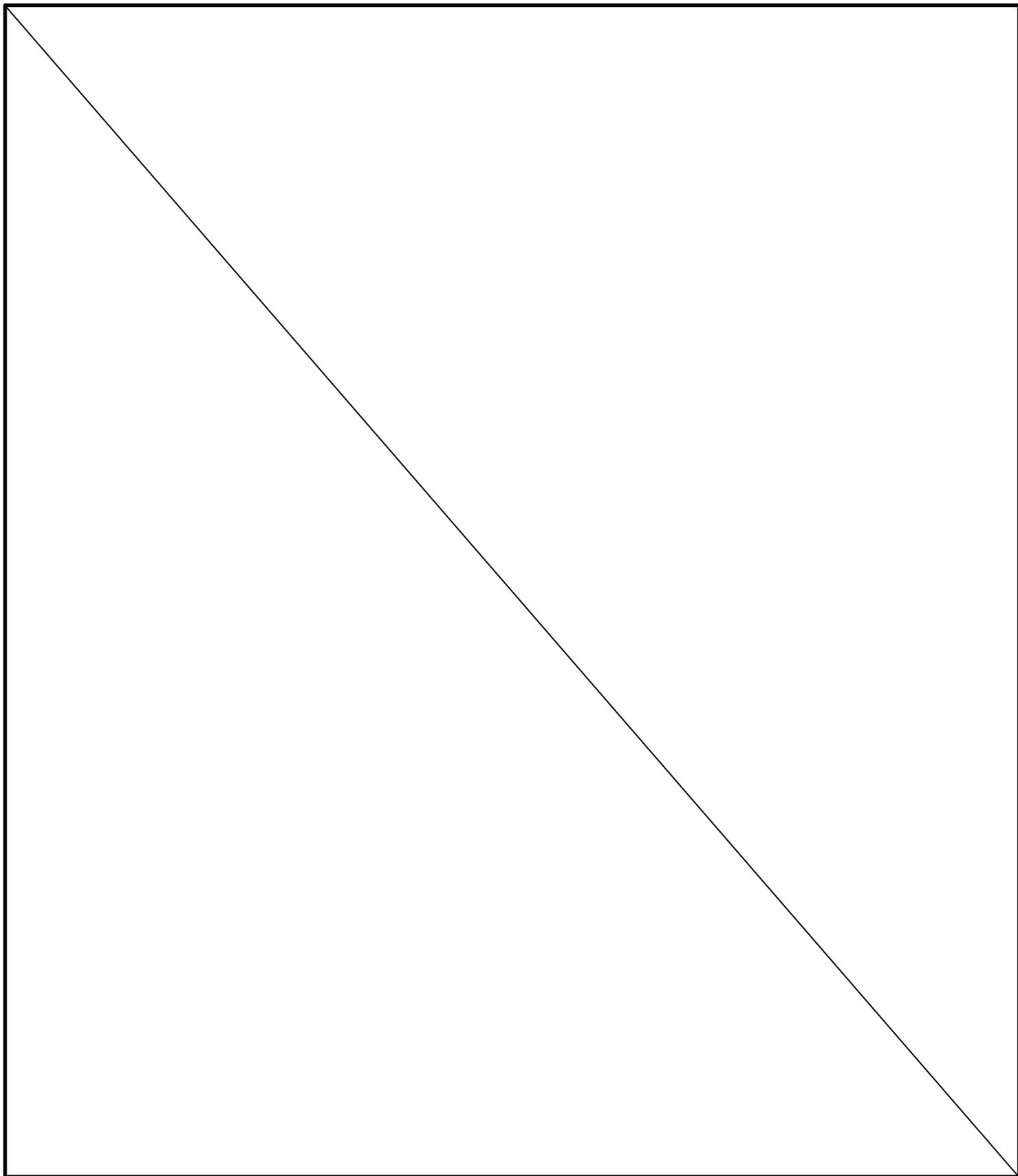
danse opératique

danze accademica

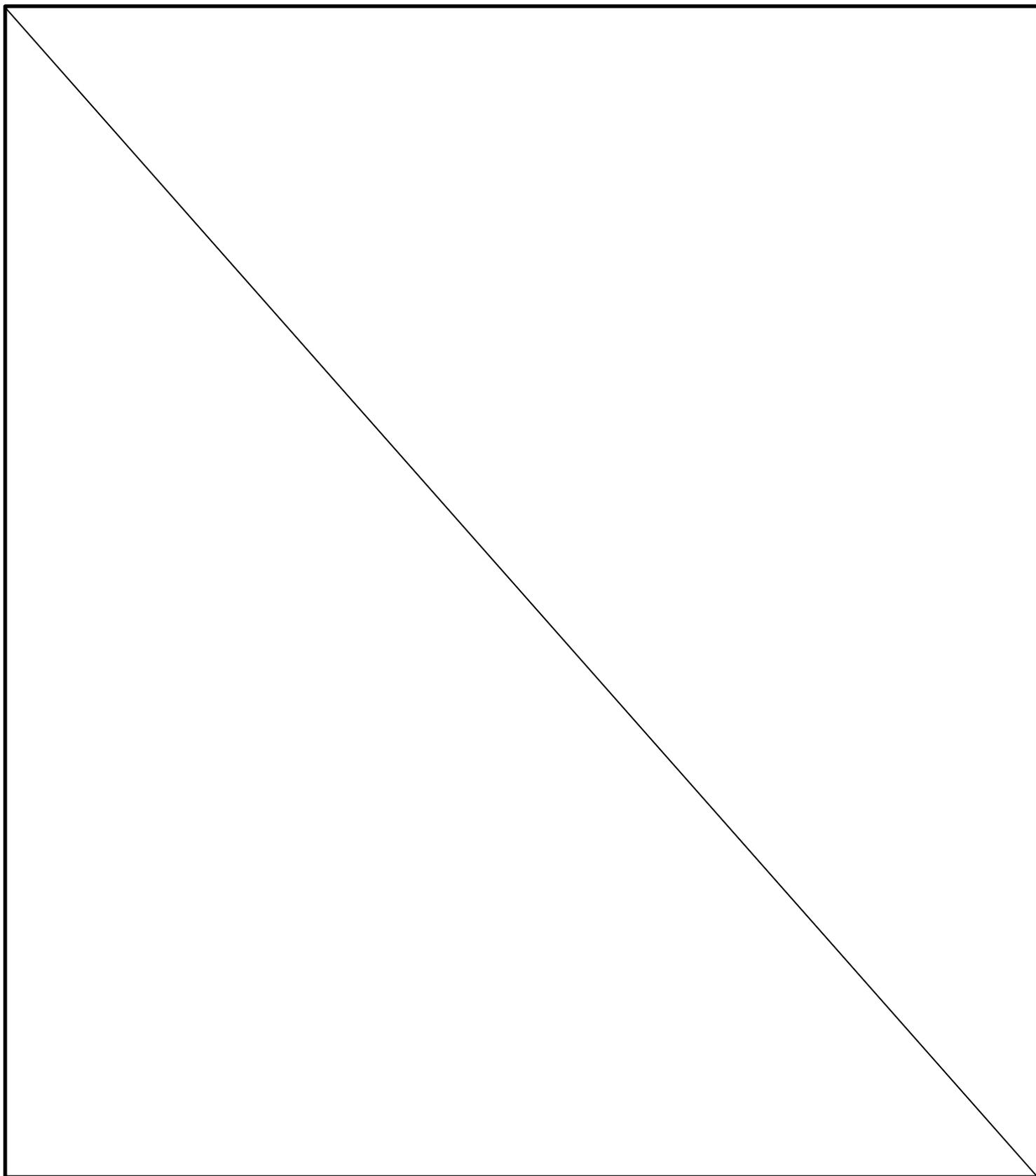
v/a : **BALLET DE COUR**

Annexe 2. Catégorie *interprétation*

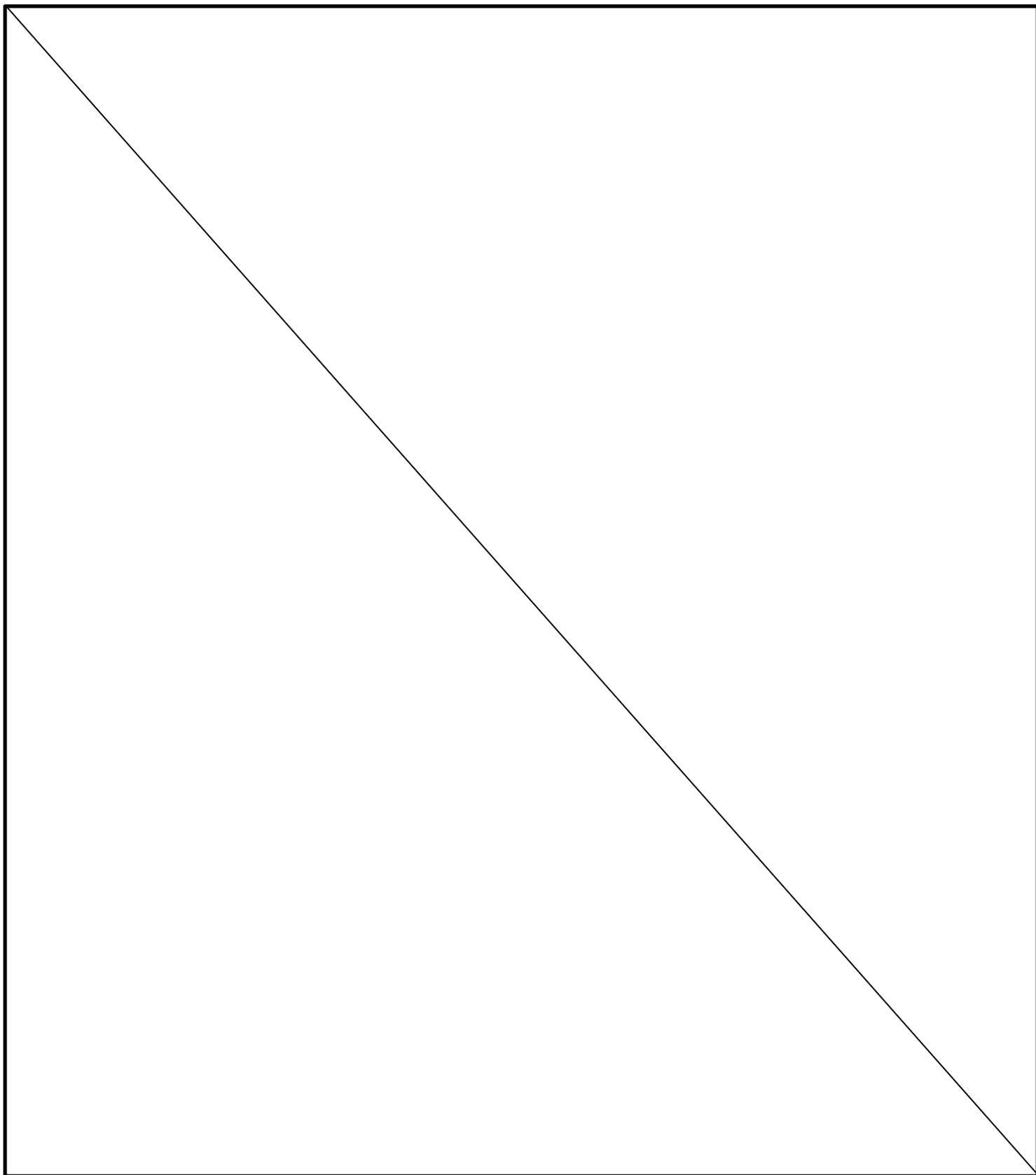




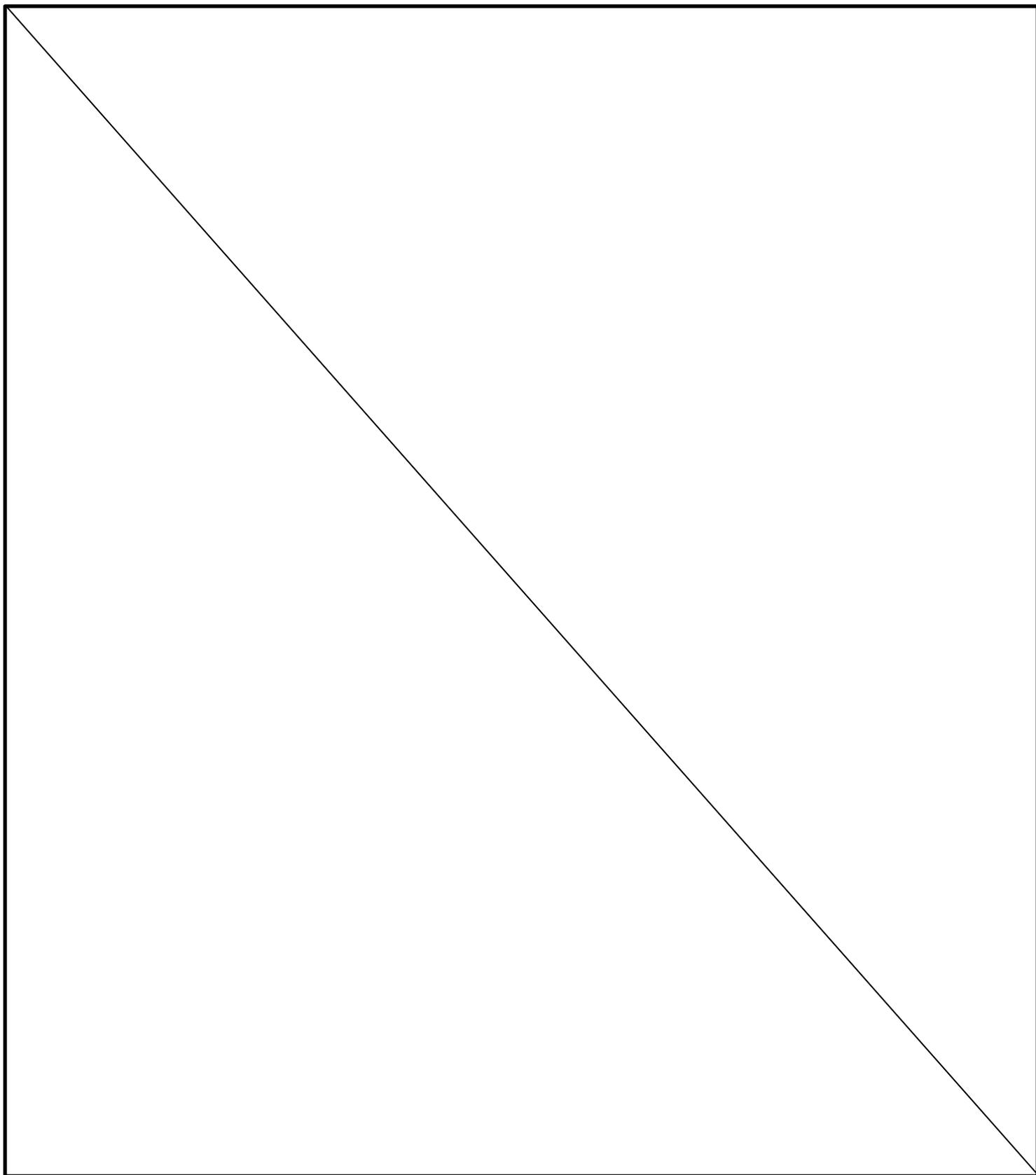
Annexe 3. Catégorie *autres arts*



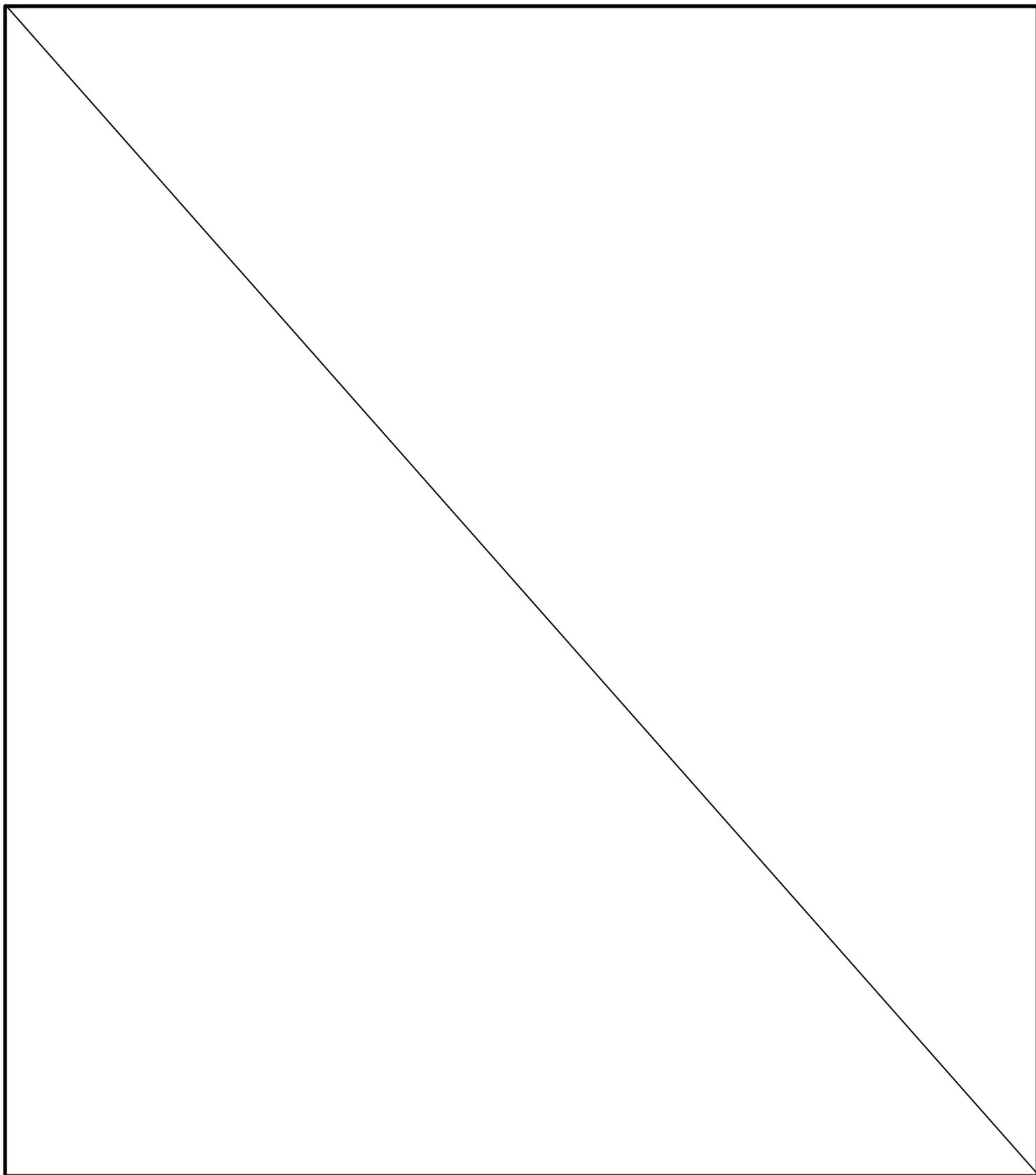
Annexe 4. Catégorie *études de la danse*



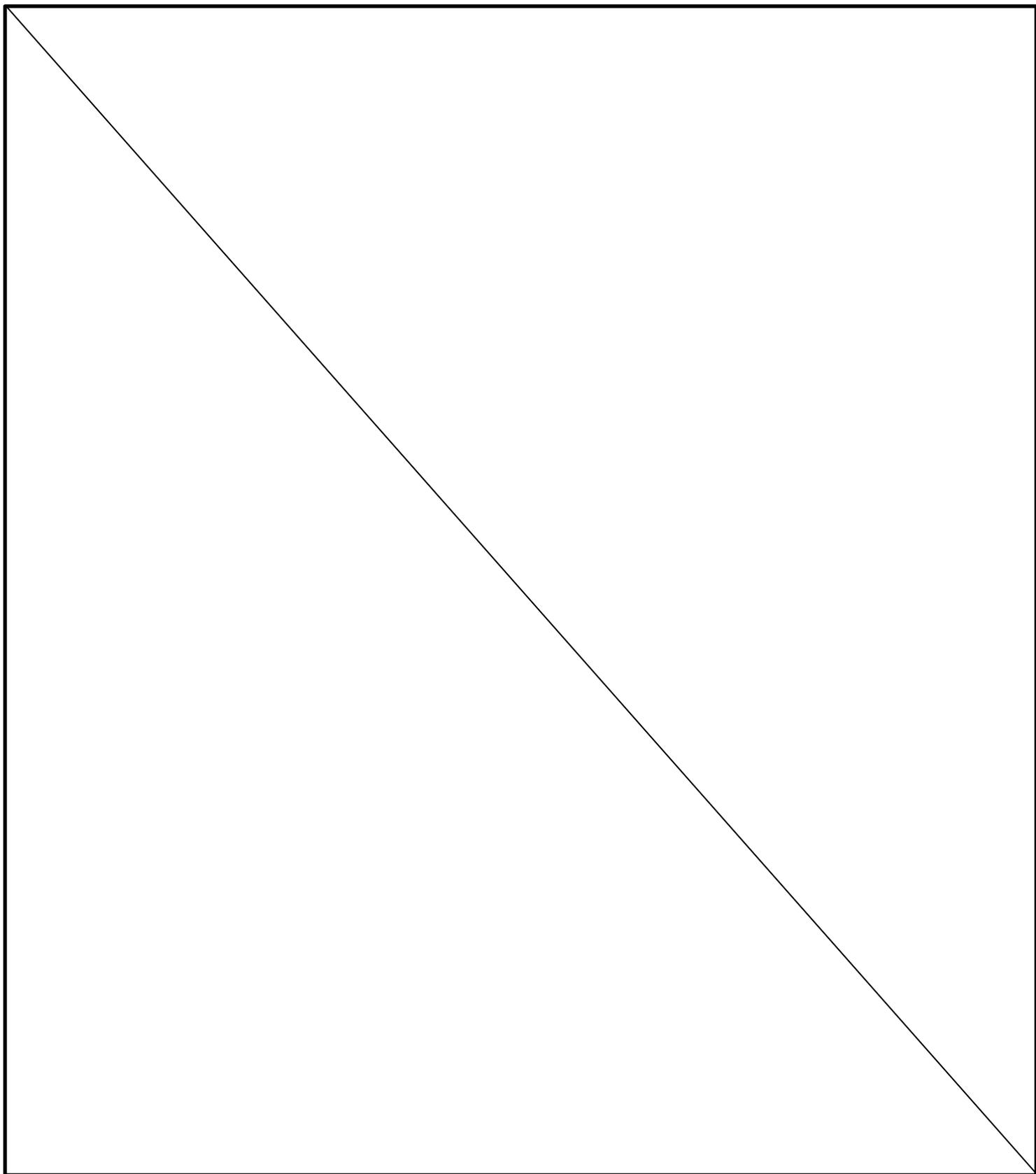
Annexe 5. Catégorie *administration/gestion*



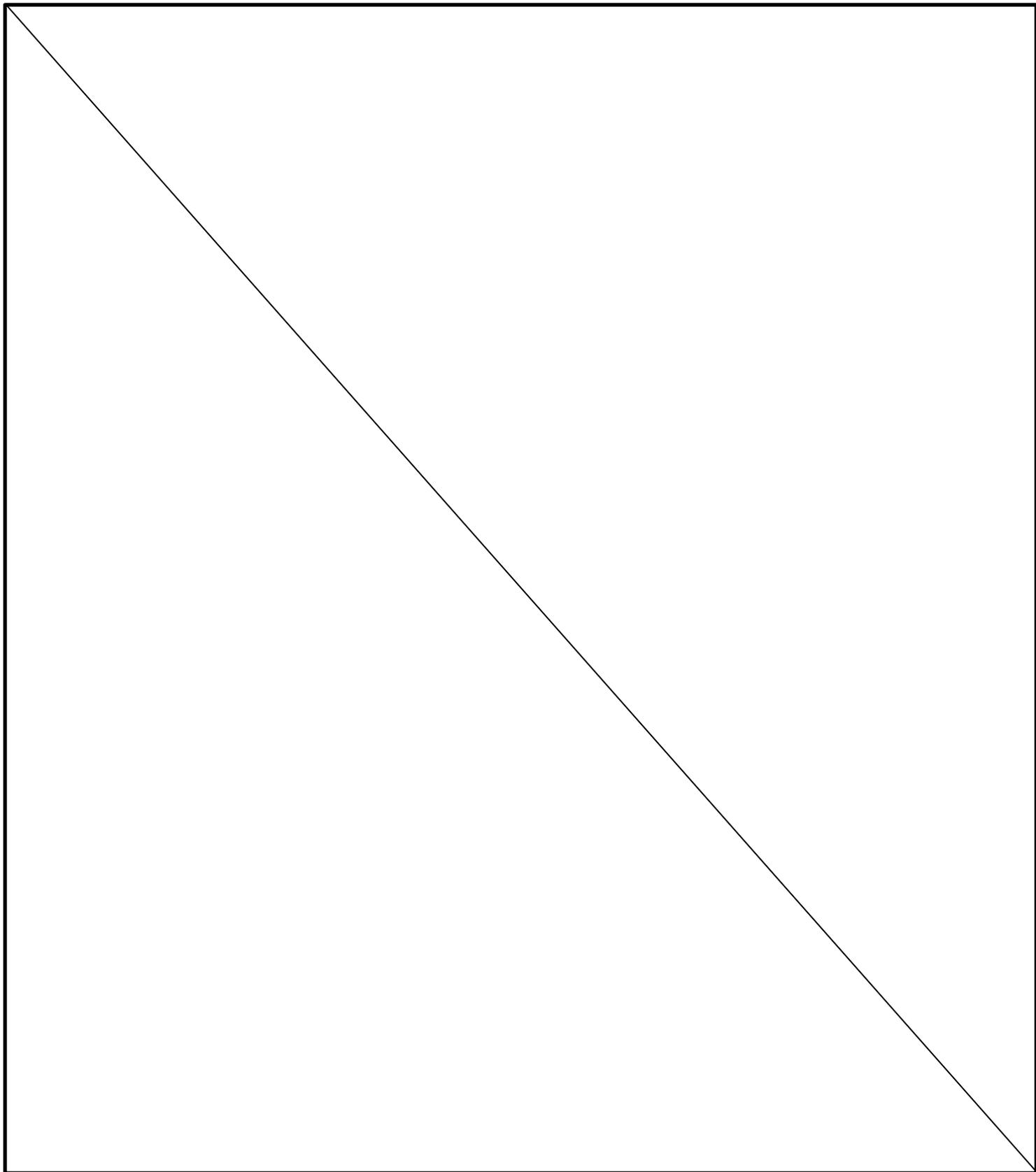
Annexe 6. Catégorie *formation*



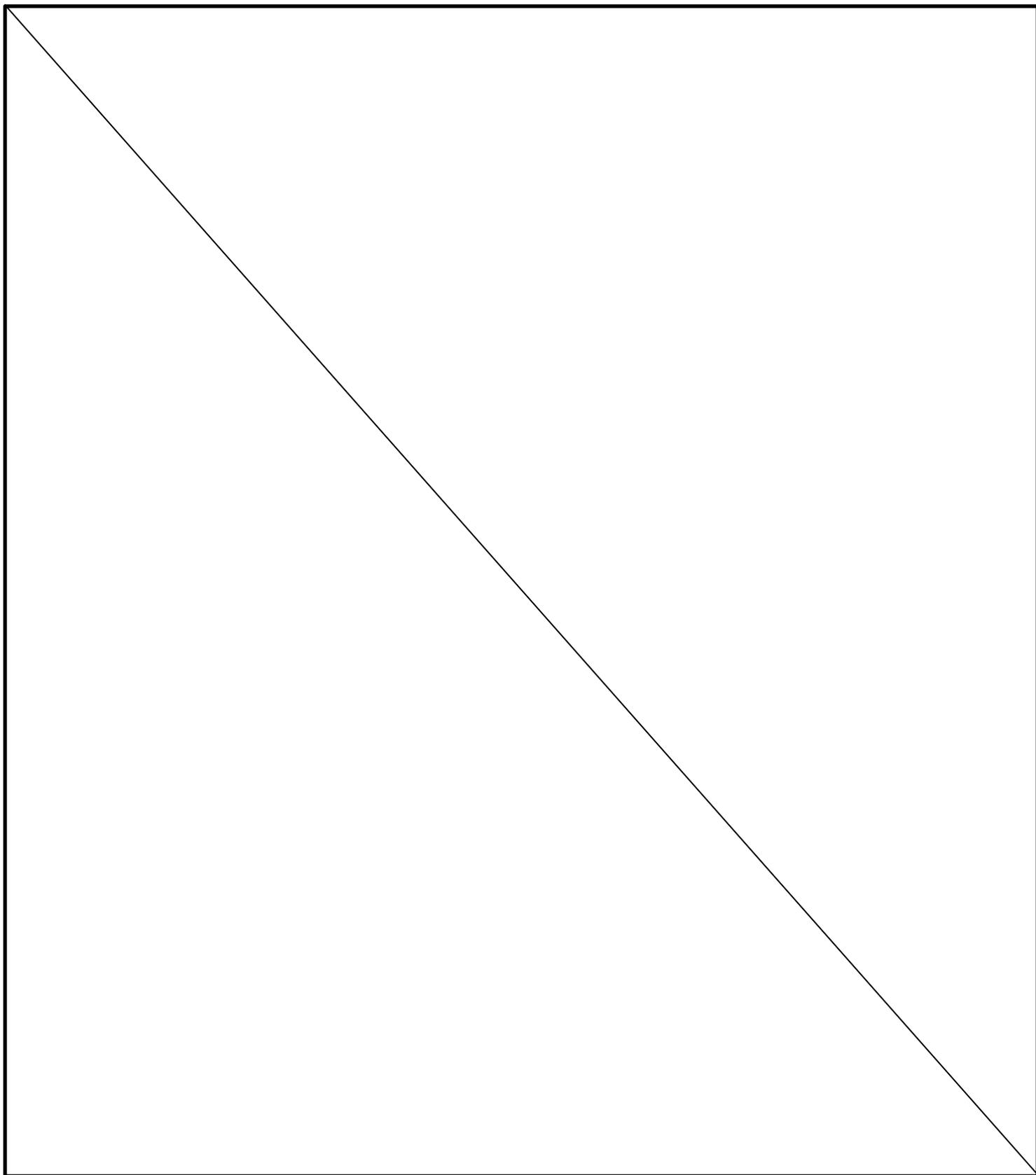
Annexe 7. Catégorie *matériel*



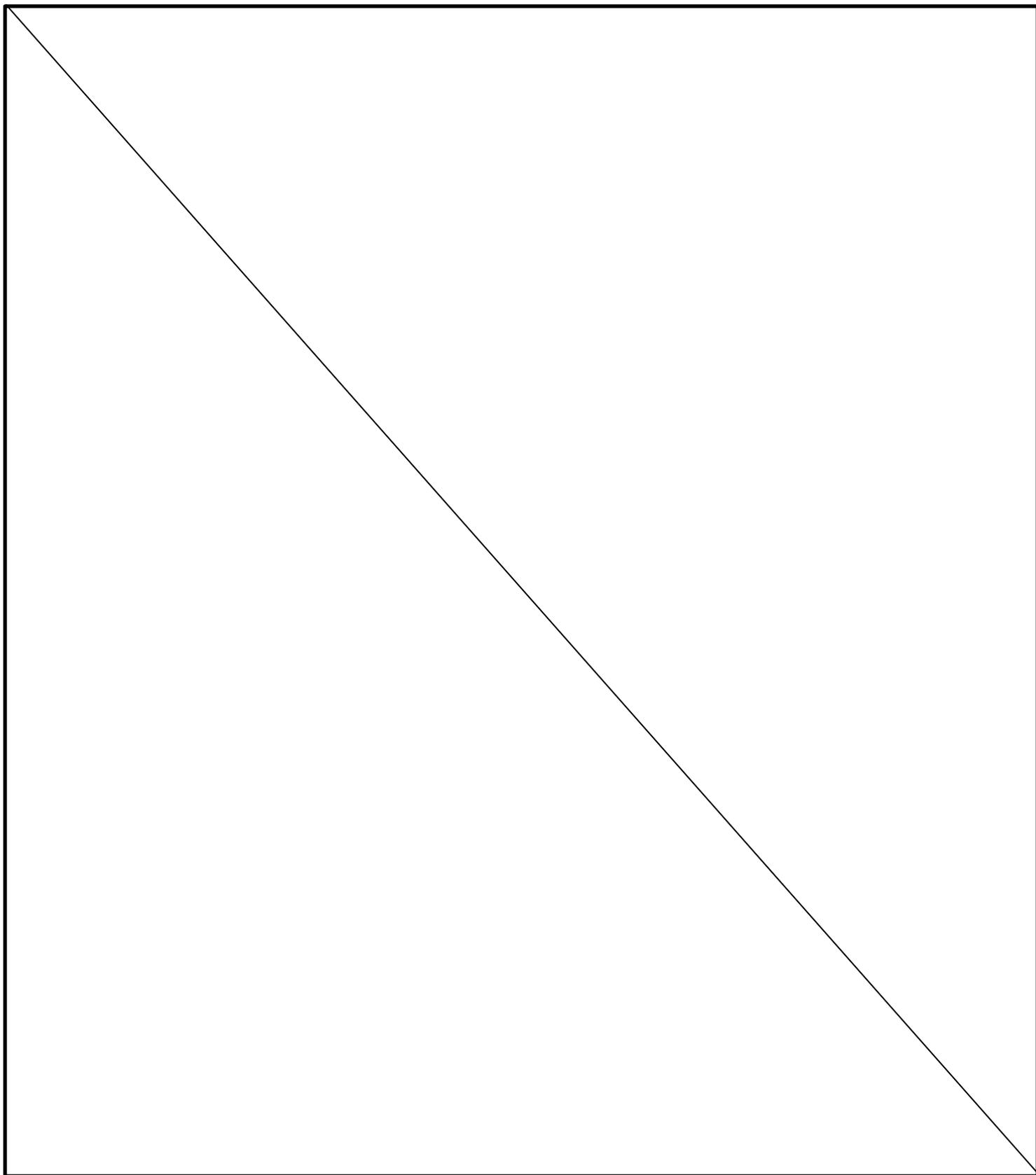
Annexe 8. Catégorie *diffusion*



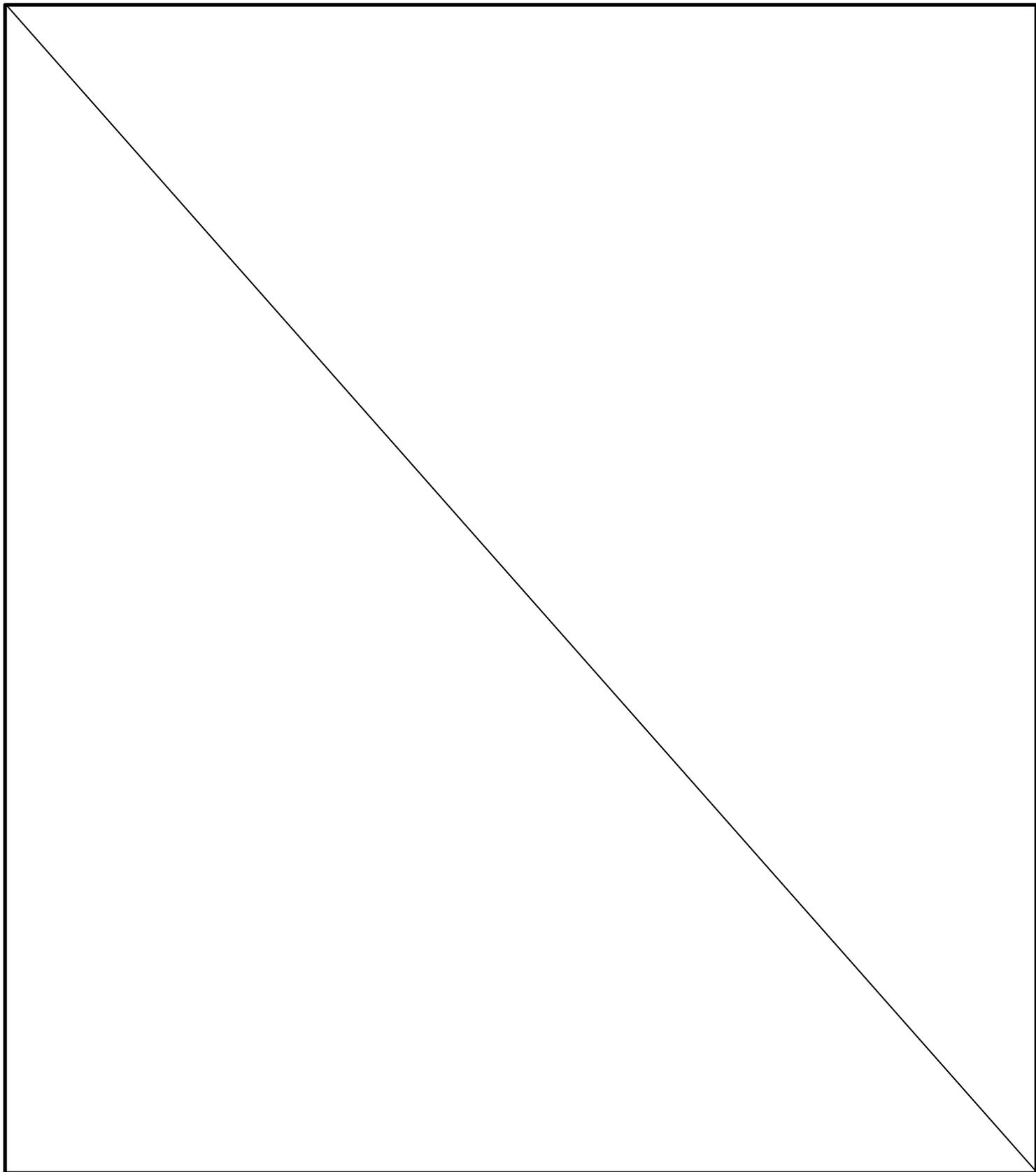
Annexe 9. Catégorie *personne(s)*



Annexe 10. Catégorie *santé*



Annexe 11. Catégorie *sport*



Annexe 12. Extrait du dendrogramme de classification du corpus

