Usage of non-conventional resources and contributive methods to bridge the terminological gap between languages by developing multilingual "preterminologies"

Mohammad Daoud

Université de Grenoble

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité: "INFORMATIQUE"

dans le cadre de

l'École Doctorale "Mathématiques, Sciences et Technologie de l'Information, Informatique"

(Doctoral School "Mathematics, Information Science and Technology, Informatics")

présentée et soutenue publiquement

par

Mohammad DAOUD

le 20 décembre 2010

Utilisation de ressources non conventionnelles et de méthodes contributives pour combler le fossé terminologique entre les langues en développant des "préterminologies" multilingues

Usage of non-conventional resources and contributive methods to bridge the terminological gap between languages by developing multilingual "preterminologies"

Thèse dirigée par

M. Christian BOITET Directeur de thèse

M. Kyo KAGEURA Codirecteur de thèse

M. Mathieu MANGEOT Codirecteur de thèse

JURY

M. Ahmed LBATH
Mme Violaine PRINCE
M. Ulrich HEID
M. Joseph DICHY
M. Fathi DEBILI
M. Christian BOITET
M. Kyo KAGEURA
M. Mathieu MANGEOT-NAGATA

Préparée au laboratoire GETALP–LIG (CNRS–INPG–UJF–UPMF)

Acknowledgements

This thesis would not have been possible without the support of many people. I wish to express my deepest gratitude to my supervisor Pr. Christian Boitet, for his support, patience and precious guidance, his encouragement and assistance enabled me to develop better research skills and a better understanding of my subject. I would like also to thank my co-supervisor Dr. Mathieu Mangeot, whose kind help and important advice improved the work on this thesis both at the technical and the theoretical levels.

I am heartily thankful to my co-supervisor Pr. Kyo Kageura (University of Tokyo), who welcomed me at his laboratory from 10/2009 to 9/2010. His invaluable experience in the domain of Terminology played an essential role since the beginning of my PhD. I thank him for his generosity and support. And I would like to thank all the members of his laboratory "Library and Information Science Laboratory", especially, Pr. Tony Hartley (University of Leeds) who was working there as a visiting professor during my stay. His experience in translation and computer-aided collaborative translation added an essential perspective to the thesis.

I also thank my brother Dr. Daoud Daoud, for introducing me to Pr. Christian Boitet in 2006 and for his constant support and encouragement. His advice and experience helped me in many levels. And I would like to thank him for providing me with the needed data to experiement with the domain of Arabic oneirology.

Cordial gratitude is also due to Pr. Asanobu Kitamoto (the National Institute of Informatics "NII", Tokyo) for giving me the chance to work at the NII for 13 months (two internships from 1/2008 to 8/2008 and 12/2008 to 6/2009) and for letting me experiment with the archive of the Digital Silk Road Project "DSR". I would also like to thank all the members of that promising project, especially Pr. Kinji Ono the leader of DSR.

I thank my colleagues at GETALP (Grenoble/France) for their support since 9/2006, especially Dr. Valérie Bellynck who co-supervised my master thesis (Master 2 of Research, UJF, 06-07): her advice and assistance during and after my master improved my research abilities and enriched my perspectives.

I would like to thank Dr. Mathieu Lafourcade for allowing me to localize JeuxDeMots into Arabic, and to experiment with it.

During the course of preparing my PhD thesis, I submitted articles to several conferences and journals, and the comments I received (even the harsh ones) helped me enhance my work and learn from my mistakes, particularly I would like to thank the anonymous reviewing committees of: ASLIB08, SNLP09, TALN09, RED09, LNCS-6162, COLING10, and LREC10.

I would like also to express my humble gratitude to the jury members, Ahmed Lbath (UJF), Violaine Prince (Montpellier), Ulrich Heid (Stuttgart & Hildesheim), Joseph Dichy (Lyon-II), and Fathi Debili (CNRS-LLACAN). I am honored that you accepted to report on my thesis and participate to the jury.

I thank my friends in Tokyo and Grenoble who made it easier for me to live abroad. And I thank my sisters, my brothers, my nieces and my nephews for their encouragements and support.

Finally, I would like to express my appreciation and gratitude to my parents who are giving me unconditional support and love.

Résumé

Notre motivation est de combler le fossé terminologique qui grandit avec la production massive de nouveaux concepts (50 quotidiens) dans divers domaines, pour lesquels les termes sont souvent inventés d'abord dans une certaine langue bien dotée, telle que l'anglais ou le français. Trouver des termes équivalents dans différentes langues est nécessaire pour de nombreuses applications, telles que la RI translingue et la TA. Cette tâche est très difficile, particulièrement pour certaines langues très utilisées telles que l'arabe, parce que (1) seule une petite proportion de nouveaux termes est correctement enregistrée par des terminologues, et pour peu de langues ; (2) des communautés spécifiques créent continuellement des termes équivalents sans les normaliser ni même les enregistrer (terminologie latente) ; (3) dans de nombreux cas, aucuns termes équivalents ne sont créés, formellement ou informellement (absence de terminologie).

Cette thèse propose de remplacer le but impossible de construire d'une manière continue une terminologie à jour, complète et de haute qualité pour un grand nombre de langues par celui de construire une preterminologie, en utilisant des méthodes non conventionnelles et des contributions passives ou actives par des communautés d'internautes : extraction de termes parallèles potentiels non seulement à partir de textes parallèles ou comparables, mais également à partir des logs (traces) des visites à des sites Web tels que DSR (Route de la Soie Digitale), et à partir de données produites par des jeux sérieux. Une préterminologie est un nouveau genre de ressource lexicale qui peut être facilement construit et a une bonne couverture.

Suivant en ceci une tendance croissante en lexicographie computationnelle et en TALN en général, nous représentons une préterminologie multilingue par une structure de graphe (Preterminological Multilingual Graph, MPG), où les nœuds portent des prétermes et les arcs des relations préterminologiques simples (synonymie monolingue, traduction, généralisation, spécialisation, etc.) qui sont des approximations des relations (terminologiques ou ontologiques) usuelles.

Un Système complet pour Éliciter une Préterminologie (SEPT) a été développé pour construire et maintenir des MPG. Des approches passives ont été expérimentées en développant un MPG pour le site Web culturel de DSR, et un autre pour le domaine de l'onirologie arabe : les ressources produites ont atteint une bonne couverture informationnelle et linguistique. L'approche indirecte par contribution active est testée depuis 8-9 mois sur l'instance arabe du jeu sérieux JeuxDeMots.

Abstract

Our motivation is to bridge the terminological gap that grows with the massive production of new concepts (50 daily) in various domains, for which terms are often first coined in some well-resourced language, such as English or French. Finding equivalent terms in different languages is necessary for many applications, such as CLIR and MT. This task is very difficult, especially for some widely used languages such as Arabic, because (1) only a small proportion of new terms is properly recorded by terminologists, and for few languages; (2) specific communities continuously create equivalent terms without normalizing and even recording them (latent terminology); (3) in many cases, no equivalent terms are created, formally or informally (absence of terminology).

This thesis proposes to replace the impossible goal of building in a continuous way an up-to-date, complete and high-quality terminology for a large number of languages by that of building a preterminology, using unconventional methods and passive or active contributions by communities of internauts: extracting potential parallel terms not only from parallel or comparable texts, but also from logs of visits to Web sites such as DSR (Digital Silk Road), and from data produced by serious games. A preterminology is a new kind of lexical resource that can be easily constructed and has good coverage.

Following a growing trend in computational lexicography and NLP in general, we represent a multilingual preterminology by a graph structure (Multilingual Preterminological Graph, MPG), where nodes bear preterms and arcs simple preterminological relations (monolingual synonymy, translation, generalization, specialization, etc.) that approximate usual terminological (or ontological) relations.

A complete System for Eliciting Preterminology (SEpT) has been developed to build and maintain MPGs. Passive approaches have been experimented by developing an MPG for the DSR cultural Web site, and another for the domain of Arabic oneirology: the produced resources achieved good informational and linguistic coverage. The indirect active contribution approach is being tested since 8-9 months using the Arabic instance of the JeuxDeMots serious game.

Table of Contents

List of Figures

List of Tables

Introduction

Construire des ressources terminologiques multilingues pour un domaine est une tâche compliquée et difficile, car la terminologie représente la structure conceptuelle d'un domaine en utilisant un ensemble d'unités lexicales dans une langue particulière. Une telle structure conceptuelle est plus dynamique et change plus vite que sa représentation symbolique (terminologie). Ce qui complique la tâche encore plus, c'est que la terminologie multilingue essaie de maintenir une représentation multilingue de cette structure conceptuelle, mais les ressources langagières ont tendance à varier en qualité et en quantité. En outre, ce ne sont pas toutes les communautés de différentes langues qui partagent le même intérêt pour le domaine. En outre, toutes les communautés de langues différentes ne partagent pas le même intérêt dans le domaine. Cependant, quand il s'agit de la terminologie des études islamiques, l'arabe a tendance à avoir de plus riches ressources terminologiques, en raison de l'intérêt de la communauté de langue arabe dans le domaine des études islamiques. Lors de la construction d'une ressource terminologique multilingue, on doit considérer la variété des ressources et des intérêts entre les langues.

L'étude des ressources multilingues et des langues avec des ressources pauvres (ou langues-pi, "langues mal équipées") est un sujet de recherche très actif au GETALP (Groupe d'Etude pour la Traduction Automatique / Traitement Automatisé des Langues et de la Parole, précédemment GETA) depuis 1980. Les expériences précédentes ont déjà élaboré un dictionnaire anglais-malais et un système prototype de TA anglais-thaï, avec de petits dictionnaires (1979-1986), suivie par le développement du dictionnaire FEM (français, anglais, malais) (Mangeot 1999), (Gaschler et Lafourcade 1994) et du dictionnaire FEV (français, anglais, vietnamien, 1990-96 et 1996-2003) (Vo-Trung, Phan et al. 2005). V. Berment a également travaillé sur des outils pour des langues d'Asie du sud-est, en particulier la langue lao (Berment 2004). Récemment, M. Mangeot a travaillé sur "Mot à Mot", un projet de construction d'une base lexicale français-vietnamien-khmer (Mangeot 2009). Il y a aussi de la recherche sur la reconnaissance automatique de la parole pour le khmer et le vietnamien, et la translittération des langues indo-pakistanaises (Malik, Boitet et al. 2008).

En ce qui concerne les techniques, le développement de ressources et d'outils pour les ressources linguistiques multilingues est un thème de recherche important au GETALP ; G. Sérasset et plus tard M. Mangeot ont travaillé sur la conception et le développement de Papillon (Boitet, Mangeot et al. 2002) (Papillon 2010) (CDM et Nadia) qui est une base de données lexicales contributive ; V. Bellynck a expérimenté des approches collaboratives dans le développement de ressources lexicales par ITOLDU (Bellynck, Boitet et al. 2005). V. Archer a étudié le traitement automatique des grands graphes linguistiques (Archer 2009).

D'après les expériences précédentes, il est évident qu'il y a un besoin de ressources terminologiques, même en présence de ressources lexicales à usage général, pour une variété d'applications et de demandes, comme la traduction automatique, la Recherche d'Information Translingue, la localisation de logiciels, et l'éducation. Ce qui augmente le besoin de ces ressources est la grand "écart terminologique" entre les langues, car les langues parlées par les communautés actives dans le domaine développent et enregistrent leur terminologie en douceur, tandis que d'autres langues ont des difficultés dans le développement et la documentation appropriée d'une telle terminologie.

Cette thèse aborde et identifie les problèmes et les enjeux dans la construction de ressources terminologiques multilingues. Elle analyse les approches actuelles de la terminologie, et propose un nouveau scénario pour la définition, la structuration et le développement de ressources terminologiques dédiées à un domaine.

Par définition, un terme est une unité lexicale d'une langue, utilisée pour désigner et étiqueter un concept dans un domaine particulier. Rassembler de la terminologie est l'étude des méthodes pour

recueillir ces unités lexicales. Comme le processus d'invention de concepts est sans fin, la terminologie n'est pas fixe (Kageura 2002).

Traditionnellement, les terminologues sont chargés d\'élaborer la terminologie. Leur travail consiste à construire les correspondances entre la représentation symbolique, le concept et les informations terminologiques qui situent le terme dans la sphère terminologique (information comme les définitions). Construire de telles correspondances est une tâche épuisante qui influe sur le coût et la couverture. D'autre part, les approches de collaboration essayent de remplacer les terminologues par des bénévoles amateurs, ce qui est une tendance prometteuse dans l'acquisition de connaissances. Toutefois, les bénévoles peuvent ne pas être en mesure d'effectuer une tâche linguistique, même s'ils sont réellement familiers avec le domaine. Les approches automatiques utilisent des ressources textuelles et lexicales pour développer une terminologie, mais elles sont limitées par les ressources disponibles et par la technique utilisée. Aussi, ce ne sont pas toutes les connaissances terminologiques qui sont disponibles en format textuel (terminologie latente). Enfin, les approches automatiques peuvent être efficaces pour trouver des termes, mais il est difficile de construire les correspondances entre ces termes.

Cette thèse suggère le développement de ressources préliminaires de terminologie, avec des correspondances lâches et faciles ((à calculer)), et d'approches non conventionnelles, appelées préterminologies, qui visent la terminologie latente et absente en augmentant la couverture des unités lexicales et le renforcement des correspondances validées en se basant sur une structure de graphe (graphes terminologiques multilingues) ; les méthodes de construction d'une préterminologie dépendent de la communauté du domaine. La communauté peut contribuer *passivement* à construire la préterminologie, en analysant ses ressources produites, et elle peut y contribuer *activement* en étant directement impliquée dans le processus de contribution.

Des scénarios et des expériences à contribution active et passive seront montrés à la fin de cette thèse, en particulier pour le développement d'une préterminologie multilingue pour les ressources culturelles archivées dans le site Web Digital Silk Road (Ono, Kitamoto et al. 2008).

Cette thèse est organisée comme suit. Le chapitre I donne des bases sur la terminologie multilingue du point de vue informatique, ses ressources, son importance, et établit un critère d'évaluation pour les ressources terminologiques. Le chapitre II présente l'état de l'art sur le développement de la terminologie multilingue, sur les systèmes actuels, et sur leurs limites. Le chapitre III définit et démontre le concept de "préterminologie" comme une proposition de solution.

Le chapitre IV introduit la notion de graphe multilingue préterminologique ; il est suivi par le chapitre V, qui montre les techniques et approches dans l'élaboration de ces graphes, et le chapitre VI présente la création d'un système pour développer et maintenir une préterminologie (SEpT).

Le chapitre VII explique les aspects techniques de SEpT, son instanciation et son expérimentation. Le chapitre VIII présente des expériences sur le développement de préterminologies grâce à des approches passives, et le chapitre IX montre le potentiel des approches à contribution active. Enfin, une conclusion et des commentaires seront présentées à la fin de la thèse.

Thesis Introduction

Constructing multilingual terminological resources for a domain is a difficult and complicated task, because terminology represents the conceptual structure of a domain using a set of lexical units in a particular language. Such a conceptual structure is more dynamic and changes faster than its symbolic representation (terminology). What complicates the task even more is that multilingual terminology tries to maintain a multilingual representation of that conceptual structure; however, resources of languages tend to vary in quality and quantity. Furthermore, not all communities from different languages share the same interest in the domain. For example, Arabic is considered as a poorly resourced language (Nikkhou and Choukri 2005) (Yassin 2003) (Diab and Habash 2009) (Ghazal 1977), especially in the domains of science and technology, while English and French have a much richer terminology in these domains. However, when it comes to the terminology of Islamic studies, Arabic tends to have richer terminological resources, because of the interest of the Arabic speaking community in the domain of Islamic studies. When building a multilingual terminological resource, one should consider the variety of resources and interest between the languages.

The study of multilingual resources and languages with poor resources (or pi-languages, "poorly equipped languages") is an active topic of research at GETALP (Groupe d'Étude pour la Traduction Automatique/le Traitement Automatisé des Langues et de la Parole, previously GETA) since 1980. The earlier experiments developed an English-Malay and an English-Thai prototype MT system, with small dictionaries (1979-1986), followed by the development of the FEM (French, English, Malay) dictionary (Mangeot 1999), (Gaschler and Lafourcade 1994) and FEV (Vo-Trung, Phan et al. 2005) (French, English, Vietnamese) dictionaries (1990-96 and 1996…2003). V. Berment also worked on tools for South-Asian languages, especially the Lao language (Berment 2004). Recently, M. Mangeot worked on "Mot à Mot" project to build a French-Vietnamese-Khmer lexical base (Mangeot 2009). This is beside the research on the automatic speech recognition of Khmer and Vietnamese, and the transliteration of indo-pakistani languages (Malik, Boitet et al. 2008).

In terms of the techniques, the development of resources and tools for multilingual linguistic resources is an important research theme at GETALP, G. Sérasset and later M. Mangeot worked on the design and development of Papillon (Boitet, Mangeot et al. 2002) (Papillon 2010) (CDM and Nadia) which is a contributive lexical database, V. Bellynck experimented collaborative approaches in developing lexical resources through ITOLDU (Bellynck, Boitet et al. 2005). V. Archer studied the automatic processing of large linguistic graphs (Archer 2009).

From previous experiments, it is apparent that there is a need for terminological resources, even in the presence of general purpose lexical resources, for a variety of applications and demands, like machine translation, Cross-Lingual Information Retrieval, localization, and education. What increases the need for such resources is the wide "terminological gap" between languages, where languages spoken by active communities in the domain are developing and recording its terminology smoothly, while other languages are having difficulties in the development and proper documentation of such terminology.

This thesis tackles and identifies the problems and issues in constructing multilingual terminological resources. It analyses the current approaches to terminology, and proposes a new scheme in defining, structuring and developing domain-dedicated terminological resources.

By definition, a term is a lexical unit of a language used to denote and label a concept in a special domain. Gathering terminology is the study of the methods to collect those lexical units. As the process of finding concepts is endless, terminology is not fixed (Kageura 2002).

Traditionally, terminologists are responsible for developing terminology. Their work involves building the correspondences between the symbolic representation, the concept and the terminological information to situate the term in the terminological sphere (information like definitions). Building such correspondences is an exhausting task which affects the cost and

coverage. On the other hand, collaborative approaches try to replace terminologists with amateur volunteers, which is a promising trend in acquiring knowledge. However, volunteers might not be able to conduct linguistic task, even if they are actually familiar with the domain. Automatic approaches use textual and lexical resources to develop terminology, but are limited to the available resources and the used technique. Also, not all terminological knowledge is available in textual format (latent terminology). Finally, automatic approaches might be effective in finding terms, but it is difficult to build the correspondences between those terms.

This thesis suggests the development of preliminary resources for terminology, with easy loose correspondences and unconventional approaches, called preterminology, which targets latent and absent terminology through increasing the coverage of lexical units and building bot validated correspondences based on a graph structure (Multilingual Terminological Graphs), the methods in constructing preterminology depends on the community of the domain. The community can *passively* contribute to preterminology by analyzing its produced resources, and it can *actively* contribute by being directly involved in the contribution process.

Active and passive contribution scenarios and experiments will be shown at the end of this thesis, particularly for the development of multilingual preterminology of the archived resources of the Digital Silk Road cultural web site (Ono, Kitamoto et al. 2008).

This thesis is organized as follows. Chapter I gives a background about multilingual terminology from the computational point of view, its resources, its importance, and establishes an evaluation criteria for terminological resources. Chapter II presents the state of the art in developing multilingual terminology, the current systems and their limitations. Chapter III defines and shows "preterminology" as a suggested solution.

Chapter IV introduces multilingual preterminological graph, followed by chapter V which shows the techniques and approaches in developing these graphs, and chapter VI shows the design of a system to develop and maintain preterminology (SEpT).

Chapter VII explains the technical aspects of SEpT, its instantiation and experimentation. Chapter VIII shows the experiments on developing preterminology through passive approaches, and chapter IX shows the potential of active contributions approaches. Finally, some conclusion and remarks will be drawn at the end of the thesis.

Part A Terminology vs. Preterminology

Introduction à la partie A « Terminologie et préterminologie »

Comme le processus d'invention de concepts est éternel, la terminologie n'est pas fixe. En fait, il s'agit d'une production massive d'étiquettes pour de nouveaux concepts, et ces «étiquettes» sont généralement exprimés dans la langue de la communauté qui a produit le concept. Il arrive que, la plupart du temps, cette langue est l'une des langues bien dotées en ressources, comme l'anglais, l'allemand, le français … Cela crée un fossé terminologique entre les langues. Certains chercheurs soutiennent même que le processus de production de l'étiquette (du "terme") n'est pas séparé du processus de production du concept, ce qui rend le processus de comblage de ce fossé encore plus compliqué. La partie A explique ce problème plus en détail.

Pour s'attaquer au problème du fossé terminologique multilingue, dans ce chapitre, nous présenterons le contexte théorique de la terminologie, et ses approches computationnelles, puis nous identifierons les problèmes de terminologie multilingue que nous traitons, avec notamment le problème, dans les systèmes terminologiques actuels, causé par l'absence de terminologie et aussi par la terminologie cachée.

Après avoir identifié les problèmes de la thèse, nous donnerons une présentation panoramique sur les approches actuelles au développement de la terminologie multilingue. Nous avons classé les approches en : classiques, automatiques, et collaboratives. Chaque approche tente de se conformer au système terminologique actuel traditionnel, et cela provoque une série de limites, en termes de coût, de couverture et de qualité.

Nous présenterons ensuite les possibilités qui peuvent être utilisées dans le développement de la terminologie, et nous proposerons un système terminologique (pour la préterminologie) qui peut facilement exploiter ces possibilités.

Cette partie est organisée comme suit. Le chapitre I donne des bases sur la terminologie multilingue du point de vue informatique, ses ressources, et son importance. En outre, il établit un critère d'évaluation des ressources terminologiques. À la fin de ce chapitre, nous définirons les problèmes de terminologie multilingue que nous nous attaquons. Le chapitre II présente l'état de l'art sur le développement de la terminologie multilingue, sur les systèmes actuels, et sur leurs limites. Le chapitre III définit ce qu'est une "préterminologie" et montre pourquoi elle peut conduire à une solution.

Introduction to Part A

As the process of finding concepts is everlasting, terminology is not fixed. In fact there is a massive production of labels to new concepts, and these “labels” are usually worded using the language of the community that produced the concept. It happens that most of the time this language is one of the well-resourced languages, like English, German, French... This creates a terminological gap between languages. Even some researchers argue that the process of producing the label “term” is not separated from the process of producing the concept, which makes the process of bridging the gap even more complicated. Part A explains this problem in more details.

To tackle the problem of multilingual terminological gap, in this chapter, we will present the theoretical background of terminology, and its computational approaches, and then we will identify the problems of multilingual terminology that we are dealing with, in particular the problem in the current terminological systems caused by the absence of terminology and even hidden terminology.

After identifying the problems of the thesis, we will give a panoramic presentation of the current approaches in developing multilingual terminology. We classified the approaches into: Classical, Automatic, and Collaborative. Each approach tries to comply with the current traditional terminological system, and that causes a variety of limitations, in terms of the cost, the coverage and the quality.

We will then present the opportunities that can be used in developing terminology, and we will propose a terminological system (for preterminology) that can easily exploit these opportunities.

This part is organized as follows, Chapter I gives a background about multilingual terminology from the computational point of view, its resources, and its importance. Furthermore, it establishes an evaluation criterion for terminological resources. At the end of this chapter, we will define the problems of multilingual terminology that we are tackling. Chapter II presents the state of the art in developing multilingual terminology, the current systems and their limitations. Chapter III defines "preterminology" and shows why it can lead to a solution.

Chapter I Multilingual Terminological Resources: Problems and Motivations

Introduction

Beside its crucial importance in human communication, terminology is essential for a variety of NLP applications. Any successful machine translation system should have a multilingual resource rich in domain-specific terms (Kübler 2002). To understand the difficulties of enriching multilingual terminological resources, we will start with an introduction to terminology from the computational linguistic point of view.

When a new idea is born, humans perceive it as a concept, and use a sign (term) for the purpose of communication. Ideas emerge endlessly, and the way we perceive them as well. That is why terminology is dynamic, which makes it challenging to maintain a repository of multilingual "signs" corresponding to a domain. Another reason is that not all the languages are equally involved in the process of producing the concepts of that domain.

In this chapter, we give an overview of the classical approaches in Terminology, and we focus on the "multilingualism" of terms as it is a fundamental point of this thesis. And then we will show the problems of the terminological system that create lexical gaps between languages.

This chapter is organized as follows. The next section introduces multilingual terminology, its structure and resources. Section I.2 focuses on the importance of multilingual terminology and its applications. Section I.3 presents a criterion to judge a multilingual terminological repository. Finally, Section I.4 summarizes the main problems of the current multilingual terminological systems.

I.1 Multilingual Terminology

I.1.1 Terminology Science and Technology

Specialists in a field use a special vocabulary which is not commonly used in other fields. This vocabulary constitutes a "special language", an average person might call it jargon, or lingo. Specialists in medicine, agriculture, and engineering… have their own jargon that they use to describe their ideas and concepts.

The following paragraph is a quotation from an article in Wikipedia about the C++ programming language.

*"**C++** (pronounced see plus plus) is a **statically typed**, **free-form**, **multi-paradigm**, **compiled**, **general-purpose programming language**. It is regarded as a "**middle-level**" language, as it comprises a combination of both **high-level** and **low-level language features**."*

It is very difficult for a normal English speaking person who has little knowledge in computers to understand the above text, because the words and expressions in bold have special meanings exclusively comprehensible by computer scientists, or people who are familiar with computer programming. In linguistics, the study of the vocabulary of such a special language is called Terminology (Rey 1995).

I.1.1.1 General Introduction and Definitions

The following sub-sections provide definitions for essential entities and concepts that we will deal with in this thesis.

a. Terms

A term is the primary object of Terminology ("T" for terminology as a science). Traditionally, it is described as "a sign to describe a concept" (Sager 1990). According to (Kageura 2002) a sign is a lexical unit that corresponds to one or more words. And according to the British Standards Institution (B.S.I. 1963) "concepts are mental constructs, abstractions". Sager gives a more general definition: "a concept is any unit of thought".

The traditional school of Terminology argues that a concept comes to existence before a term represents it, which is widely accepted within the traditional schools of Terminology. However, (Temmerman 2000) argues against that. But what is clear is that a concept is not useful socially without a term that describes it.

b. Terminology

A terminology is the vocabulary (set of signs) of a domain. (Note again that the spelling Terminology is used to denote the discipline of Terminology, which is the study of terms).

c. Terminologists

A terminologist is a person involved in the formation, interpretation, and organization of terms in a domain. The terminologist forms a term by identifying a concept and finding its corresponding sign. S/He interprets a term (that already has been constructed) by giving a definition to its concept, and s/he organizes the terms by structuring them in a dictionary format. A terminologist is not necessarily interested in a language in particular.

d. Domain

A domain is a subject field like medicine, law, and engineering that has its own set of concepts, and thus its own set of terms to represent these concepts.

e. Community

A set of persons who use the terminology of a domain, and responsible directly or indirectly of its evolution.

I.1.1.2 Terminology Concept System

In order to know how a term describes a thought, the semiotic triangle has been adopted. Figure 1 shows a semiotic triangle (also known as the triangle of reference or the triangle of meaning (Ogden and Richards 1923) which was influenced by Firdinand de Saussure who founded the concept of the sign/signifier/signified/referent forms) that describes the relation between the thought (object) in the real world, its perception (concept) and its linguistic representation (sign, term).

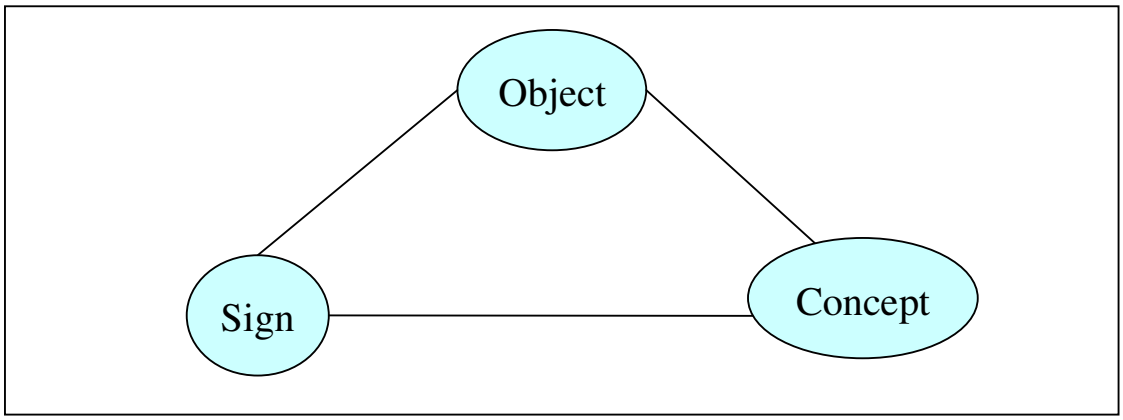


Fig. 1: Triangle of meaning

In a domain, objects are organized according to the way we perceive them (concepts). Within that domain concepts are structured with defined relations amongst them. Consequently, terminology should capture this structure.

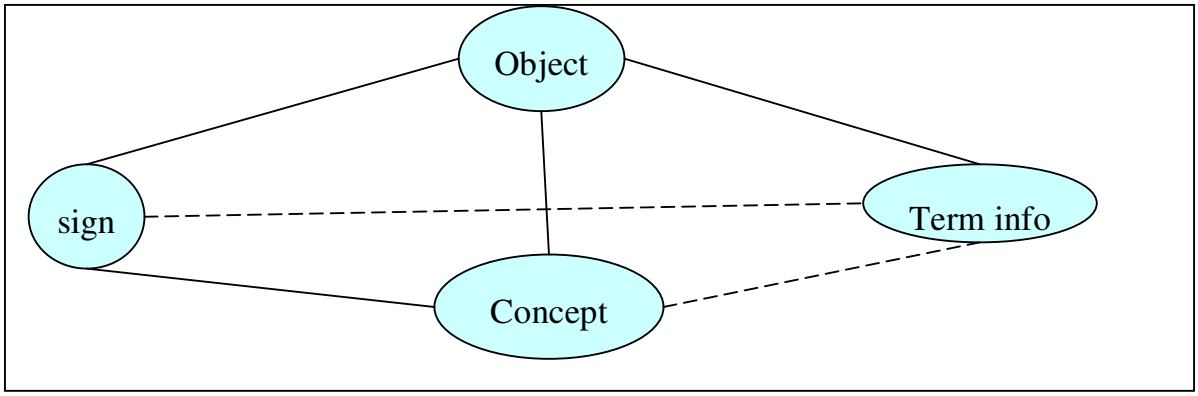


Fig. 2: Modified triangle of meaning

The semiotic triangle was not enough to exactly describe terminology, as terminologists believed that terms are not merely linguistic signs, but terminological information is needed. Specifically, a term definition that relies on already defined concepts is needed, see figure 2. This way, a term is produced and defined in a concrete manner, based on the already existing and defined terms and concepts. For example the definition of C++ (shown in I.1.1) relies on the definitions of other (already defined) concepts written in bold.

I.1.1.3 Lexicography and Terminology

It is important to emphasize the difference between a word and a term. A term may consist of one or more words, but it should uniquely represent one concept within the controlled domain. Terminology deals with terms, while lexicography deals with the analysis and compilation of the vocabulary (lexicon) of a language (Dubuc 1992). The major differences are shown in table 1.

Table 1: Terminology vs. Lexicography

	Lexicography	**Terminology**
Atomic unit	Word	Term (representation of a concept)
Descriptive information	Grammatical information, suggested usage	Definition, context, prescribed usage
Sphere	Vocabulary of a general language	Vocabulary of a domain-specific special language
Organization	Alphabetical order	Concept structure

These differences illustrate the distinction between a lexicon, a terminological repository, and, more relevant to this thesis, a multilingual dictionary, and a multilingual terminological database.

I.1.2 **Multilingualism**

I.1.2.1 Modified Concept System

a. *Multilingual Terminology*

Concepts are language-independent, which is equivalent to say that all domain specialists with different linguistic backgrounds share (or are supposed to share) the same concepts, but they represent them using different terms (from different languages).

However in reality, domain specialists might need a lingua franca or a pivot language to communicate with each other. For example, in Islamic studies, Arabic is the base for communication. In international conferences on computer science, English is used as a pivot language. One of the main characteristics of a pivot language is to be understood by domain specialists, and to have widely adopted terms to describe the concepts of that domain. Furthermore, some of the concepts may remain exclusively described using terms of one language.

The process of the lexical translation of the terms into new languages is called terminology multilingualization. Thus, if terminology is concerned with the terms and its relations with the concepts and amongst each other, multilingual terminology extends that interest to the relations between terms and their lexical translations.

b. *Multilingual Concept System*

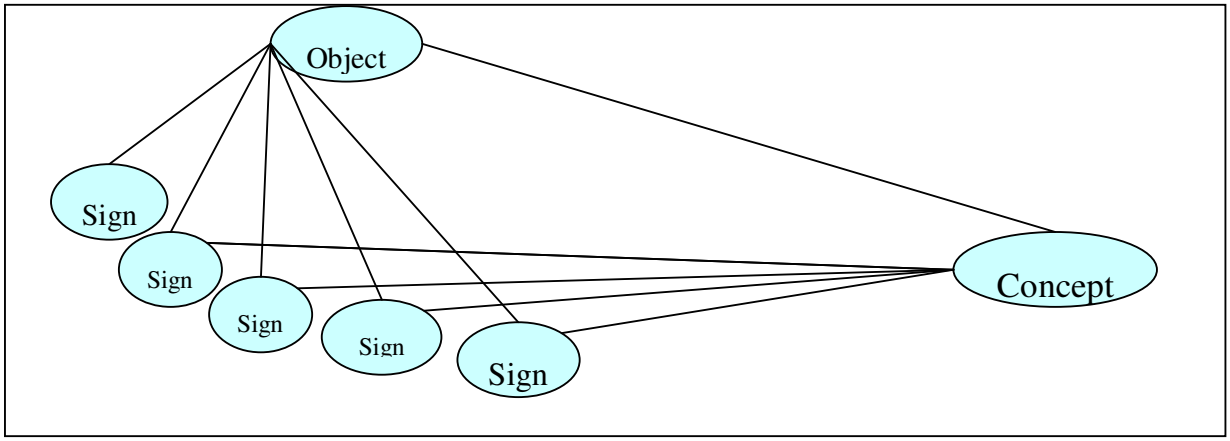


Fig. 3: Multilingual triangle of meaning

Having extended the interest of terminology, we need to extend the semiotic triangle to have more signs (in various languages) to represent the same concepts, as shown in figure 3.

I.1.2.2 Challenges

In reality, concept founders coin terms in their own language (and in the lingua franca, if different), as shown on figure 4, so there is a need to translate the term into the other languages.

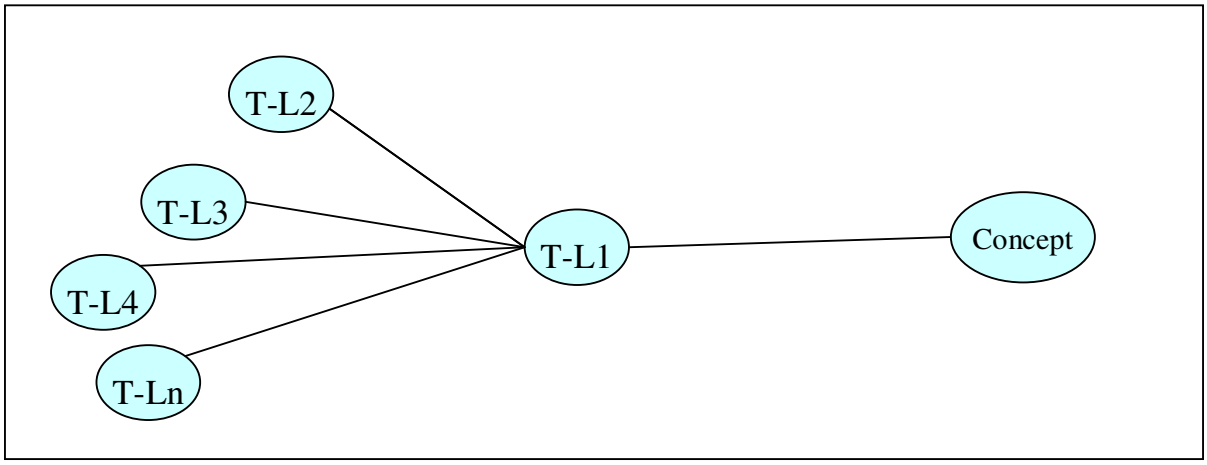


Fig. 4: Terminology production through term translation

What complicates the multilingualization process is that the term is not a sign only (according to the extended semiotic triangle) but it needs descriptive information, meaning that terminology multilingualization is not only concerned with lexical translation, but the introduction of a concept and its sign to different languages.

I.1.3 **Importance and Applications of Multilingual Terminology**

This subsection illustrates the importance of multilingual terminology nowadays in different fields and applications.

1) Communication:
   a) Scientific: a high percentage of the text of scientific articles consists of domain-specific terms, so that misunderstanding terms will definitely lead to knowledge loss.
   b) Media: whether it is written or non written, terminology is essential for classical media services. And with the new alternative media (blogs, forums, twitter…), terminology became more dynamic, and its users became uncontrolled.
2) Knowledge transfer: creating a multilingual terminological resource is the first step towards transferring knowledge in different languages, and also for localizing knowledge, whether it is software localization (Mozilla 2006) (Sun-Microsystems 2007) or content localization (Bernardi, Bocsak et al. 2005). For example, the main problem in enriching the non-English Wikipedias is the lack of multilingual terms, which is a problem that a Wikipedia contributor might not resolve.
3) Accessing knowledge: any CLIR (Cross-Lingual Information Retrieval) model depends on a multilingual resource rich with terminology.

I.2 Terminology Management Systems

A terminology management system is a software that provides facilities to store, manage, and maintain terminological data (Schmitz 2006). It is usually centralized by a database (a terminological database). This subsection explains the main features and requirements for maintaining terminology, and most importantly the challenges of multilingual terminology.

I.2.1 Terminological Databases

Also known as term base or term bank, a terminological database is a set of information about terms and concepts of a domain, electronically stored, and structured to capture the actual conceptual relations.

I.2.1.1 Structure

There are two main areas where a termbase needs a defined structure:

1. The structure of the relation of terms, when one is referring to another, *Macrostructure*.
2. The specific structure of each terminological entry, its *Microstructure*. This deals with the term and its descriptive information needed for each term.

a. *Macrostructure*

Usually, a macrostructure of a term base is concept-oriented, so that the key in the organization structure is not the orthography, but the semantics of the term and its related terms (the position in the semantic net).

Terms can be related by two clusters of relations (Trippel 1999):

- Semantic coordinating relations: synonym, antonym…
- Ontological relations: ISA, PART OF relations…

A termbase should illustrate these relations in the macrostructure level.

b. *Microstructure*

The microstructure is the structure of a terminological entry in the termbase, such microstructure contains the following information:

- the orthographic structure of the term;
- the descriptive information;
- the conceptual level (link to macrostructure).

I.2.1.2 Functionalities

The main functionalities implemented in a termbase from the point of view of a user are as follows.

- Term search and retrieval: either concept-based or orthography-based retrieval, depending on the need and the available data.
- Receiving feedback: by rating and editing the entries.

From the point of view of an administrator:

- Macrostructure modification.
- Microstructure modification.

- Term creation, deletion, updating and manipulation.
- Import and export terminology.

I.2.1.3 Bilingual and Multilingual Term Banks

The structure of a multilingual term base is more complicated, as many information at the microstructure level might not be useable for all languages. Hence, there is a risk of inconsistencies in the information carried at each monolingual microstructure level. Linking a term with its translations is a problem.

G. Sérasset (Sérasset 1994) (Sérasset, Brunet-Manquat et al. 2006) suggests the usage of multilingual acceptions to represent the equivalence links between terms and their lexical translations, see figure 5.

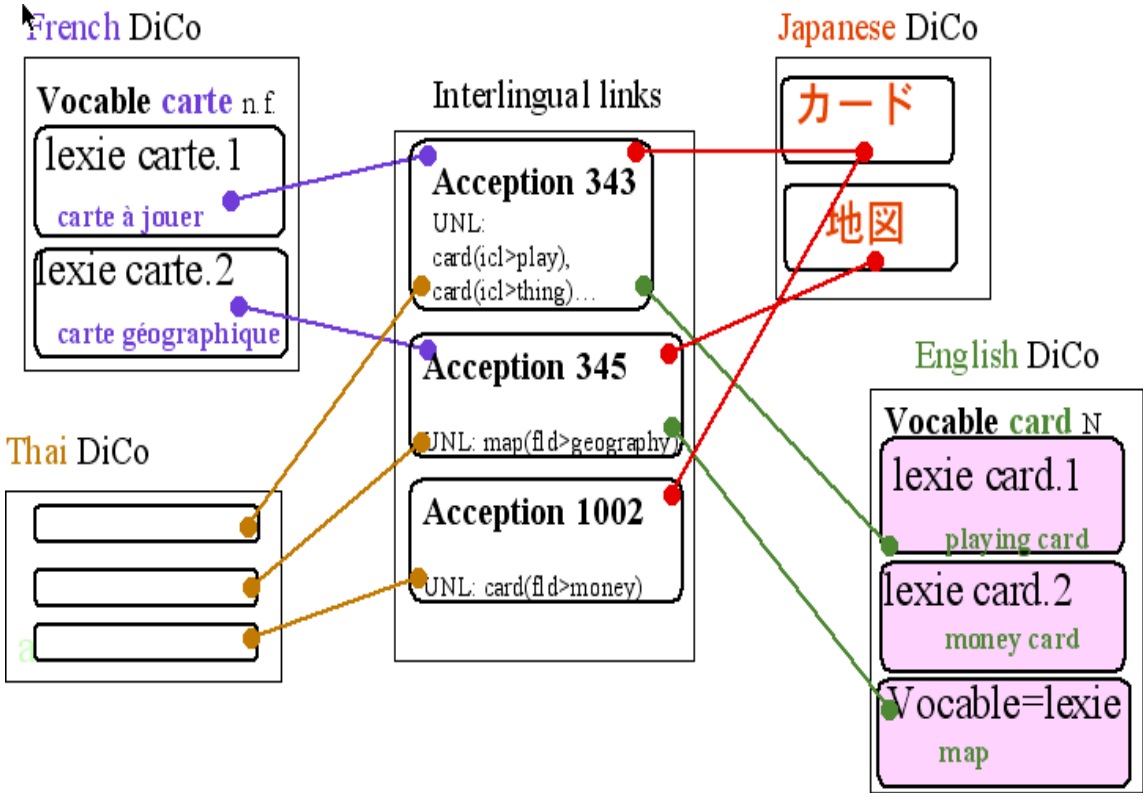


Fig. 5: Interlingual links in a Jibiki-based term base

These acceptions link lexies (microstructures) from different dictionaries. This macrostructure has been adopted in the Jibiki platform which is an online generic dictionary development platform (Mangeot 2001) (Sérasset 2004).

I.2.2 **Examples of Terminological Databases**

I.2.2.1 IATE

IATE (IATE 2008) is the EU multilingual term base; it has 8.4 millions terms, covering all the 23 EU official languages. It is a compilation of previous term bases, notably, EURODICAUTOM and EUTERPE.

The terms are categorized into 21 domains and tens of sub-domains, like information technology, law, agriculture…

IATE offers also a description, and references to the term. It also provides a reliability measure for each term. (See figure 6).

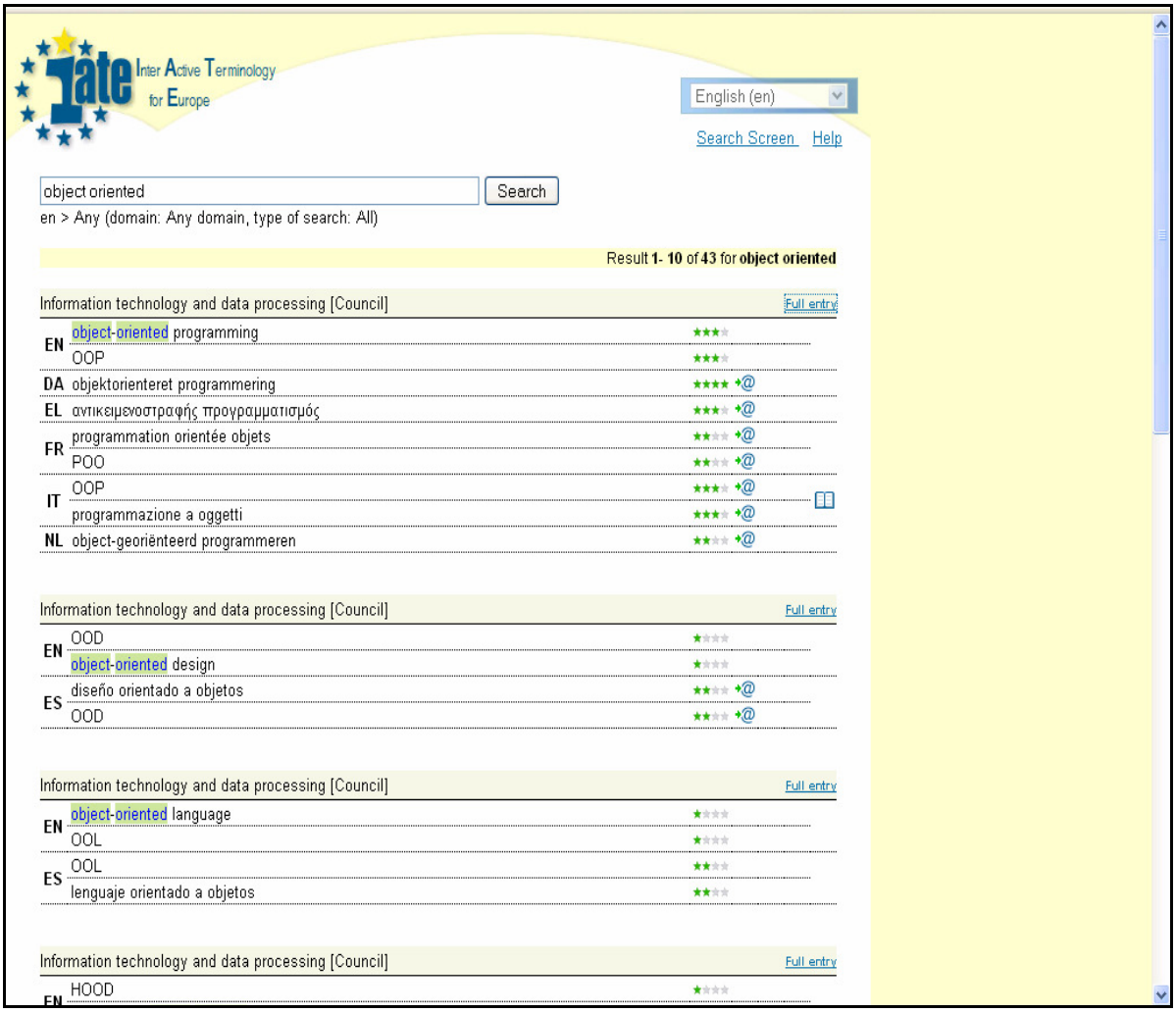


Fig. 6: IATE result screen

I.2.2.2 UNTERM

UNTerm (UN 2008) is a multilingual terminological database dedicated to the United Nations demands. It has 80,000 terms in 6 languages (the 6 UN official languages), the structure of the data is simple, and the platform provides advanced searching techniques.

I.2.2.3 Electropedia

Electropedia (IEC 2008) is an online electrical and electronic terminology database; it is produced by the International Electrotechnical Commission. The database contains more than 20,000 terms in 9 languages, categorized into 75 categories.

I.2.2.4 The GDT (Grand Dictionnaire Terminologique)

It is a terminological database for three languages (French, English, and Latin) produced by the Quebec board of the French language (Office Québécois de la Langue Française). It contains more than 3 million terms classified into 200 categories (OQLF 2008).

Formerly the dictionary was available in various formats; currently it is only available on the Web.

I.2.2.5 FAOTerm

Produced by the FAO (Food and Agriculture Organization) (FAO 2008), FAOTerm contains 58,000 terms in 7 languages. Figure 7 shows a screenshot.

Fig. 7: FAOTerm

I.2.2.6 LexALP

LexALP (EURAC 2007) is a Jibiki-based (Mangeot 2006) term bank dedicated to Spatial Planning and Sustainable Development in the framework of the Alpine convention. The bank provides terminology in four languages, German, French, Italian and Slovene (Sérasset, Brunet-Manquat et al. 2006). The project serves terminologies for 6 national legal systems (Austria, France, Germany, Italy, Slovenia and Switzerland). Terminology work is carried out by team of terminologists representing the different legal systems and languages. Terminologists manipulate terms through an online Jibiki environment.

I.3 Informational and Linguistic Coverage in Multilingual Terminological Resources

I.3.1 Dynamics of the Multilingual Terminological Sphere

I.3.1.1 Terminological Sphere

a. Lexical Sphere

The lexical sphere of a language is the set of all the meaningful lexical units of that language. A lexical unit may have one or more words, but it is not as complex as a phraseological unit which combines more than one concept to express a complex situation.

To give an idea about the size of this sphere, an average bilingual pocket dictionary has around 50,000 entries (some of them might include phraseological units). But what about the total number of lexical units in a well-resourced language like English? (Oxford dictionaries 2010) has tried to answer this question:

"*How many words are there in the English language?*

There is no single sensible answer to this question. It is impossible to count the number of words in a language, because it is so hard to decide what counts as a word. […]

The Second Edition of the Oxford English Dictionary contains full entries for 171,476 words in current use, and 47,156 obsolete words. To this may be added around 9,500 derivative words included as subentries. Over half of these words are nouns, about a quarter adjectives, and about a seventh verbs; the rest is made up of interjections, conjunctions, prepositions, suffixes, etc. These figures take no account of entries with senses for different parts of speech (such as noun and adjective). This suggests that there are, at the very least, a quarter of a million distinct English words, excluding inflections, and words from technical and regional vocabulary not covered by the OED, or words not yet added to the published dictionary, of which perhaps 20 per cent are no longer in current use. If distinct senses were counted, the total would probably approach three quarters of a million. "

English WordNet V3 (Miller 1995) has around 150,000 words. Note that some dictionaries may have hundreds of thousands or even millions of entries. For example, the ATLAS-II v.14 MT system (Fujitsu 2009) has 5.7 million entries in its technical dictionaries. Another example is the "Honyaku no osama" MT system (IBM 2010) which has several hundreds of thousands of entries in its terminological base. However, a dictionary entry is not always a word, it might represent a combination of lexical units, or even phraseological units. Hence, although there seems to be a reasonable number of words in a language, the lexical sphere extends these words to include lexical units made of several words. And the terminological sphere of a domain is a subset of the lexical sphere.

b. Lexical Sphere in Relation with Concepts

In a domain where concepts are dynamic and alive, every year thousands of new concepts are introduced, and thousands of lexical units are chosen to represent them. Figure 8 shows a few terms from the conceptual sphere of the domain of programming languages. The figure features 3 interrelated concepts. On the other side, the words of English form the general purpose lexical sphere. The terminological sphere is a part of the lexical sphere.

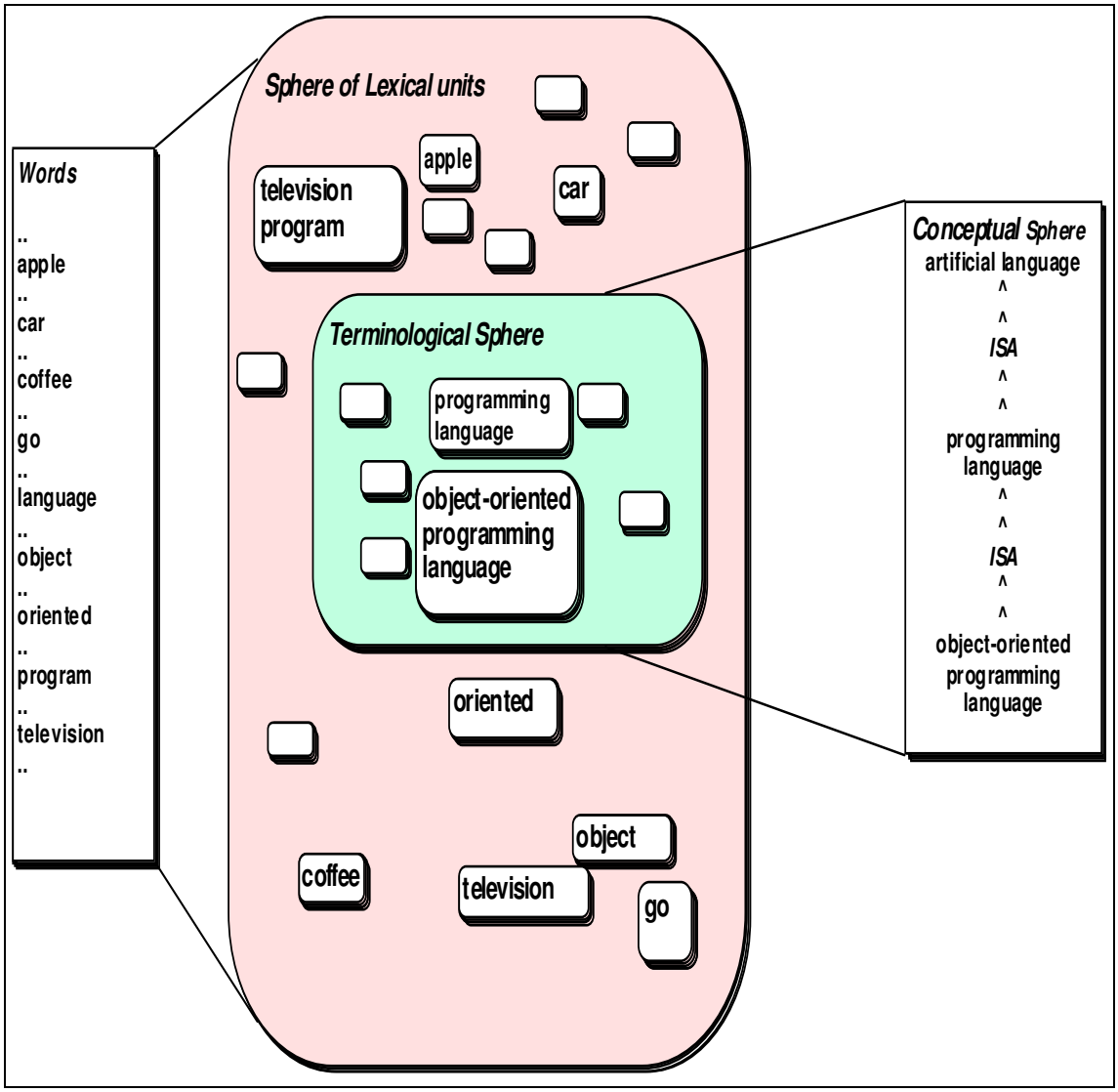


Fig. 8: Terminological sphere vs. lexical sphere

I.3.1.2 Multilingual Terminological Sphere

The multilingual terminological sphere of a domain is the union of the monolingual terminological spheres that correspond to the conceptual sphere of the domain.

The traditional terminology school teaches that concepts are our perception of the real world, and this perception is represented through communicative signs. Thus, in theory, a term is language-independent. However, in reality, it depends on the perspective of the person who deals with the term:

1. Term developer: s/he interacts with the object in the real world, perceives it and describes it in her/his own language, coining a new term.
2. Readers (domain experts, translators, localizers): they usually deal with the term first, then they try to find what it represents, so for them a term is actually not language-independent. This will not really effect our definition of a multilingual terminological sphere, but it changes the multilingualization process of a terminology.

Hence, terminology multilingualization is not always the process of finding a lexical unit that represents the concept in each language, but rather the process of lexical translation of the term that has been developed to represent the concept.

I.3.2 **Linguistic Coverage**

This subsection discusses criteria for judging a multilingual terminological repository, based on its coverage. There are two types of coverage: linguistic, and informational (Daoud 2007).

I.3.2.1 Linguistic Coverage Formula

For a terminological repository, the linguistic coverage is the ratio between the number L of languages included in the termbase that has at least certain number of entries K and the number of the needed languages N, where $N \geq L \geq 1$, and $k>0$.

$$M_{lang} = L / N$$

This measure is important to compare termbases based on the number of languages they serve. For example, a domain shared by a community of English, French and Arabic speakers should have a termbase that contains at least k term in each of the three languages.

I.3.2.2 Examples

IATE covers the 24 languages including the 23 official languages of Europe, thus it has 100% linguistic coverage, but when K is very small, meaning that not all languages at IATE have the same number of entries. For example, the Bulgarian language has 3.500 entries only, while English has more than 1,400,000 entries. When K = 25,000:

$M_{lang} = 12 / 24 = 50\%$

Because only 12 languages have more than 20,000 entries.

Also UNTERM (UN-Geo 2002; UN 2008) has a linguistic coverage of 100%, when K is small, as it covers all the six official languages of the United Nations.

But a problem in this measure is that even though the linguistic coverage is high in average, the amount of information in each language is not equal, meaning that there are concepts that do not have terms in some languages, that is why we introduce the concept ofinformational coverage in the next subsection.

I.3.3 **Informational Coverage**

I.3.3.1 Informational Coverage Formula

Informational coverage (lexical coverage) is the ratio of entries (terms) in each language of a termbase to the concepts of the domain.

$$IC = (\sum_{k} I_{TK}) \ / \ (N * I_s)$$

Where IC is the informational coverage, K is a language in the termbase, I_{TK}, is the number of entries in K, I_s is the number of entries in the terminological sphere, and N is the number of the languages in the termbase.

I.3.3.2 Terminological Sphere Size Estimation

To measure the informational coverage for a language, we need a reference of the existing bases to estimate the size of the terminological sphere of a domain.

A good candidate could be a resource of a rich language like English. For example, in the domain of law, if we learn that the term base has 1000 terms in English, and 200 terms in Arabic, we assume that I_s is 1000.

a. Informational Coverage of the Arabic Language

Although it is the native language of around 250 million persons, Arabic is considered a pi-language due to the shortage of its lexical resources. This is illustrated by the scattered Arabic termbanks, with their notably low informational coverage and outdated entries.

For example (Arabization.org 2010) offers a terminology service (considered one of the best termbases for Arabic). However, its specialized terminology depends on compiling old terms. For instance, in the domain of mathematics and astronomy altogether, the term base depends on a reference published in 1990, and has only 4,067 entries, while since then many terms have been introduced in those domains. In fact the Wikipedia portal on mathematics only has more than 24,000 articles. And the gap seems bigger if we compared Arabic terminological resources to a rich bilingual terminological resources like the terminology (and phraseology) of ATLAS-II v.14 (Fujitsu 2009) which has 5.7 million entries in its "technical dictionaries".

b. Informational Coverage of the languages of the UN at UNTerm

UNTerm has 85000 multilingual entries, however not all entries are actually multilingual, as there are some terms missing in some of the six UN official languages, as figure 9 shows.

The following figure shows that in 12 multilingual entries, 10 monolingual terms were missing. In that case, if we assume that I_s is 12, then IC on UNTerm, according to this estimation, is equal to (12 * 6) – 10 / (12 * 6) = 62 / 72 = ~86%.

Again, this is under the assumption that the 85,000 entries represent the whole terminological sphere of the terminology of the UN, which is not true. Hence, the informational coverage can give a measure of the size of multilingualization, but for the size of the corresponding terminology, the terminological sphere should be estimated carefully.

	technology, statistics	hypothesis testing; significance testing	test d'hypothèse; vérification de la signification	verificación de significación [prop.]	выверение гипотез	假设检验；显著性检验	اختبار الفرضيات
	landmines and mine action	recognition pole	piquet de repérage électromagnétique; piquet témoin	-	-	-	عمود التعرف
	landmines and mine action	International Test and Evaluation Programme for Humanitarian Demining	Programme international d'essai et d'évaluation des techniques de déminage humanitaire	-	-	国际人道主义排雷测验和评价方案	البرنامج الدولي للاختبار والتقييم لإزالة الألغام للأغراض الإنسانية
	narcotic drugs	Fischer-Morris test	test de Fischer-Morris	prueba de Fischer-Morris	анализ Фишера-Морриса	菲舍尔－莫里斯试验	اختبار فيشر-موريس
	health and medicine, logistics and supplies	color vision chart	atlas pour le test de la vision des couleurs	tarjetas de los mapas de colores; láminas de Ishihara	карточки для проверки цветового зрения	色觉检查图	شرائح اختبار رؤية الألوان
	accounting and auditing	substantive procedure	-	-	процедура проверки по существу	实质性程序	إجراء موضوعي
	accounting and auditing	test of detail	-	-	детальный тест	细节测试	فحص التفاصيل
	-	hot acceptance test	essai de recette à chaud; essai à chaud	ensayo de recepción en caliente; prueba de aceptación en caliente	приемочное испытание с запуском двигателя	热验收测试	تجربة القبول مع تشغيل المحرك
	budget and management, economics, science and technology	predictive value	1. valeur prédictive [science]; 2. valeur de prédiction; valeur de prévision; valeur prédictive [écon.]	valor de predicción	ценность для прогнозирования	预测值	القيمة التنبؤية
	narcotic drugs	cobalt thiocyanate test	test au thiocyanate de cobalt	prueba de tiocianato de cobalto	анализ с применением гуанидиния кобальта	硫氰酸钴检验	الاختبار بثيوسيانات الكوبلت
	narcotic drugs, statistics	comparison-of-means test; t-test	test de comparaison des moyennes (test-t)	prueba (test) de comparación de medias o prueba (test) t	сравнительный анализ методов	平均数比较检验；t－检验	اختبار المقارنة بين متوسطين
	narcotic drugs	odour test	test olfactif	ensayo del olor	испытание на запах; проба на запах	-	اختبار الرائحة

Fig. 9: UNTerm incomplete multilingual records

I.4 The Lexical Gap between Communities and Terminological Repositories

I.4.1 Overview

After defining the concepts of multilingual terminology, term base, and coverage, this section describes the main problems in the current Terminology as a consequence of the low linguistic and informational coverage.

Due to scientific and social needs, concepts are introduced constantly, and that requires new signs (terms) to represent these concepts. The produced terms are coined in one language. In order for these terms to be used by all the members of the domain community, it is necessary to multilingualize them. However, the rate of term production is faster than their multilingualization. This causes a substantial decrease in the informational and linguistic coverage, which leads to a situation where many concepts do not have signs to describe them in some languages.

I.4.2 Absence of Terminology

I.4.2.1 Definitions

A *lexical gap* or *lacuna* is the absence of a word or a compound word in a particular language. A terminological lacuna is the absence of a term for a concept in a particular language.

Sometimes, when a concept is developed in a specific community, it is described using a term of the language of that community and then this term is multilingualized to other languages. In the

case of a terminological lacuna, the term is never translated into the considered languages, figure 10.

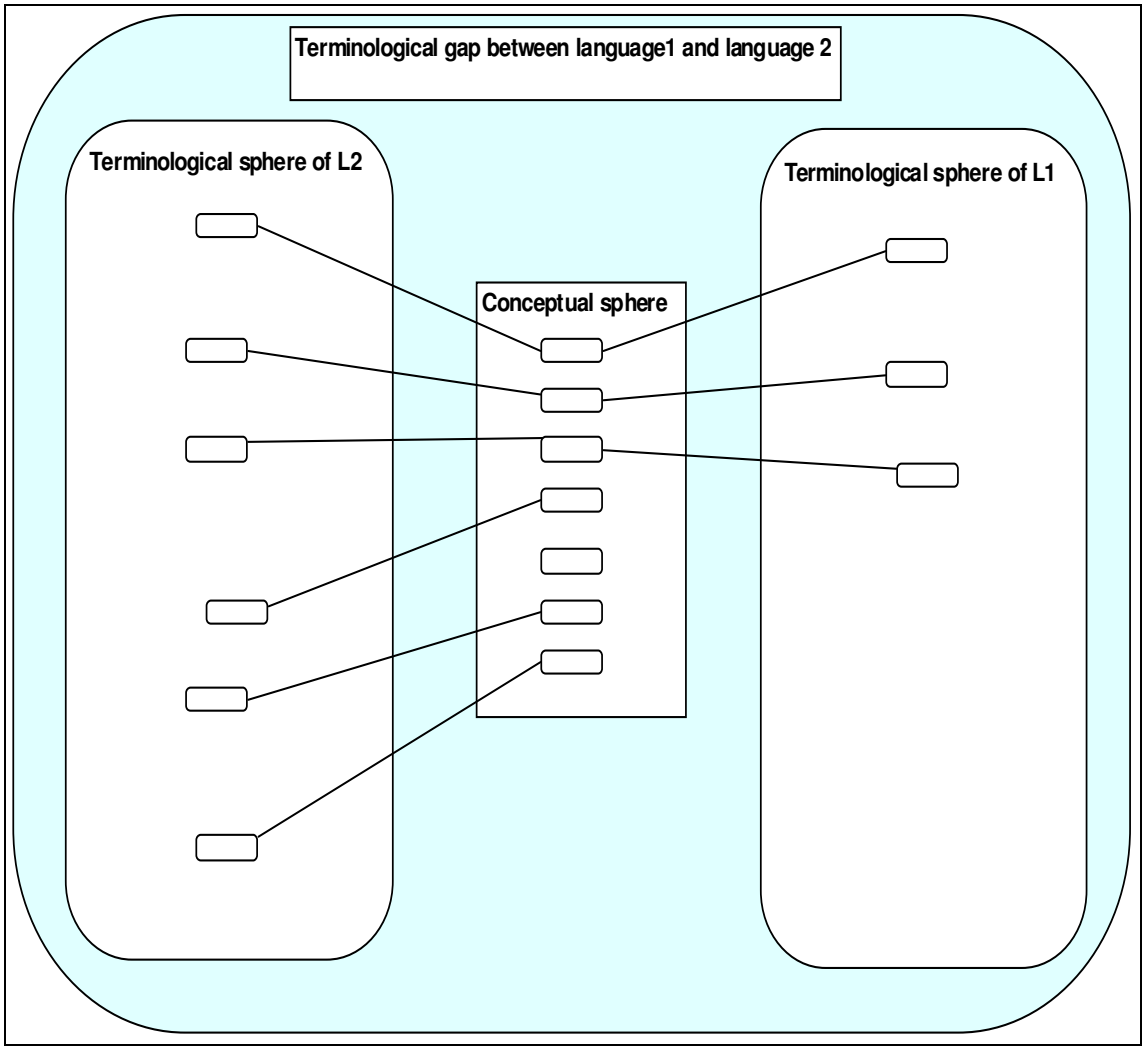


Fig. 10: Terminological gap between dominant languages and poorly equipped languages

I.4.2.2 Absence of Terminology and Knowledge Transfer

A term is considered to be the smallest piece of information in a knowledge structure. Transferring knowledge from a language to another starts by finding a glossary of terms and standardizing them. The absence of terminology is considered one of the main issues that prevent publications in some domains in many languages. It is also very difficult to produce knowledge in these languages, due to the difficulties in finding proper terms.

Wikipedia's (Wikipedia-A 2008) articles are concept-based. Each article focuses on one concept. The availability of terms in a language makes it easy for the community to produce Wikipedia articles. Hence, we may conjecture that the size of Wikipedias in different languages gives an indication of the available terms in each language. Figure 11 shows statistics about the first 25 Wikipedias.

№	Language	Language (local)	Wiki	Articles	Total	Edits	Admins	Users	Active Users	Images	Depth
1	English	English	en	**3,328,300**	20,678,978	395,237,259	1,730	12,570,379	143,741	849,360	519
2	German	Deutsch	de	**1,083,421**	3,091,593	78,881,754	289	1,018,643	24,726	193,197	88

100 000+ articles [edit]

№	Language	Language (local)	Wiki	Articles	Total	Edits	Admins	Users	Active Users	Images	Depth
3	French	Français	fr	**960,746**	3,904,213	57,882,127	189	855,185	16,200	40,193	139
4	Polish	Polski	pl	**708,702**	1,269,716	23,237,078	165	371,355	5,942	1,106	11
5	Italian	Italiano	it	**699,426**	2,174,565	36,361,923	99	525,236	8,927	77,954	74
6	Japanese	日本語	ja	**684,725**	1,795,007	33,265,748	65	429,928	11,714	74,919	49
7	Spanish	Español	es	**611,109**	2,795,463	40,642,388	138	1,519,063	16,201	0	186
8	Dutch	Nederlands	nl	**607,393**	1,466,716	21,701,234	68	322,331	5,285	18	30
9	Portuguese	Português	pt	**587,140**	2,283,437	21,027,214	38	746,354	6,358	0	77
10	Russian	Русский	ru	**551,157**	1,967,854	26,883,557	87	505,677	12,860	101,918	90
11	Swedish	Svenska	sv	**359,974**	946,684	12,291,331	103	179,290	3,342	0	34
12	Chinese	中文	zh	**313,172**	1,010,370	13,615,844	79	849,771	5,865	28,193	67
13	Norwegian (Bokmål)	Norsk (Bokmål)	no	**263,851**	640,333	7,653,345	65	164,168	2,213	627	24
14	Catalan	Català	ca	**248,217**	596,173	5,573,484	23	62,518	1,460	5,393	18
15	Finnish	Suomi	fi	**241,727**	652,045	8,883,371	48	151,285	2,306	24,215	39
16	Ukrainian	Українська	uk	**214,344**	669,132	4,568,642	20	72,654	1,578	35,530	31
17	Czech	Čeština	cs	**165,850**	430,440	5,696,662	27	116,139	2,151	170	34
18	Hungarian	Magyar	hu	**163,956**	545,518	8,073,733	36	141,533	2,392	35,116	80
19	Turkish	Türkçe	tr	**145,663**	714,210	8,275,060	23	291,390	2,419	23,954	177
20	Romanian	Română	ro	**145,600**	602,134	4,339,847	25	147,928	1,609	41,564	71
21	Korean	한국어	ko	**138,756**	403,570	5,658,583	24	118,397	2,224	7,830	51
22	Esperanto	Esperanto	eo	**130,803**	289,941	2,959,896	18	37,363	449	11,978	15
23	Danish	Dansk	da	**130,721**	346,726	4,276,779	41	106,634	1,268	9	34
24	Arabic	العربية	ar	**127,914**	717,591	6,609,881	22	290,409	2,033	8,033	196
25	Indonesian	Bahasa Indonesia	id	**126,762**	454,140	3,918,463	14	183,915	1,520	20,211	58

Fig. 11: Sizes of Wikipedias

The size of a Wikipedia is not only affected by the absence of terminology, but also by many other factors, like the number of native speakers. Although the number of speakers of Arabic is larger than the number of the native speakers of French, German, or Japanese for example, the Arabic Wikipedia is significantly smaller. In the case of Arabic, we think that one of the main reasons for this problem is the absence of proper terms which makes the production of articles more difficult because the authors have to coin new terms.

I.4.3 **Hidden Terminology**

I.4.3.1 *Definition of Latent Terminology*

In the case where a term is absent in a particular language, the community resorts to one of the following solutions.

- Using a circumlocution instead of a term. A circumlocution is a descriptive sentence for the corresponding concept.
- Using a foreign term.
- Using a term without proper recording nor standardization, and this is what we call *latent terminology*.

I.4.3.2 Example of the Arabic Latent Terminology and Discussion

Social networks introduced many concepts, like: status update, group invitation, friends request… These concepts were not initially translated nor recorded properly into Arabic; however, the Arabic users of social networks, like Facebook (Facebook 2010) succeeded to find Arabic terms to represent these concepts. Now they are coined, but still not officially recorded.

In the case of Arabic, one can not find proper terms for many concepts in the standard terminological Arabic sphere. However one can find them hidden somewhere in the lexical sphere of Arabic, see figure 12.

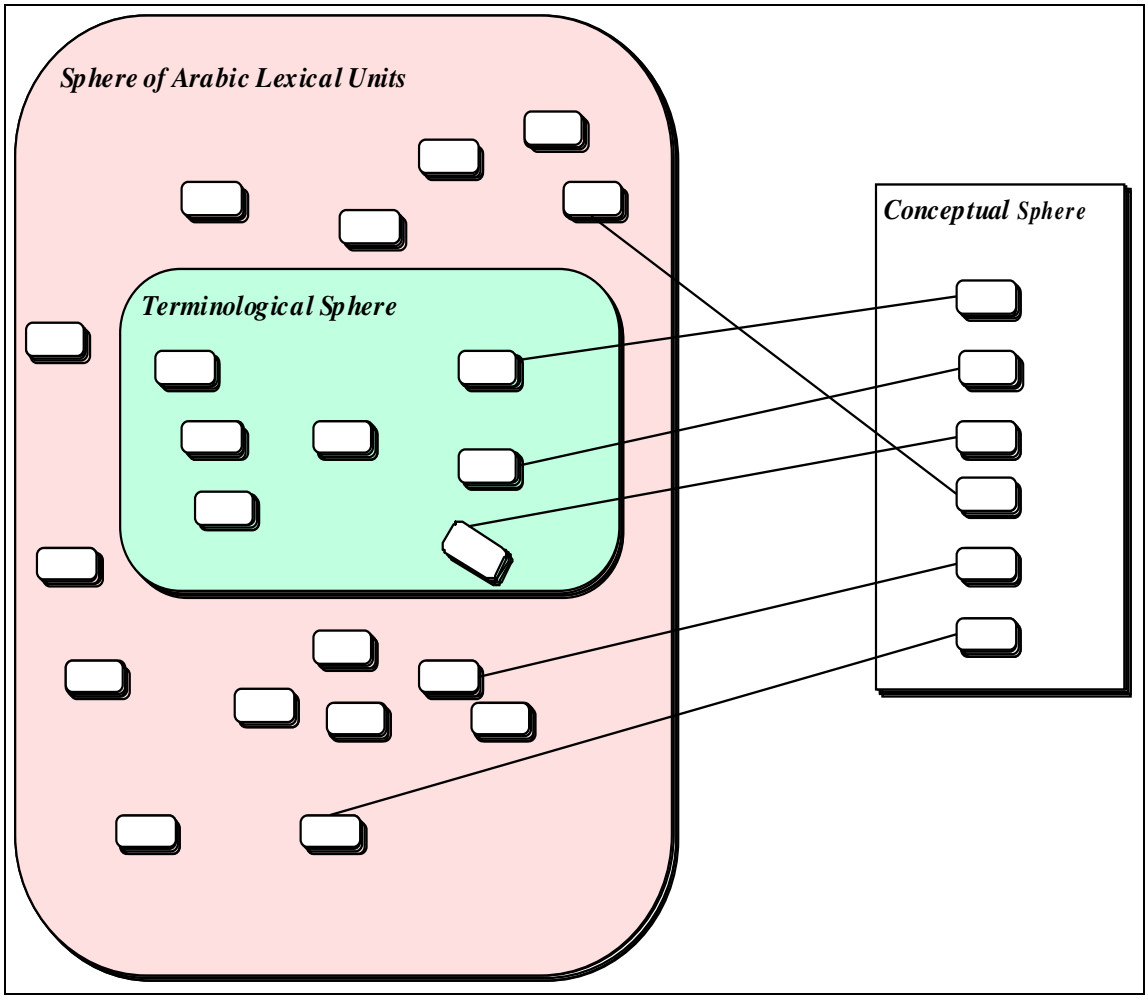


Fig. 12: Arabic Lexical sphere

Some of the concepts have their terms outside the terminological sphere: they are not standardized, and have no descriptive information.

Conclusion

A term is a sign to describe a concept; a terminology constitutes the sphere of terms that describes and represents the concepts of one domain. An **object** or an idea is perceived by the human brain as a **concept**, and represented as a **term**. This triangle has been modified to add terminological information, particularly definitions, to situate a term within a terminology, and to define its relations with other terms. Multilingual terminology is a collection of multilingual

spheres of terms representing the concepts of the same domain in different languages. Usually, such spheres are maintained and managed in a computerized database system or "term base" that organizes terminology and offers functionalities like modification and retrieval.

A concept system is usually more dynamic than its representation. That is why a terminology might not really cover all the concepts of a domain. Linguistic resources vary, that is why not all multilingual spheres cover the same amount of concepts. In fact, lexical and linguistic coverage is a big issue in multilingual terminological repositories. When a concept has a word or term to express it in language A, but not in language B, there is a lexical or terminological lacuna in B with respect to A. Many communities do not depend on a terminological repository to maintain their terms, but they find equivalent terms and use them without proper registration. In that case, we speak of "latent terminology".

Chapter II State of the Art in Developing Multilingual Terminology

Introduction

Last chapter gave the background regarding multilingual terminological resources, and illustrated the problem of the terminological gap between languages. This chapter shows the current approaches to deal with the problem of lexical gap through the development of multilingual terminological resources.

The approaches are classified into three categories:

- Classical (professional) approaches.
- Automatic approaches.
- Collaborative approaches.

This chapter is organized as follows: the next section presents the classical approaches which depend on human terminologists, the second section shows automatic approaches used for terminology, the third section discusses the recent trends in exploiting volunteer contributions to develop knowledge bases, and finally section 4 describes the limitations of these approaches and the required framework.

II.1 Classical Approach

The classical approach in constructing a multilingual term base depends on the effort of *terminologists* to define the boundaries of the domain, and discover the terms for each language.

Mostly, terminologists are not the term developers. In fact they often do not interact directly with the concepts of the domain, but they deal with already established terms and try to find their corresponding concepts, in order to define the terms and situate them within the terminological sphere.

That is a fundamental difference between term coiners (who are experts in the domain) and terminologists. Building a term base involves a substantial amount of recording already known and defined terms, while less effort is needed in labeling concepts by finding their terms.

II.1.1 General Process

Figure 13 shows a general view of the traditional approach in constructing a multilingual terminological database (Cabre and Sager 1999) (Kim, Yang et al. 2005). Usually one starts with a thematic analysis of the domain, to find its logical components. After that, a team of terminologists consults some related documents to find the most important terms for each sub-domain. Then for each targeted language, a team of terminologists tries to find equivalences of each extracted term in the target language (Hartley and Paris 1997), along with descriptive information. Finally, each entry is verified to reach an agreement on its correctness.

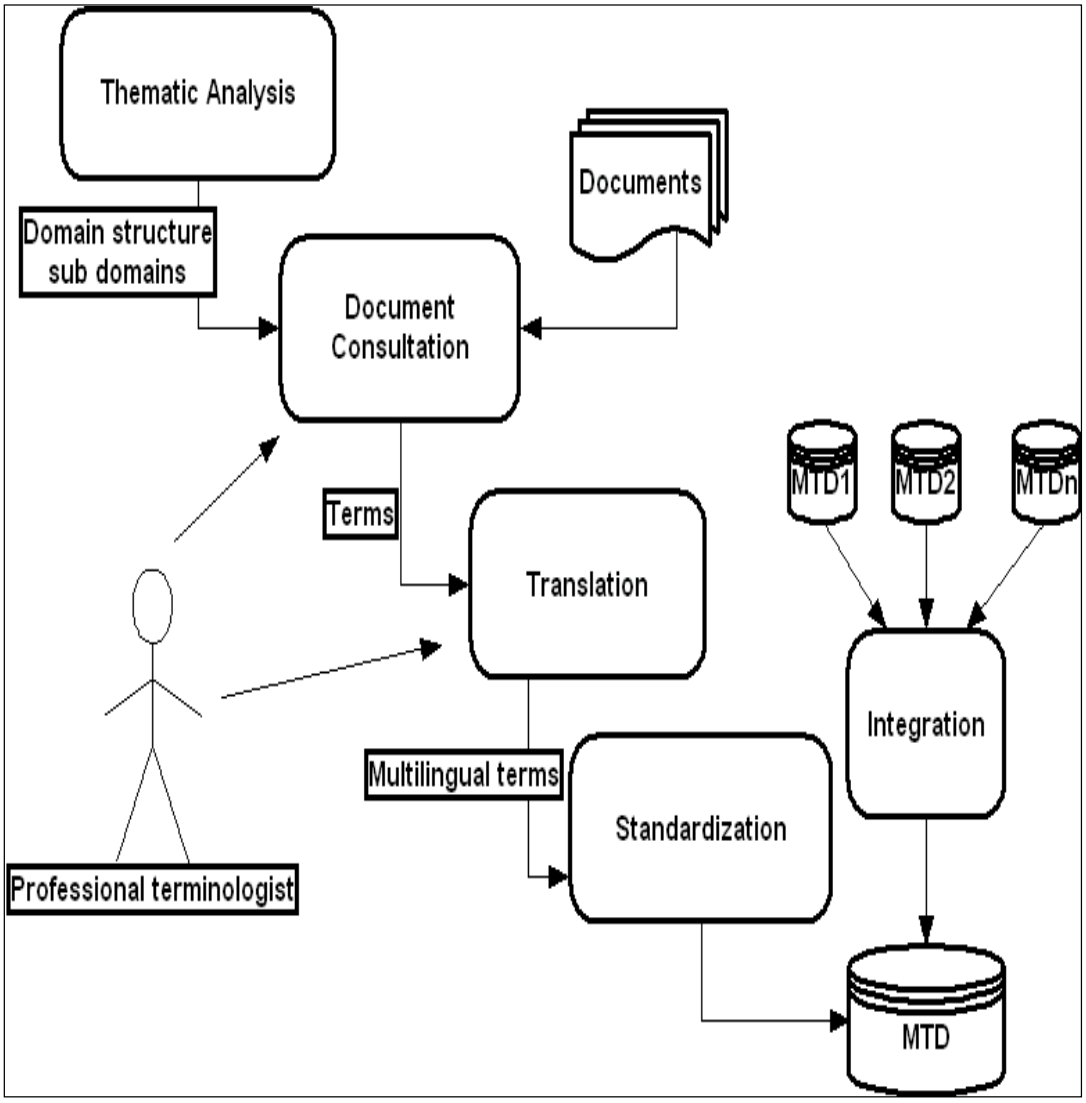


Fig. 13: Classical approach – general process

II.1.1.1 Thematic Analysis

The thematic analysis of a domain is the process of finding the logical components and the boundaries of the domain. This step divides the terminology work into smaller pieces and produces a taxonomy for the domain.

a. Domain Taxonomy

A domain taxonomy is the hierarchal structure of the domain. Usually this structure is organized by generalization-specialization relationships, or less formally, parent-child relationships. Finding the taxonomy effectively organizes the concepts of the domain.

b. Example

Figure 14 shows a part of a taxonomy of programming languages. Anyone who is interested in developing a termbase for programming languages should utilize such a taxonomy.

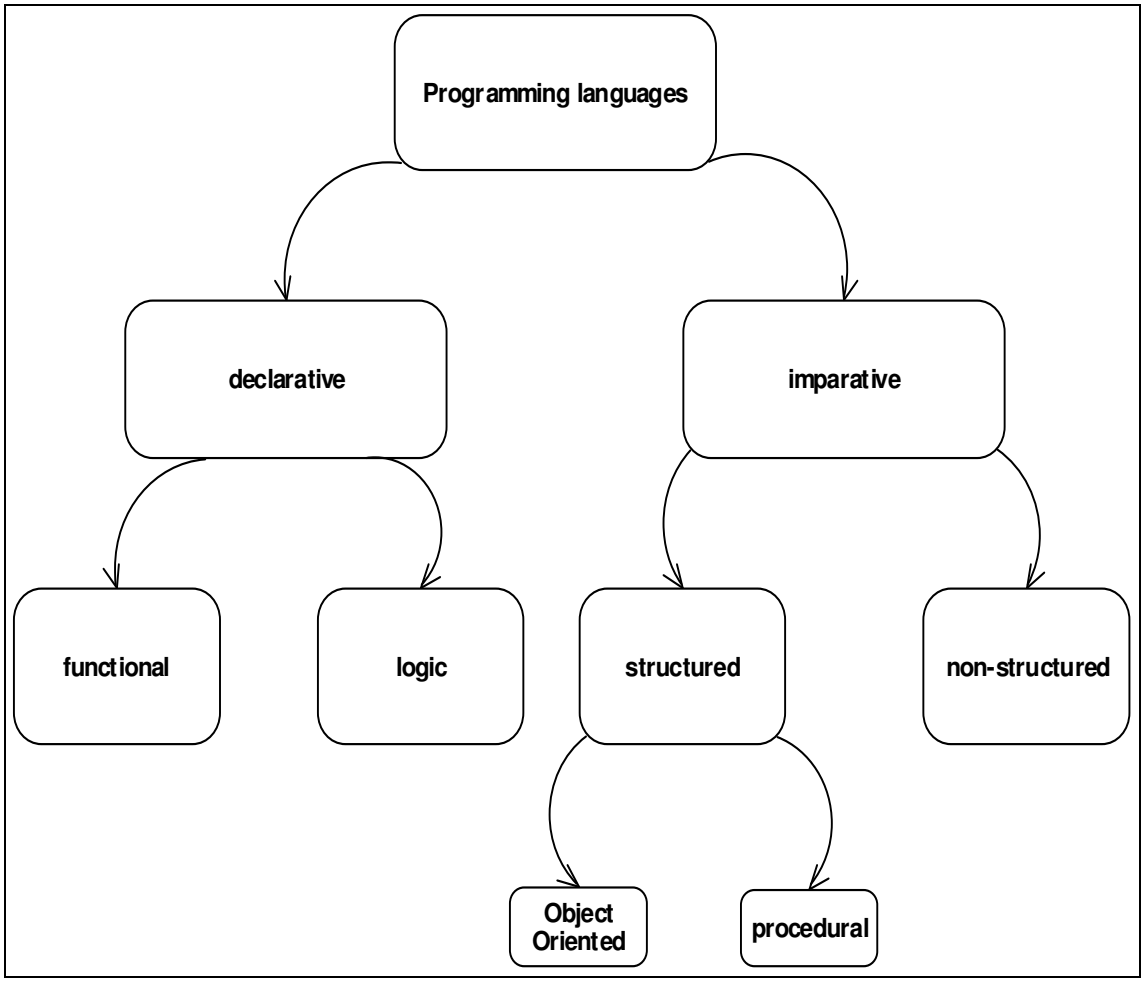


Fig. 14: Taxonomy of programming languages

Each subcomponent has its own concepts and terms, and finding them for the smaller component is easier and more organized.

II.1.1.2 Documents Consultations

After finding the logical components for the domain, terminologists compile relevant documents for each sub-domain. These documents are then studied to identify the main terms (note that in this case the terminologists do not interact with the concept directly but rather with its sign). The extracted terms constitute the candidates for the termbase.

II.1.1.3 Terminology Translation

Different groups of terminology translators of each language try to translate the terms extracted from the previous step. Terminology translation is difficult as it needs a lot of reference information (original context, usage, definition…). Terminology translators should not only be bilingual, but they should be familiar with the domain that they are dealing with.

II.1.1.4 Standardization

Terminology standardization is the process of producing a consensus between delegations representing various concerned groups in the domain representing different cultural, economical, and political points of view.

Standardization provides a common technical language between those groups, which eases the communication between them. Standard terminology is usually produced by official standardization bodies.

II.1.2 Example of Current Classical Terminological Multilingual Databases

The above approach needs a lot of resources, particularly, human resources. Only large organizations are able to conduct such kind of work. Table 2 provides some examples of online multilingual terminological databases built by large organizations. It is clear that the providers have mature resources and experience to build such databases, like (IDRC 2009) (EVROTERM 2010) (EUSKALTERM 2007) (EuroFIR 2008). In fact, these online systems are continuations to efforts started decades ago, and they have been compiled using material from existing databases. For example, IATE which is provided by the EU has been constructed by compiling previous databases, namely EURODICAUTOM, EUTERPE, and (TIS).

Table 2: Some existing multilingual terminological online databases

Name	Number of terms (ML)	Number of Languages	Domains	Provider
IATE (IATE 2008)	8.400,000 terms	23 languages	General, 155 domains	EU
UNTerm (UN 2008)	80,000 terms	6 languages	100 subjects related to the UN	UN
FAOTerm (FAO 2008)	58,000 terms	7 languages	FAO related domains and bodies	FAO
Electropedia (IEC 2008)	20,000 terms	9 languages	Electrical terminology in 75 cats.	IEC
The Great Terminological Dictionary (OQLF 2008)	3000,000 terms	Fr, En, and Latin	200 categories	OQLF
EuroTermBank (ETB 2009) (EuroFIR 2008)	~100,000 terms	27	22	TILDE, IIM
LANGUAL (Jean A.T. Pennington 1994)	25,000 terms	6 languages	14	EuroFIR
International Monetary Fund terminology (IMF 2010)	4,500 terms	5 languages	1	IMF
The unified health lexicon (WHO-EMRO 2009)	~130,000	3 languages	90 sub-lexicons	WHO

Not only are the conventional approaches in building a multilingual terminological database very expensive, but it is usually difficult to achieve good coverage (either informational or linguistic), especially in particular or specific domains. Besides, terminologists are more prone than domain experts to introduce inaccuracies.

II.1.3 Limitations

As a result of depending on professional terminologists, building a multilingual term base is very expensive, especially when it comes to identifying and extracting a term, defining it, and

translating it into another language. An entry may need hours of professionals' time. The following points are the rest of the most important limitations.

1. **Coverage**. (a) *Lexical coverage*: in the traditional approach, extracting the terms is done manually by consulting related documents, which might not have a reasonable amount of technical terms of the domain. Beside, human terminologists may not extract all the available terms. (b) *Linguistic coverage*: there might not be enough terminologists for some language pairs, which means the database will not be available in all the targeted languages.
2. **Involvement of domain experts**. The traditional approaches may create a gap between the terminologists and the subject matter experts, as they are not directly involved in the decision and the technicality of creating the base entries.
3. **Rigidity of the approach**. Terminology is alive, thousands of terms are produced yearly, and a rigid approach can not solve the problem.

II.2 Automatic Approaches

Beside the traditional approach, automatic approaches are used in the process of developing a multilingual termbase. In the overall general classical process, automatic approaches try to decrease the human effort, specifically in consulting the text, finding the terms and translating them.

This section discusses these approaches, as well as their resources and techniques, and presents their limitations.

II.2.1 Machine-Readable Dictionaries

II.2.1.1 Overview and History

A Machine-readable dictionary (MRD) is a dictionary stored electronically instead of being only printed on paper. That means that its data can be read and manipulated automatically through an application software (Louise, James et al. 1996).

In the 1960s, the Merriam-Webster Seventh Collegiate Dictionary and the Merriam-Webster New Pocket Dictionary were produced as machine-readable tapes (Merriam-Webster 2010). Since then, tens of monolingual and multilingual dictionaries and lexicons have been published as MRDs. The main interest of the research on MRDs is the extraction of knowledge automatically from these dictionaries. Nowadays, these dictionaries can be accessed online or offline. This section discusses the use of such general purpose dictionaries in the development of termbases.

II.2.1.2 Offline MRDs

An offline MRDs is a dictionary in digital format. Its data can not be accessed through Internet, but the whole dictionary can be placed in the users' machine (Copestake, Briscoe et al. 1994).

What is special about MRDs is that they do not need special software functionalities to be used (search, edit, etc.). They should only follow an understandable format so they can be used manually or automatically through a separate application software.

Each dictionary has different purposes and characteristics because each has its information and structure, and that creates many inconsistencies between MRDs. Such inconsistencies decrease the reusability of these MRDs and of the tools built to manipulate them.

There have been many attempts at finding a standard XML format can that handle all the needs of different dictionaries, like DML (Dictionary Markup Language) (Corréard and Mangeot

1999), Lexical Markup Framework (LMF) (Francopoulo, George et al. 2006), TBX (LISA 2010) and DicML (DicML 2003), which provide facilities to describe entries of monolingual and multilingual dictionaries. The availability of conversion and manipulation tools dedicated for these formats is important to make them useful.

XDXF (XML Dictionary eXchange Format) (XDXF 2010) is another format that has been adopted to unify all existing open source dictionaries. So far, the XDXF project has 630 dictionaries, and conversion tools from other XML or non XML formats like SimpleDict (SimpleDict 2010) and StarDict (StarDict 2010).

Here is an example of an English-Arabic dictionary entry formatted using XDXF:

```
<ar>
<k>Absenteeism</k>
الغياب :: التغرب :: التغيب :: حالة التَّغيّب عن العمل او أداء الواجب
</ar>
```

The Text Encoding Initiative (TEI) is a project that maintains a standard or the representation of texts in digital form. Chapter 9 of its guidelines suggests an XML format for the macrostructure and microstructures of digital dictionaries (TEI 2009). Each microstructure has an entry and several senses, and translations for each sense:

```
<entry>
 <sense n="1"/>
 <sense n="2"/>
</entry>
```

II.2.1.3 Online Dictionaries

Recently, many dictionaries have been made available as free online services (Systran 2009) (WordReference 2008). They vary from monolingual to multilingual and from domain-specific to general-purpose. Their performance (in terms of quality and coverage (lexical and linguistic)) depends on the language pair in question.

For example, Word Reference (WordReference 2008) has dictionaries for more than 20 languages pairs, while Google dictionary now has around 50 language pairs (Google 2010).

The entries of these dictionaries can not be processed in batch like offline dictionaries, but they can be accessed through Internet one entry at a time.

II.2.1.4 Automatic Acquisition of Lexical Data from MRDs

a. Online Dictionaries Acquisition

Although one can not have access to the whole data set of an online dictionary, usually these dictionaries are accessed through an http request, see figure 15.

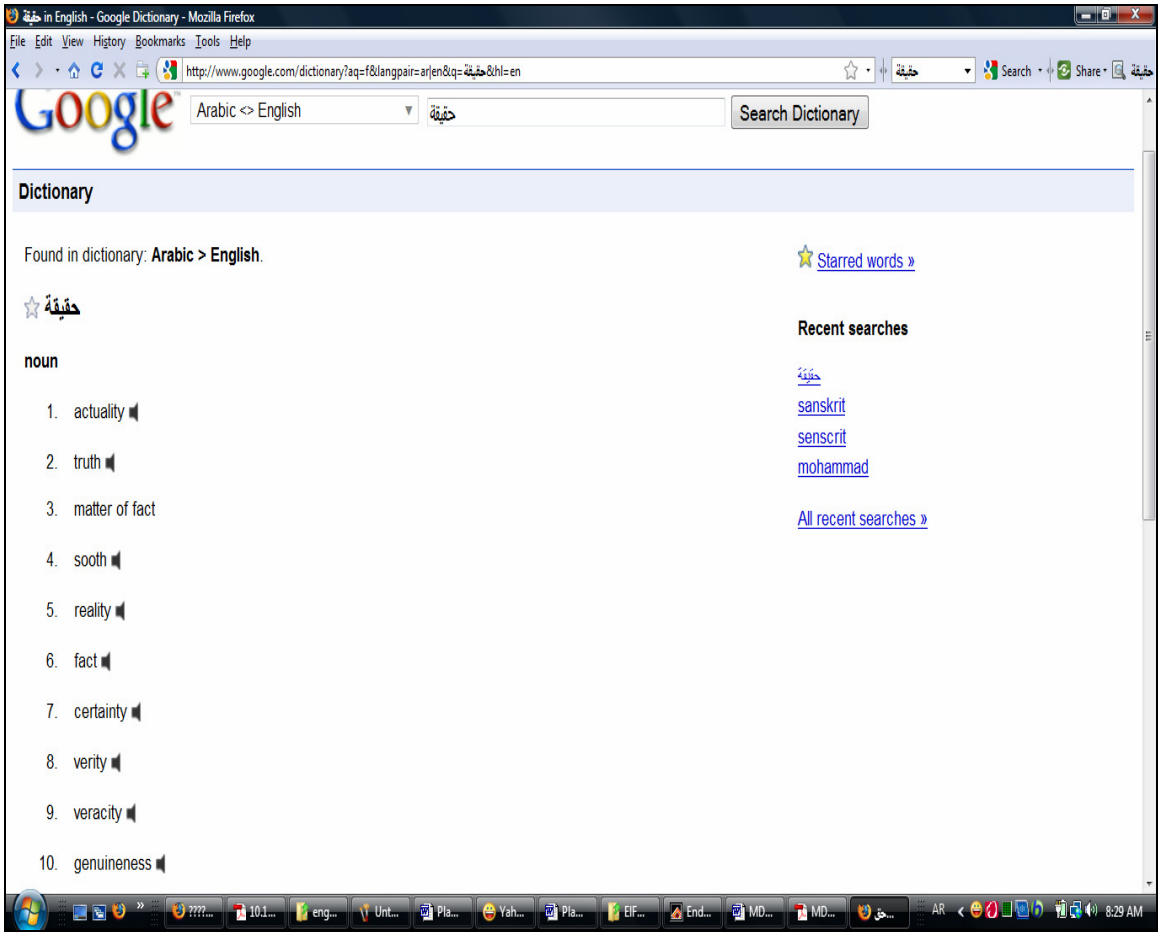


Fig. 15: Google Dictionary

Many systems use APIs to access online dictionaries by sending automatic http requests and analyzing the returned html page to retrieve lexical information.

b. Offline Dictionaries Acquisitions

An offline dictionary is available as an offline document, hence it can be automatically used by building an application software that analyses the entries. An XML parser is a typical software to search and retrieve lexical data from an offline dictionary, furthermore it should identify lexical and translational relations between the extracted lexical units. That is one of the motives to unify the structure of MRDs so there would be a possibility to reuse the tools developed for the analysis and maintenance of each dictionary.

II.2.2 Corpus Analysis for Terminology

A text corpus is a large and sometimes structured set of texts stored electronically and processed automatically (Tercedor and López-Rodríguez 2008). A monolingual corpus has text in a single language, while multilingual corpora have aligned text from different languages (Huynh, Boitet et al. 2008).

Corpora are heavily used in statistical analysis of texts to prepare SMT (Statistical Machine Translation) system. They are also used for validating translations and linguistic rules, and for checking occurrences of a specific text fragment to allow to study its context and the way it is used. Hence, bilingual corpora are used as a reference to the language. Translators and

terminologists use them in order to understand the source text and its context, and to validate their produced text.

There have been many attempts at utilizing domain-specific parallel or comparable corpora to extract multilingual terminology (Sadat, Yoshikawa et al. 2003) (Aussenac-Gilles and Jacques 2008) (Tufiş 2004; Aussenac-Gilles and Sörgel 2005) (Tiedemann 2003; Huang, Zhang et al. 2005) and even monolingual corpus, like (Prince and Ferrari 2000). An analyzer takes the corpora as an input, and based on the frequency of the terms and a morpho-syntactic pattern (or simply a co-occurrence pattern), it produces candidate entries for a multilingual termbase.

Hence, corpus analysis is useful for the terminological work in three ways:

- it provides validation and checking for the terminologists.
- it gives a definition of the boundaries of the domain.
- it helps the recognition of possible multilingual terms (next subsection).

II.2.3 **Multilingual Terminology Extraction**

II.2.3.1 Automatic Term Extraction

Automatic term recognition (ATR) is the task of automatically extracting terms or keywords from (monolingual) text (Kageura and Umino 1998) (Halskov and Barrière 2008) (Calberg-Challot, Candel et al. 2008).

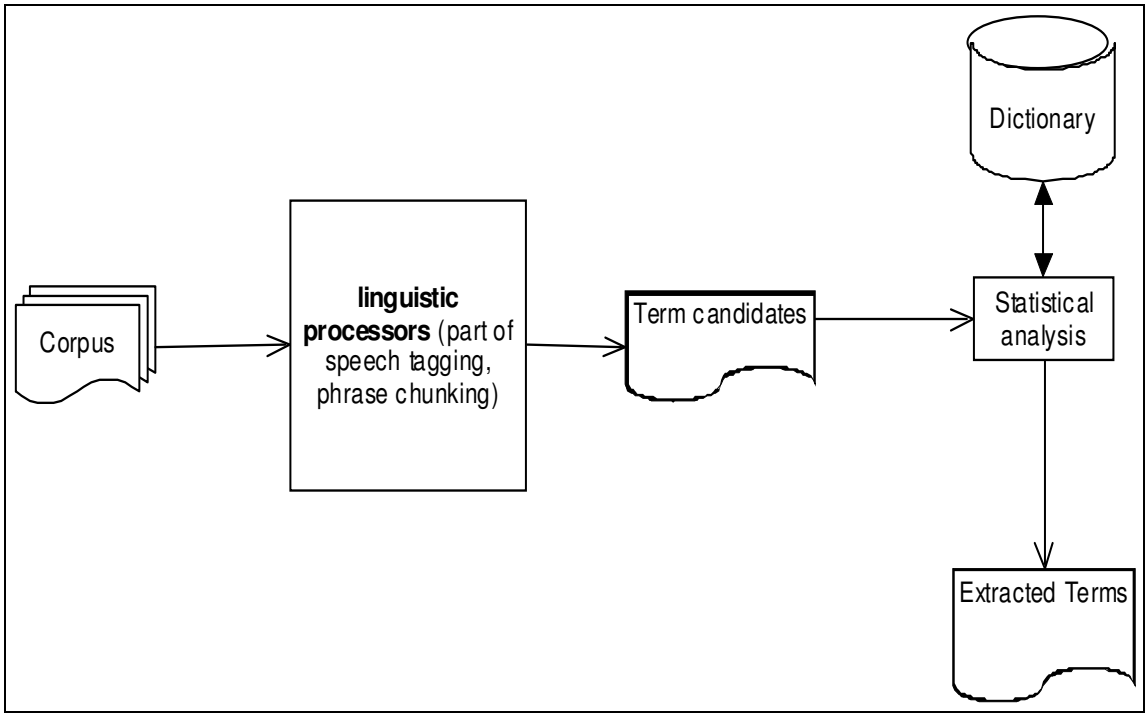


Fig. 16: Automatic term extraction

Term extraction starts by processing the text linguistically, and by annotating it (Kawazoe, Jin et al. 2008), one then produces candidates, which are finally filtered by applying statistical methods, see figure 16.

Some of these systems are available as online services, that can be used for free for a specific task, like Yahoo! Terms (Yahoo 2008), TExtractor (Valderrábanos, Belskis et al. 2002) and

TermExtractor (LCL 2010). Figure 17 shows an example. There are also lighter version of term extractors, called "tag cloud generator", for example, (TagCrowd 2010), which is a simple application that measure the frequency of the words and visualize them (see figure 18).

Fig. 17: Yahoo! Term

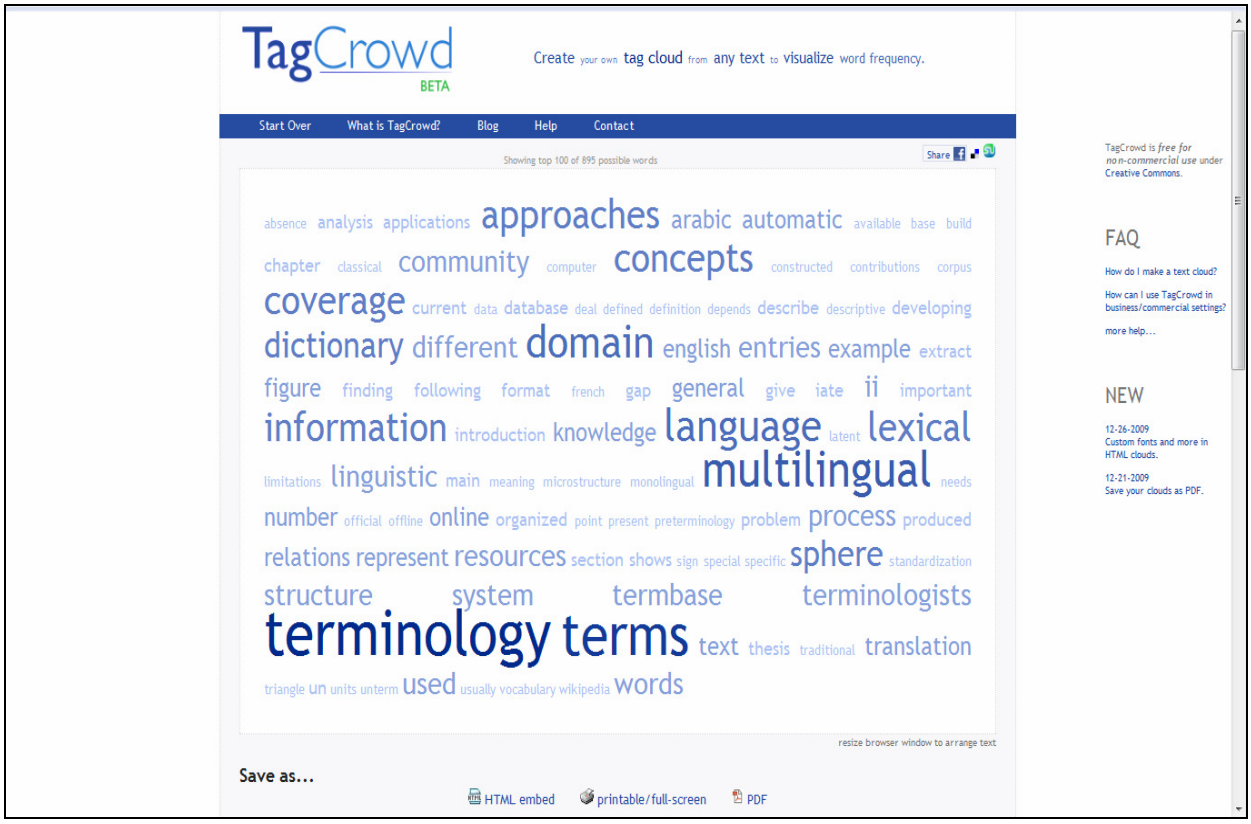


Fig. 18: Tag cloud

II.2.3.2 Bilingual and Multilingual Term Extraction

Multilingual Term Extraction is the process of analyzing a parallel corpus to find a term and its translation(s) in different languages (Auger and Barrière 2008) (Heid, Jauß et al. 1996). Figure 19 shows a typical bilingual term extraction process (Kwong, K.Tsou et al. 2002).

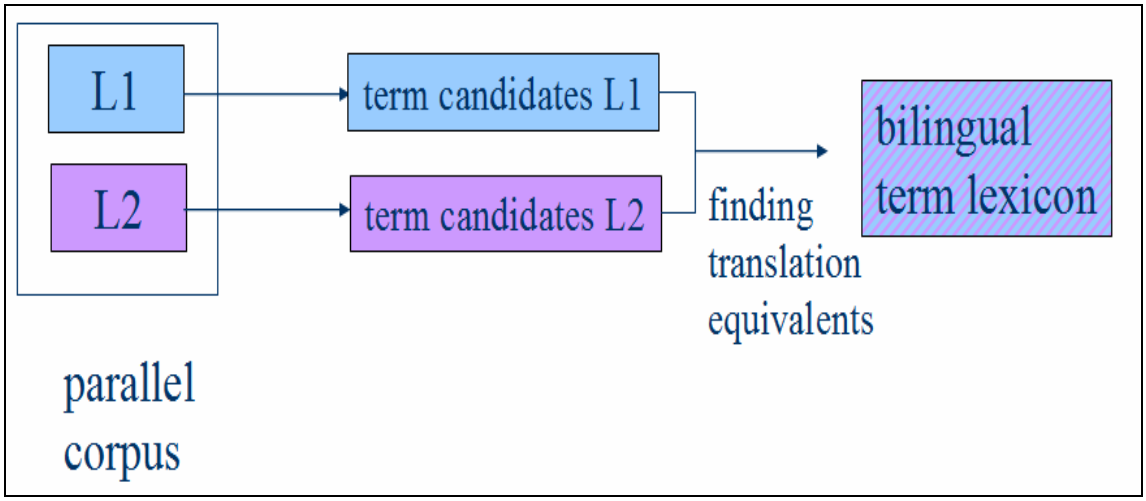


Fig. 19: Bilingual term extraction

The first step in finding bilingual terms is to find term candidates as described earlier. The next step is to match the term with its translation using term aligning techniques described in (Och

and Ney 2003) using one of the following tools: Twente (Hiemstra 1998), Egypt/Giza/Giza++ (Och and Ney 2003), PLUG (Tiedemann 2003) etc.

II.2.4 Limitations

There are three major limitations in the usage of automatic approaches.

1. **The performance of the technique**. Whether we are extracting terminology from text, structured text, or even an MRD: the relevance of the extracted data to the domain is still an issue. Not every lexical unit with a certain frequency qualifies to be a term, and not even every extracted term actually belongs to the domain. The same observation applies to general purpose dictionaries whose entries can not be supposed to denote concepts of a specific domain.

   Termhood measures (Kit and Liu 2008), are based solely on the characters of the text, as these techniques rarely deal with semantics.

2. **The digital resources used in the techniques**. Even if the tools and text analyzers were perfect, the big limitation is the available resources: they are only as useful as the corpus they have (Resnik and Smith 2003). The main issues of creating terminology by corpus are as follows:

   a. Relevance of the text to the domain.

   b. Coverage of the domain: the assumption that the terminology of a domain exists in the corpus is often not true, because not all terms exist in machine readable format.

   c. Consistency and timeline of the text: outdated texts often contain outdated terms.

3. **Terminology creation**. It is natural that terms are at some point in time not yet created for some concepts in all languages. Hence, it is impossible to find them in corpora. In another word one cannot solve the problem of terminology lacunae by looking into corpora. Also not all developed terms are actually available in digital format; they might be only used by the community without proper digital recording.

Although automatic analysis of digital resources is helpful to make good use of what is available, it is not enough for building a sufficient termbase.

II.3 Collaborative Approaches

II.3.1 Collaborative Knowledge Gathering: General Overview

Some tasks prove to be rather easy for a human, while challenging and difficult for computers (Ahn 2005). Many researchers are studying the human factors in handling such tasks (Richardson and Domingos 2003) (Chklovski and Gil 2005-a), in particular the task of building knowledge bases.

Extracting knowledge automatically faces many obstacles. The first one is the availability of data to be analyzed to extract knowledge, as human knowledge is not always available in a proper format. The second one is the performance and complexity of the software that extracts knowledge.

In this section, we will present the new trend in computation, which is to depend on humans to gather knowledge. We show how that could be implemented in the case of multilingual terminology (assuming multilingual terminology is a special kind of knowledge).

II.3.1.1 Human Computation

Human-based computation is a process whereby problem is solved by outsourcing some of the steps to be performed by humans. This approach benefits from some of the computational abilities and knowledge of humans, while reducing the cost. In traditional approaches, the computer receives the problem from the human, while in human-based computation the roles are reversed: the computer provides a task to a human or a potentially large group of humans, then it gathers their output, integrates it and analyses it.

Recently, and with the facilitations of Internet technologies, collaborative efforts have produced a variety of successful applications and knowledge repositories, like open source software, and Wikipedia, where volunteers from all over the world are contributing with the help of communication-facilitating techniques and platforms.

The concept of human contribution, collaboration and computation has already been utilized in many applications and scenarios. The work of Luis von Ahn made a breakthrough, especially in reCAPTCHA (reCAPTCHA 2009) and ESP (ExtraSensory Perception) games (Ahn and Dabbish 2008) (Ahn 2006). Human computation (crowdsourcing, volunteer contribution) is now seriously considered to be able to solve large computational problems (Speer 2007).

A normal architecture of the computation model is shown in figure 20, where the human is a user of a knowledge base, who contributes to that base, and may use it as well.

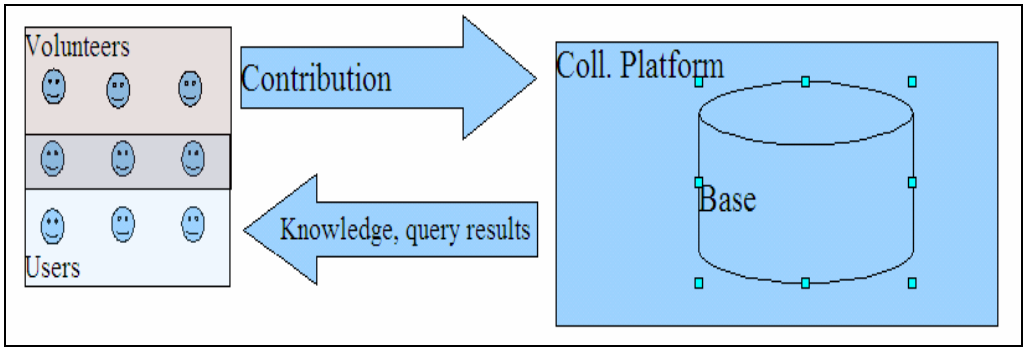


Fig. 20: Collaborative approach in knowledge gathering

Although this approach relies on many scenarios of interaction, we can categorize them into two categories: spontaneous and non-spontaneous contribution.

II.3.1.2 Spontaneous Contribution

Spontaneous scenarios associate the computation and contribution process with an activity done by the human on a regular basis.

The main incentive of the contributor here is not the contribution itself, but the regular activity he is involved in.

In reCAPTCHA, the user who wants to create an account to some website is asked to enter two words, to validate that he is not a spamming robot. One of the words is already digitized, while the other is a word scanned from a book that needs digitization. Figure 21 shows an example.

Fig. 21: reCAPTCHA

Another example is to depend on massive contributions from volunteers through an online game. GWAP stands for "Game With A Purpose" (Ahn and Dabbish 2008). There is variety of games in different domains now, like Open Mind Common Sense (Singh, Lin et al. 2002), ESP games, Learner (Chklovski and Gil 2005-a) (Chklovski and Gil 2005-b), FACTory-CYC games (FACTory 2010), (CYC 1991), Mindpixel (Mindpixel 2009), JeuxDeMots (Joubert and Lafourcade 2008), Google Image Labeler (Google 2009), etc.

II.3.1.3 Non-Spontaneous Contribution

The more traditional form of contribution is the non-spontaneous one, where the user is purposely contributing while he is fully aware of the fact that he is contributing. This approach does not attach the process to any other regular one, and the incentive for contribution varies. Wikipedia is one of the biggest successes in this category. Millions read Wikipedia. Some contributors edit its articles for various political and social reasons, and others merely because they want to support a cause or a project.

II.3.2 **Collaboration Factors**

For a collaborative system to be successful and do its function in human computation, one should consider the following *Contribution Factors (CFs): A volunteer V provides a Contribution Unit CU, because of a Motivation M, through a Collaborative Environment CE* (Daoud, Kageura et al. 2010).

Many systems succeeded by finding a suitable set of (CFs); table 3 shows some examples.

Table 3: Examples of collaborative systems and their contribution factors (CFs)

System	CU	V	M	CE
Wikipedia	Articles	Direct: encyclopedia visitors	Ideological, Repetitive action,	Through a wiki environment
ESP Games	Image tags	Indirect Contribution	Enjoyability, Creditability	Online game
reCAPTCHA	One word from a book	Indirect Contribution	Repetitive action	Registration forms

This subsection presents each of these factors, to understand the requirements for a system dedicated to multilingual terminology creation, maintenance and use.

II.3.2.1 Volunteers (Contributors)

Volunteers are the most important factor. They are internauts often having fair computer skills. Their main task usually involves providing the system with specialized knowledge (knowledge easily known by humans, but very difficult to construct and extract by machines). Their contributions are usually validated through the wisdom of the crowd (wisdom of group of the volunteers).

II.3.2.2 Contributed Unit

A Contributed Unit is the smallest possible piece of knowledge received from a volunteer. In Wikipedia it is a part of an article, in reCAPTCHA it is a word from a scanned book. In general, the size, format and shape of the CUs can affect the success of the system. For example, a Wiktionary CU consists of a structured dictionary entry, which makes it more difficult for contributors.

II.3.2.3 Motivation

The motivation is the reason why a volunteer contributes to the system.

a. *Spontaneous*

A spontaneous contribution comes from the original task that the contributor is doing. For example, contributors in reCAPTCHA are not really motivated to digitize books, but their incentive is to create an account on some Website on the Internet. They are induced to contribute while they are creating such an account.

b. *Non-Spontaneous*

In the case of non-spontaneous contribution, the system is required to offer an incentive to the volunteers to contribute either by giving them encouragement for their contribution through online community ranking, or by giving the freedom and power to be influential, like in the case of Wikipedia.

II.3.2.4 Contribution Environment

The Contribution Environment (Shimohata, Kitamura et al. 2001) is the subsystem supporting the communication between the system and the human contributors. Usually, it is an online system that enables the user to contribute, like a wiki environment (Désilets, Gonzalez et al. 2006) (Desmet and Boutayeb 1994), BEYTrans (Bey 2008) (Bey, Boitet et al. 2006), or an online serious game.

II.3.3 **Collaborative Systems for Lexical Resources**

II.3.3.1 Particularities of Lexical Knowledge

This subsection gives an overview of contributive approaches for gathering knowledge in general. We are particularly interested in the use of collaborative work to gather lexical resources, in contrast to the previous two approaches (classical, and automatic). Based on the fact that many project succeeded to build large knowledge bases, we conjecture that is it possible to build a large base for multilingual “pre” terminology using human computations.

What is promising about this approach is that it is not limited by the abilities of professional terminologists nor by the availability of linguistic resources: it directly gathers knowledge from persons who use terminology. However, a lexical resource is in general inherently difficult to construct, because its CU should follow a specific structure, the volunteers should have a

minimal linguistic knowledge, and, in the case of multilingual terminology, they should have knowledge of the domain and the involved languages as well.

This subsection shows and discusses the limitation of many attempts to utilize these approaches to gather various lexical resources.

II.3.3.2 Current Systems

a. ITOLDU

It is a light web service used for the collaborative construction of a bilingual lexicon by a manageable small community (typically, a batch of students) and for a specific domain. In one year, 17,000 English-French terms have been created by 250 students (in paper, ink, and graphics industry), divided in 15 classes taught by 6 teachers (involved in the project). It could be adapted and used as a "front-end" to a quite more complex terminological database (Bellynck, Boitet et al. 2005). The limitation here is that users quickly find ways to deceive the system by scoring more points with less accurate contributions, as the main incentive is scoring points.

b. Wiktionary

It is a wiki-based collaborative project to produce a free, multilingual, general-purpose dictionary with definitions, etymologies, pronunciations, sample quotations, synonyms, antonyms and translations. Wiktionary (Wiktionary 2008) is the lexical companion of the open-content encyclopedia Wikipedia.

Currently, the English dictionary (En-N) contains 192,025 entries with definitions. A limitation is that the structure of a wiki-based document is that of a typical Wikipedia article, while it is a problem to consider it as an editable dictionary entry.

c. Papillon

The Papillon project aims at creating a large open-source multilingual lexical database (Sérasset 2004) (Mangeot and Sérasset 2007) (Sérasset 1994). Papillon-CDM[1] has about 2M entries from contributed dictionaries, in 8 languages (Japanese, German, English, French, Thai, Lao, Chinese, and Vietnamese). It has collaborative facilities that enable the volunteers to contribute to the dictionary through the web. On the technical side, it is built over the Jibiki platform (Mangeot 2006; Nguyen, Boitet et al. 2007), a sort of framework to develop web-based collaborative dictionaries and lexical databases of any structure.

d. JeuxDeMots

JeuxDeMots is a serious game that targets the construction of a monolingual lexical network collaboratively (Joubert and Lafourcade 2008). The concept of the game is rather simple: the system shows the player a word, and the player should input any word coming to his mind and related to the shown word, if his answers were similar to someone else's answers he will have more point. The result of this game is a lexical network with even functional lexical relations.

The French version now has 221,824 French terms (LIRMM 2010).

[1] CDM (Common Dictionary Markup) is an XML DTD able to represent the content of any entry in the open source online dictionaries encountered so far.

II.3.4 **Limitations**

The systematic problem in collecting lexical resources collaboratively lies in CU, *V* and *M*. The 3 attempts described above depended on non-spontaneous contributors, with no real motivation (except in the case of ITOLDU, where students get 1/3 of their English grades from the system).

A second common limitation is that the contribution unit (dictionary entry) is structured in a way that an average volunteer might not be able to deal with (this is a fundamental reason why Wiktionary is not as successful as Wikipedia).

Table 4: Examples of lexical collaborative systems

System	**Motivation for Contribution**	**Notes**
ITOLDU	Student graded activity	Engineering terminology
Yakushite (Murata, Kitamura et al. 2003)	Improve the suggested equivalents and the MT system itself	Bilingual dictionaries
Papillon	Dictionary consultation	General dictionary
JeuxDeMots	The fun of playing	Related word collection, not multilingual
Wiktionary	Wikipedia	Wiki-based

The challenge is not to motivate anybody to contribute, but to motivate those who have knowledge in the domain, and the languages, which has proven to be difficult. Average volunteers can not do the job of professional knowledgeable terminologists. Table 4 shows some examples of collaborative systems.

In short, the conclusion is that, if we are seeking a non-spontaneous approach, we should make sure that we provide a good reason for volunteers to contribute.

Problems and Discussions

Classical approaches in multilingual terminology development are not dynamic enough to handle the constant changes that happen in the terminological sphere of a domain. On the other hand, automatic approaches can improve the static state of classical approaches by relying on digital corpora, but they need large digital resources. Beside latent terminology is not always in digital format.

Table 5: Approaches comparison

Approach	Latent terminology	Absence of terminology
Classical	Terminologists' abilities, and references are limited to cover new terms in use	Classical approaches deal with this problem by human translation of terms, but translating absent term is

		difficult because it involves creating new terms for different languages, which is very expensive and needs domain experts' involvement
Automatic	Not all latent terminology are in digital format	Automatic approaches translate absent terms automatically, which is not sufficient as automatic resources do not contain absent terms
Collaborative	Very effective approaches for latent terminology, as humans are the source of latent terms. However it is very difficult to attract relevant contribution	Using crowd wisdom to establish terms is ideal. However, to come to a consensus is very difficult, unless contributors can actually meet (virtually on the web)

Table 5 shows each approach, and how it tries to solve the main problems in building multilingual terminology (latent and absent terms and), and its limitations.

Chapter III Preterminology

Introduction

The previous chapter presented the various approaches to build multilingual term bases and solve the problems of latent terminology and absent terminology. Another main problem is that the traditional definition of terminology emphasizes two standards:

- Terminology should capture the conceptual structure of the domain.
- In the modified terminological semiotic triangle, terminology should accompany detailed terminological and structural information.

This complicates the problem of building monolingual terminology, and terminology multilingualization.

To satisfy the above standards, it is difficult to build multilingual resources using any approach other than the classical one (which has many limitations).

In this chapter, we study alternative approaches, and an alternative definition of terminological resources that can be easily constructed, accept latent terminology, and resolve the absence of terminology.

This chapter starts by showing opportunities in constructing multilingual terminology. Then in section III.2 we define "preterminology" a new kind of terminological resource which is easier to construct using the opportunities shown in the first section. Then section II.3 presents the resources for preterminology. Finally, section III.4 shows possible applications of preterminological databases.

III.1 Exploiting Suggested Opportunities in Constructing Multilingual Terminology

This section discusses various unconventional resources and approaches that can be used in building unconventional terminological resources.

III.1.1 Recycling and Exploiting Digital Resources

III.1.1.1 Multilingual Terminological Databases (MTDs)

a. *Features*

Terminological databases such as (UN 2008) (FAO 2008), are built by large standardization bodies and international organizations. Although they tend to have less linguistic and informational coverage, and contain less hidden terminology, they focus on the quality and the completeness of terminological information. So it is important to use them as a seed.

b. *Technical Feasibility*

As MTDs are available online and accessible through http requests, such a service can be used as an automatic term translator as follows:

$Term_{source}$→http request→MTD (generate result)→html (to be analyzed)→ $Term_{target}$

III.1.1.2 General Purpose Online Multilingual Dictionaries

Although large online dictionaries do not aim at completeness in any technical domain, they contain the most frequent terms of many domains, because they are frequently found in text written for general purpose use. Bilingual dictionaries provide translations for such terms, and these translations may often be used as approximations, for example: minerai (fr) → ore (en).

Their linguistic coverage and online accessibility make them very convenient to easily build a first draft of absent terminology.

III.1.1.3 Machine Translation as a Tool to Build Lexical Resources

a. Machine Translation as Dictionaries

Many researchers have studied the possibilities of using MT systems as bilingual dictionaries (Nagata, Saito et al. 2001), which is sensible as these MT systems depend on lexically rich dictionaries, and most importantly sometimes they depend on linguistic resources like parallel corpora that contain many technical terms.

b. Technical Feasibility

The automatic use of such MT systems can be easily implemented (Vo-Trung 2004) (Callison-Burch and Flournoy 2001) as most MT systems have online services like (Google 2008). MT system can act as a term translator: it receives the term along with some other parameters (encoding, language pair) and it produces results in an html page to be analyzed to extract the translation.

III.1.1.4 Unconventional Multilingual Online Resources

a. Local Online Glossaries

For a specific domain, not only big organizations produce standardized terminology, but small businesses, websites, individuals can publish their own bilingual or multilingual terminology as glossaries. Recycling such glossaries is very important to target hidden terminology that is in use without proper recording.

b. Online Encyclopedias

A multilingual encyclopedia, notably Wikipedia, can be a very rich resource of terminology, because of the following reasons (Jones, Fantino et al. 2008) (Daoud, Kitamoto et al. 2008; Daoud, Kageura et al. 2009).

- It has good linguistic and lexical coverage. As of December, 2009, there were 279 Wikipedias in different languages, and 14,675,872 articles. There are 29 Wikipedias with more that 100,000 articles and 91 languages have more than 10,000 articles.
- It is built by domain experts to meet the requirements of their peers, hence the contents, including terminology, are quite acceptable.
- It is updated easily, and regularly, so that new terms that might not be found in dictionaries could be available in Wikipedia.
- Wikipedia's articles are organized and classified by the community.
- Its structure can be easily used to extract multilingual content.

III.1.2 **Human Interactions**

III.1.2.1 Overview

Adoptability of the entries of a term base is essential, even though the resources mentioned in the previous subsection are important; they need human validation and contribution (Popescu-Belis 2003) (CYC 2008) (Chklovski and Gil 2005-a). In this subsection, we present further approaches to compute human opinion on valid terminology.

III.1.2.2 Explicit Interaction

"Explicit" means that the volunteer is actively involved in the contribution process, whether he is aware of that (non-spontaneous) or not (spontaneous). These interactions are important for two tasks:

- lexical translation of terms.
- validation of automatic entries automatically (or by volunteers).

But validation needs either an expert opinion or a massive amount of contributions to build a consensus or opinion.

III.1.2.3 Implicit Interaction

The second cluster of interactions is the implicit one, where the contributor is not actively involved in the contribution process, but his actions are analyzed to extract trends and knowledge. Recently, there is an interest in analyzing users' experiences on the Web in an automatic way to find interesting facts (Mindpixel 2009) (Mihalcea and Chklovski 2003). For example, a recommendation system (recommender system) is an application that analyses users' behavior to build a knowledge base that associate Internet items (videos, pages, images, books…) with each other. Such knowledge base could also be built by analyzing access log files to find out patterns of interest. Indeed if many users are buying book A and book B, then there is a possibility these books are associated, so that next time a user buys book A, the recommender system will advise him to buy book B.

This technique seems promising to strengthen and boost the abilities of both the collaborative and automatic approaches.

III.1.3 **Structuring Preliminary Raw Data for Multilingual Terminology**

Building multilingual terminology is not only a process of compiling lists of lexical units, but it is a process of capturing the linguistic representation of the conceptual sphere of a domain which contains interrelated concepts (Desmet and Boutayeb 1994) (Temmerman 2000). This process can be further analyzed along 3 cases:

- the lexical unit of the term is recorded, but its information is not;
- the term is hidden;
- the term is absent.

This subsection discusses each case.

III.1.3.1 Recorded Terminology

As described in chapter II, recorded terminology is produced by large organizations and can be found in traditional terminological database. For example, a term like "North Atlantic Treaty Organisation (NATO)" is a well-established term that can be found in several databases. IATE

provide a description to that term and its translation in 15 languages, here is the En and Fr entries:

--------en----------

Term: North Atlantic Treaty Organisation
Reliability: 3 (Reliable)
Ref. http://europa.eu/smartapi/cgi/sga_doc?smartapi!celexplus!prod!CELEXnumdoc&lg=en&numdoc=32001E0555
Date: 02/06/2004
Abbreviation: NATO
Reliability: 3 (Reliable)
Ref. http://europa.eu/smartapi/cgi/sga_doc?smartapi!celexplus!prod!CELEXnumdoc&lg=en&numdoc=32001E0555
Date: 02/06/2004

---------fr----------

Term: Organisation du traité de l'Atlantique Nord
Reliability: 3 (Reliable)
Ref. http://europa.eu/smartapi/cgi/sga_doc?smartapi!celexplus!prod!CELEXnumdoc&lg=fr&numdoc=32001E0555
Date: 02/06/2004
Abbreviation: OTAN
Reliability: 3 (Reliable)
Ref. http://europa.eu/smartapi/cgi/sga_doc?smartapi!celexplus!prod!CELEXnumdoc&lg=fr&numdoc=32001E0555
Date: 02/06/2004

III.1.3.2 Unrecorded Terminology (Hidden)

Latent terms may exist in 2 cases.

a. *Digital*

1. A word or word combination may exist in digital format, but it not yet registered as a term anywhere, although it is used by domain experts.
2. No word (combination) exists to denote a concept in language L2, so speakers always borrow a foreign term in L1.

b. *Non-Digital*

The second type of latent terminology is the unrecorded terms that are not in digital format. Such terms are used by the community and recognized as terms, but are not available in any corpus. Thus automatic approaches can not detect them, and building correspondences needs extracting such terms from the people who use them.

III.1.3.3 Lacuna

We are interested here in the case of "total lacuna", i.e. when no latent term in L2 exists for a concept C, which happens for many concepts of new technologies, it is necessary to produce one or more candidates, using translation and adaptation. (Boitet and Nédeau 1997) argues that, to prepare a term in French, one should prepare a calque of the English term and avoid existing words. For example, Google Dictionary translates "butiner" (fr) by "to browse", which is a problematic translation, because butiner refers to bees and flowers, but "parser" (fr) is OK for "parse".

III.1.3.4 The Need for a Structure

Collecting all semantic relations between terms mirroring those between concepts is ambitious, and rather difficult. At the other extreme, producing only a stream of lexical units will reduce the term base to a list that does not really represent the nature of the conceptual system.

A compromise can be found: in fact, the unconventional resources discussed in this section can be used not only for finding multilingual terminology, but also to find its monolingual correspondences, as the next subsection will show.

III.2 Preterminology: Formal Definitions

Introduction

Traditional terminology has a specific definition of terminological resources which disallows the integration of unconventional resources and structures of raw lexical units. That is why a classical standard terminological repository suffers from a lack of linguistic and informational coverage, and can not deal flexibly with hidden or absent terminology. The problem becomes even harder when the community is smaller and does not have enough resources.

In this section, we introduce a new kind of terminological resource that can contain un-validated terminology obtained from unconventional terminological resources, called "preterminology" (Daoud, Boitet et al. 2009).

A preterminology can be easily constructed so that coverage is increased and cost is reduced. In this subsection, we will give the basic definitions concerning preterminology, and hints to develop them.

III.2.1 Multilingual Preterminological Sphere

III.2.1.1 Domain Specific Preterminological Sphere

The preterminological sphere is contained in the lexical sphere of a language and contains the terminological sphere, as shown in figure 22.

For developing the preterminological sphere, one tries to build correspondences between concepts and lexical units that are not yet recognized as terms (latent terms) for these concepts (a preterm is denoted by "L", see figure 22).

One also tries to bring lexical units closer to the preterminological sphere by building correspondences between them and the concepts without associated terms to resolve the absence of terminology; such terms are denoted by "A" in figure 22.

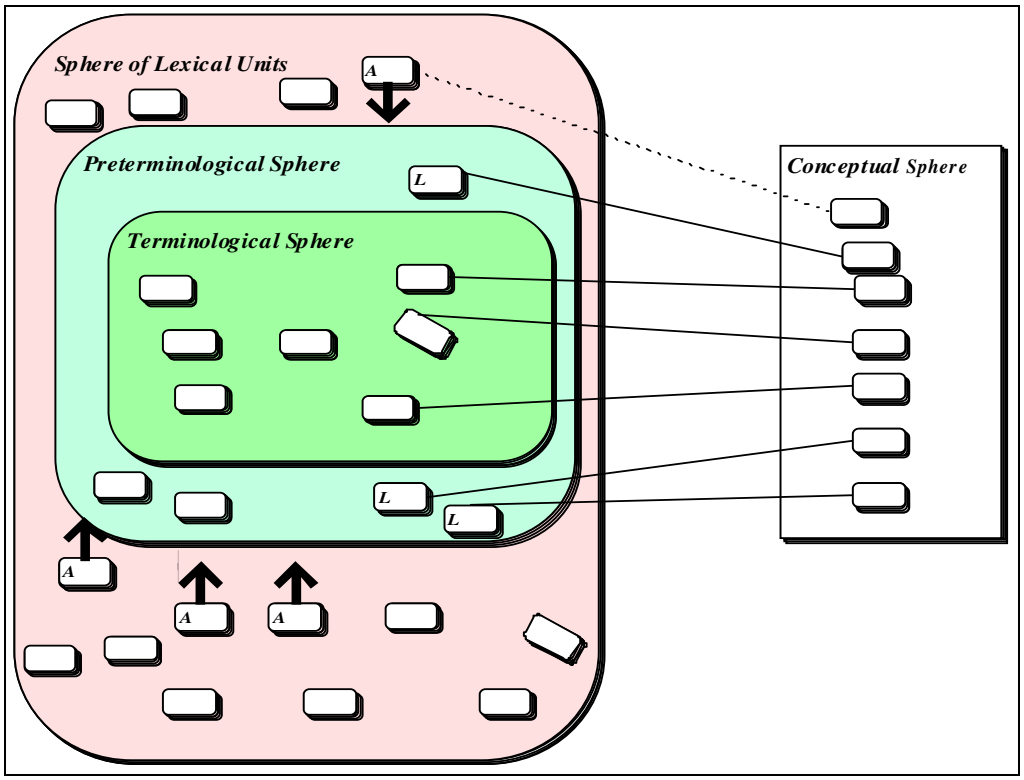


Fig. 22: Preterminological sphere

III.2.1.2 Multilingualism

A multilingual preterminology is the set of the preterminological spheres of a domain for certain languages, with correspondences between the preterms on those spheres. From a practical point of view, the developer of a multilingual preterminology does not deal with the conceptual sphere directly, but rather with a preterminological sphere of a language with rich resources.

Developing a multilingual preterminology involves moving the lexical units towards the preterminology of a domain. While the term information and definition are essential in the classical semiotic triangle of terminology (discussed in chapter I), the modified system I am proposing for preterminology focuses on the multilingual signs and their correspondences.

As shown in Figure 23, the concept itself is not dealt with as a separate entity, but it is denoted using an established term in some well- resourced language. That is more realistic, as concepts are usually presented to us as terms.

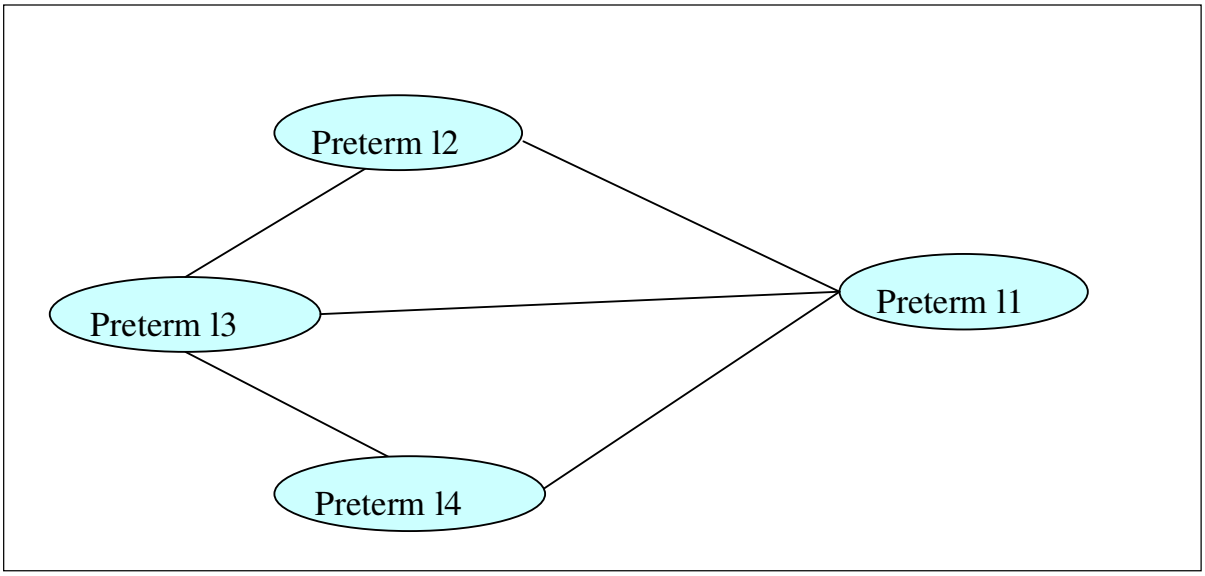


Fig. 23: Multilingual preterminology

III.2.1.3 Latent Sphere

The latent sphere is the sphere of those lexical units that have been used as terms without being identified yet as terms. It is located somewhere in the lexical sphere, but it is not contained in the terminological sphere, otherwise its elements (preterms) would already have been registered terms.

Having no associated terminological information, latent terms do not satisfy the semiotic triangle of terminology and cannot be simply added to a terminology by standardization bodies. While that information is not added, these terms must remain in the preterminological sphere, as they represent raw material for terminologists.

In with the absence of terminological information, we should utilize the resources in the lexical sphere to structure the preterminology, even if the terminological and conceptual relations are not yet available.

III.2.2 **Preterms**

Having defined the general context of preterminology and its sphere; this subsection gives a detailed presentation of the main concepts associated with a preterminology, and the corresponding terms we will propose for them.

III.2.2.1 Definition

a. *Preterm*

A preterm is a lexical unit of one or more words that denotes a concept. Contrary to a term which is a validated and standardized sign and has associated terminological information, a preterm is only an un-validated lexical unit that can denote a new concept.

b. *Preterminology*

A preterminology is a set of preterms corresponding to a domain.

III.2.2.2 Preterminological Information

Preterminology can have the following information.

a. ***Preterm***

A preterm is a string of characters constituting a word or a compound word that corresponds to a concept.

b. ***Relations***

Preterms may be linked by relations that are in general less precise than the terminological relations. For example, "X is more specific than Y", (kind of, instance of) relation. Table 6 shows the precise terminological relations and their preterminological interpretations.

Table 6: Correspondences between terminological and preterminological relations

Term. relations / Preterm. relations	Multilingual synonymy	Synonymy	General
Translational equivalence			
Synonymy			
Acronym			
Hyponymy			
Antinomy			
Hypernymy			
Other (related, substance of, …)			

c. ***Context***

The context of the preterm can be given by the source of the term.

d. ***Terminological Information***

If there is already terminological information, it should be preserved.

III.2.2.3 Examples

Figure 24 shows an example of some terms used in the domain of social networks. While some of these concepts are denoted by established terms in Arabic, like E-mail and Chat, some other new concepts have hidden terms, used by the people who use social networks.

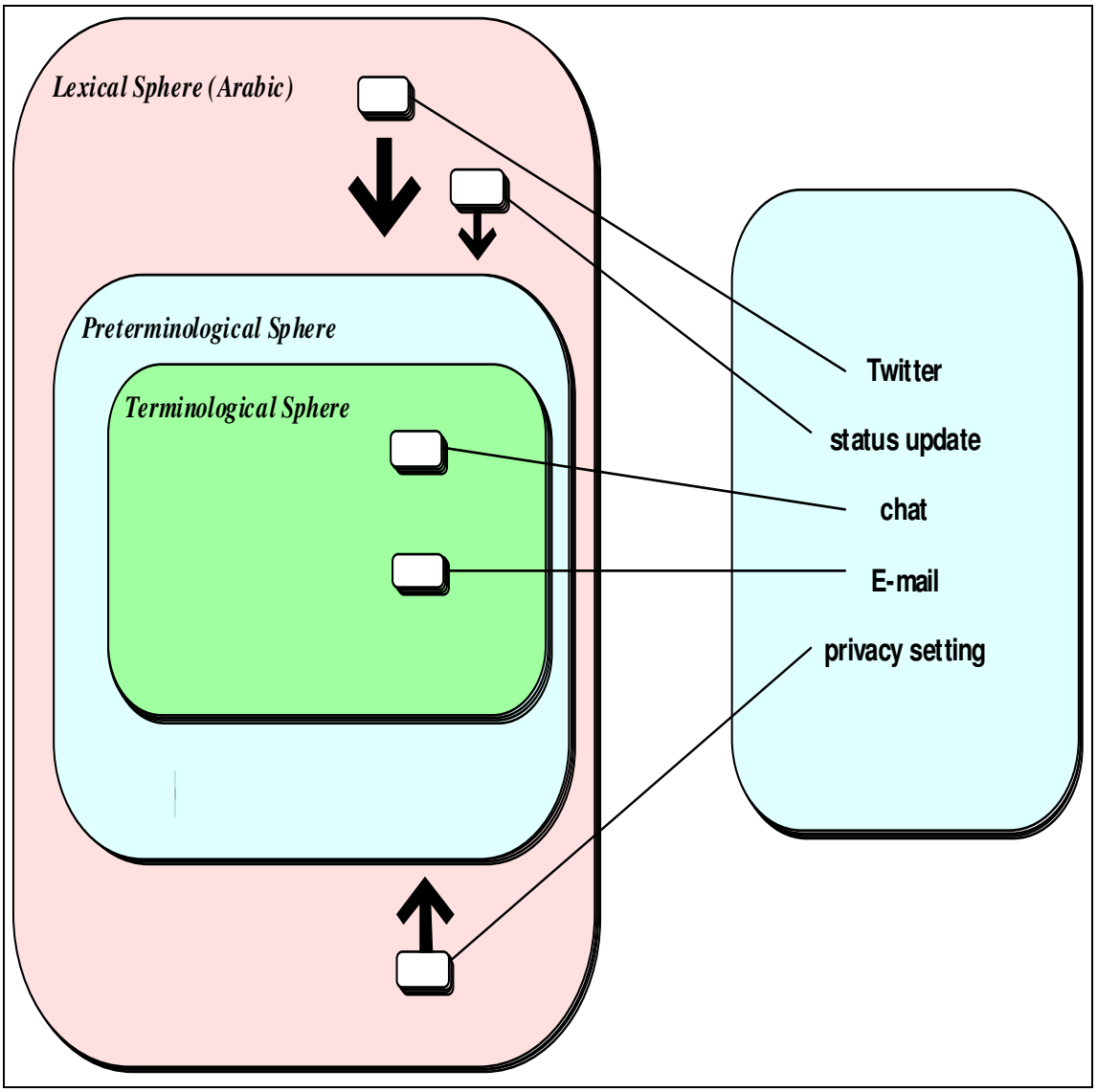


Fig. 24: Latent terminology and terminological gap

III.2.3 **Multilingual Preterminology**

This subsections deals with the particularities and details of multilingual preterminology.

III.2.3.1 Definition

a. *Multilingual Preterminology*

Multilingual preterminology is the collection of monolingual preterminological spheres of a specific domain connected with translational relations between translationally equivalent preterms.

b. *Multilingual preterm*

A multilingual preterm is a set of translationally equivalent lexical units that correspond to a concept in a domain.

III.2.3.2 Intrernal Correspondences

Figure 25 shows the bilingual terminology of a domain for 2 languages, L1 and L2.

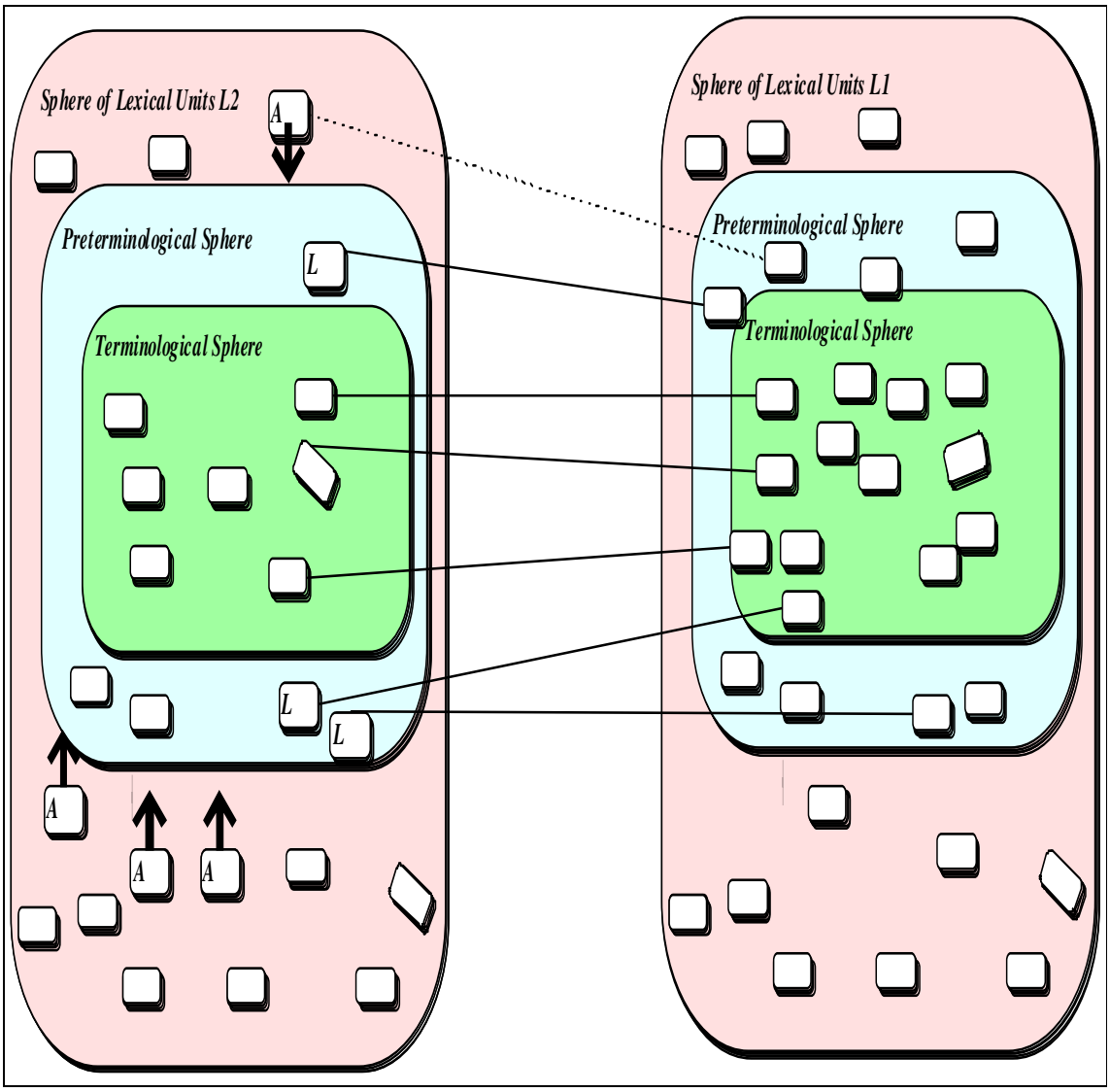


Fig. 25: Preterminological gap

Note that the correspondences between the two spheres are not between concepts but between preterms; such relations are translation relations. In figure 25, L1 has more terminological resources, so it acts as the source of conceptual information. The construction process of any new preterminological sphere involves the utilization of the preterminology of L1, as the preterms of L1 are translated into lexical units of L2; lexical units are promoted to preterms and move towards the preterminological area (dotted line in figure 25).

III.3 Resources of Preterminology

There are three goals for preterminology (Daoud, Boitet et al. 2009):

- making good use of what is standardized and available;

- making good use of hidden terminology, whether it is digitized (available in corpus) or not;
- making good use of lexical unit that can represent absent terms.

So, one distinction between the sources is whether preterms are in digital form or are only used by the community without recording.

Hence, possible resources for preterminology are humans, computerized lexical resources, online dictionaries, and special files containing lexical units like logs of visits to websites.

III.3.1 **Human Resources**

III.3.1.1 Communities

A community is a group of people with common characteristics and interests. Recently, the notion of an online community became very important. An online community is a virtual community that exists in the Internet and whose members enable its existence through taking part in membership rituals (Kim 2000). An online community consists of an information system where members can post content, such as a Bulletin Board system or one where only a restricted number of people can initiate posts, such as Weblogs. Online communities offer means of communication between people even if they do not know each other in real life.

Online forums, social networks like (Facebook 2010), and even news portals like CNN.com and Aljazeera.net, have a community of loyal groups of members having a common interest or objective.

Such human resources could contribute directly or indirectly to preterminology.

III.3.1.2 Individuals

There are several categories of persons who can contribute to a preterminological base for a domain.

- Domain experts: their contribution and validation is essential, but they are difficult to find and motivate.
- Terminologists, who know how to build new terms.
- Casual term users: casual users are not domain experts but they have a certain interest in the domain, and they use the terminology of the domain (without actively participating in the development and validation of such terminology). What is important about them is that they are not rare as domain experts, and more inclined to contribute.
- Translators: translators spend a lot of effort researching and studying references to know the meaning of a term or expression. Their experience and legacy of previous translation tasks is very important and useful for preterminology.

III.3.2 **Digital Content**

Digital resources are very important to build a preterminology. We can distinguish between three kinds of potentially useful digital resources: lexical, textual, and non-textual.

III.3.2.1 Lexical Resources

Lexical resources are the easiest to extract and identify.

They contain general-purpose lexical units and their translations (Etzioni, Reiter et al. 2007), however it is common that a lexical unit gives an indication of the meaning of the term. Figure 26 is a screenshot from Google Dictionary showing the meaning of "e-mail"

Fig. 26: Google Dictionary "E-mail"

III.3.2.2 Textual Resources

Textual resources contain lexical units in a very complicated set of expressions, which makes it more difficult to extract the relevant lexical units. However, a textual corpus dedicated to a domain is usually produced by persons who are familiar with the domain and its terminology, so utilizing it will gives a rich resource of lexical units that qualify as preterms.

Textual material can be found as a result of the following:

- Publication: books (digitized books also), journals, webblog… any publication by the community of the domain is bound to contain many relevant lexical units.
- Communication: discussions amongst the community members are also important as they contain casual language that has lots of un-validated and hidden terminology.
- Education: tutorials and training material related to the domain can be used as a source of preterminology.

III.3.2.3 Non-Textual resources

Textual formats are understandable by an average human being, however in the era of machine, much digital information is stored in a way that is not comprehensible by humans, but it is meant to be useful for computers. However, some of these resources can be used as a source of terminology (Barrière 2009).

In particular, one can use access log files or server logs (Stermsek, Strembeck et al. 2007) which are digital documents automatically created and maintained by a Web server that registers all the activity performed by it. A typical example is a web server log which maintains a history of page requests, and referred sites.

Recently the majority of the traffic on the Web depends on search engines, and keywords and their logs register search engine requisites and referrals, which contain search terms used by users to access the desired content (Richardson 2008). These search terms usually target particular concepts, so in a way the search terms can show how the general public names certain concepts.

III.4 Applications of Preterminology

As we discussed above, terminology is used in a variety of important applications. Despite the incompleteness or absence of the terminological information in preterminology, a preterminology can still be useful for most of these applications.

III.4.1 Education

Starting from kindergarten, for each level and for each subject, teachers introduce new concepts through terms. Hence, building a glossary for each level is essential to prepare the students for the next level.

Then, starting from junior high school, students can not rely only on school books and their glossaries. They need to use external references, often from the Internet, and in the case of speakers of a poorly informatized language, students need to use references in foreign languages. Thus, the classical school book glossary (even if it has full terminological information) is not enough. A multilingual preterminological resource can enrich the classical glossary and help student find external educational resources.

III.4.2 Knowledge Transfer

(Argote and Ingram 2000) defined knowledge transfer as the process through which one group is affected by the experience of another. One of the problems and difficulties in knowledge transfer is the language. Because knowledge is mainly interpreted in a symbolic representation (textual), such representation depends heavily on special concepts and their terms, transferring such knowledge into another language should involve finding a term base for that knowledge. This may be the main reason for making the process of transferring knowledge into “pi-languages” (poorly informatized) very difficult, especially in technical domains.

To demonstrate the size of the problem: the translation of EOLSS (EOLSS 2008) (Encyclopedia Of Life Support Systems), which contains about 200,000 pages, probably requires to find or create translations for at least 250,000 Arabic terms, an effort which is estimated at about 25,000 working hours.

III.4.3 Consultation Service

Preterminology is a lexical resource which is supposed to offer lexical translation and consultation service, although it does not have complete terminological information, it has better coverage to serve the following services:

- general purpose consultation: particularly for latent terminology;

- professional lexical translation: professionals need confirmed translation, however coverage is essential;
- search term translation: no need to precise terminological relations, however a large multilingual lexical resource like preterminology can be effective;
- reference for translators: when translators encounter new terms or absent terminology, they can not find any reference in order to translate the encountered term;

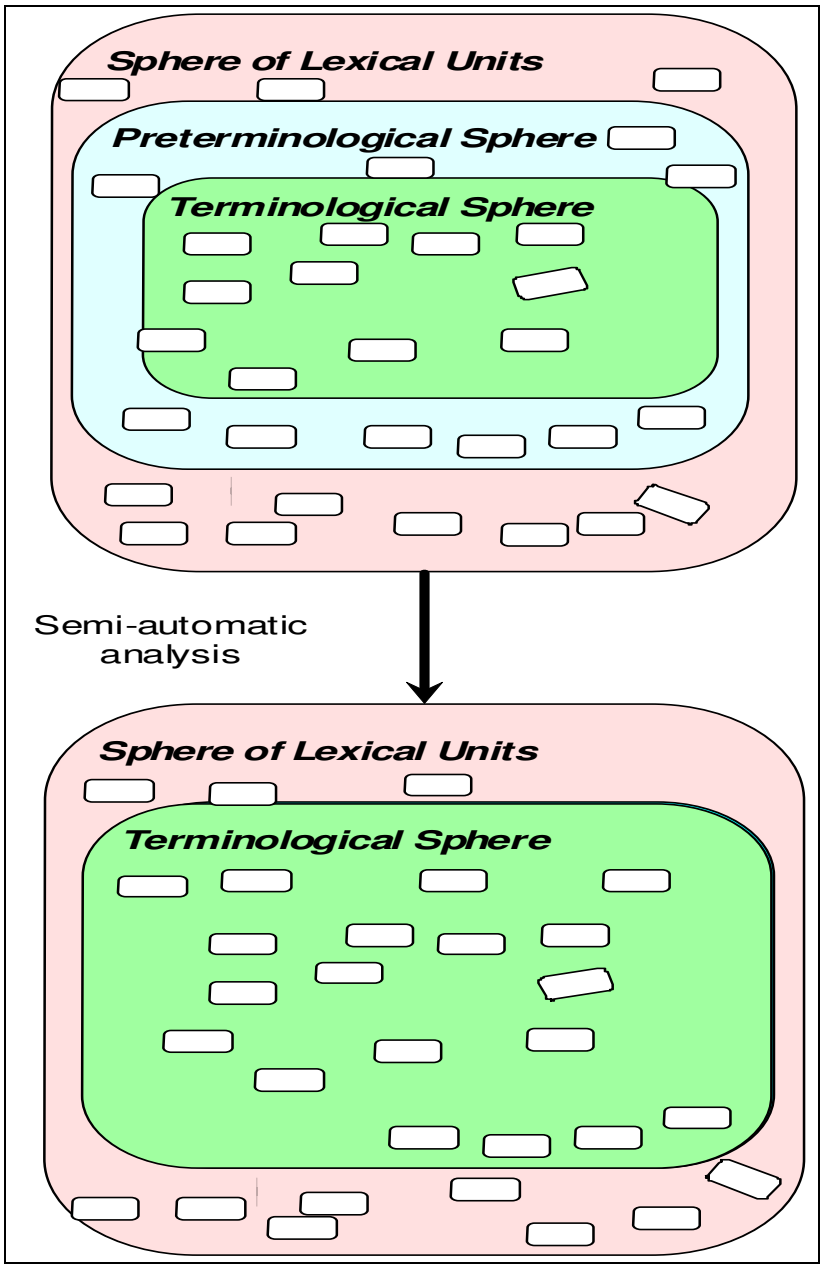


Fig. 27: Preterminology to terminology

Conclusion

Conventional terminology has a precise definition of terminological resources which disallows the integration of "preterms" found in unconventional resources (presented in this chapter, like

implicit and explicit interactions with possible human contributors) and of "raw" lexical units. That is why a classical standard terminological repository suffers from lack of linguistic and lack of informational coverage, and can not deal flexibly with hidden or absent terminology, even though they are built by large standardization and international bodies (UN, EU…) with huge resources. The problem becomes even bigger when the community is smaller and does not have such resources.

This chapter presented a novel terminological resource that can associate un-validated terminology from unconventional terminological resources into a base of preliminary terminology, called "preterminology". Preterminology can be easily constructed and it aims at increasing the coverage and reducing the costs of constructing terminology. Figure 27 illustrates the process of "paving the way to terminology building by collecting preterminology".

Preterminological work structures simple or compound words unrecognized as terms in the lexical sphere by analyzing digital (non-textual) resources, and involves the community in easy (non-linguistic) contribution activities.

Conclusion of Part A

A term is a linguistic representation of a concept. A terminology is the vocabulary of a particular domain. One can speak of the "sphere of terms" that describe and represent the concepts of that domain. An **object** or an idea is perceived by the human brain as a **concept**, and represented as a **term**. This triangle has been modified to add terminological information, particularly definitions, to situate the term within a terminology, and to define its relations with other terms. Multilingual terminology is a collection of such spheres of terms representing the same domain in different languages. Usually such spheres are maintained and managed by a computerized database system, a "term base", that organizes terminology and offers functionalities like modification and retrieval.

Concept systems are usually more dynamic than their representation. This affects the lexical coverage of a multilingual term base. Linguistic resources vary, that is why not all multilingual spheres cover the same amount of concepts. In fact, the lack of lexical and linguistic coverage is a big issue in multilingual terminological repositories. Many concepts have a representation in language A, but do not have any representation in language B (Lacunea). On the other hand, we have seen that many communities do not depend on terminological repositories to maintain their terms, but they find equivalent terms and use them without proper registration (latent terminology).

This part showed various approaches to solve these problems. Classical approaches in building multilingual terminology depend heavily on terminologists, which increases the cost, and affects the coverage as the terminologists' abilities are limited. On the other hand, collaborative approaches try to replace terminologists with amateur volunteers, which is a promising trend in acquiring knowledge. However, volunteers are usually not able to perform a sufficient linguistic activity, even if they are familiar with the domain. Automatic approaches use textual and lexical resources to develop terminology, but are limited to the available resources and the used techniques. Also, not all terminological knowledge is available in textual format (latent terminology). Finally, automatic approaches might be effective in finding terms, but it is difficult to build the correspondences.

Building terminology with the specific traditional definition and requirements is very difficult. That is why Part A suggested the development of a novel terminological resource that can associate un-validated terminology from unconventional terminological resources into a base of preliminary terminologies; called (preterminology). Preterminology can be easily constructed and it aims at increasing the coverage and reducing the costs of constructing terminology.

Part B Multilingual Preterminology: Maintenance and Structure

Introduction à la partie B « Préterminologie multilingue : gestion et structure »

La terminologie multilingue est difficile à construire en raison de ses définitions et de ses approches rigides traditionnelles. C'est pourquoi les bases terminologiques traditionnelles souffrent d'un manque de couverture lexicale et linguistique. Le concept de préterminologie a été proposé dans la partie précédente pour tirer parti des ressources terminologiques non reconnues et étendre la couverture en trouvant la terminologie latente et en comblant ainsi le fossé terminologique.

Cette partie donne plus de détails sur la préterminologie, sa structure, et ses aspects techniques. Construire des correspondances (terminologiques ou traductionnelles) est l'une des principales raisons de l'augmentation de la difficulté de produire une base terminologique, car cela exige des connaissances spécifiques au domaine et linguistiques. Cette partie présente l'utilisation d'une structure de graphe qui gère les correspondances d'une manière facile. Elle traite également de méthodes pour construire de tels graphes.

Cette partie est organisée comme suit. Le chapitre IV introduit les graphes multilingues préterminologiques, le chapitre V présente les techniques et approches dans l'élaboration de ces graphes, et le chapitre VI décrit la conception d'un système pour développer et maintenir une préterminologie (SEpT).

Introduction to Part B

Multilingual terminology is difficult to construct due to its rigid traditional definitions and approaches. That is why traditional term bases are suffering from a lack of lexical and linguistic coverage. Preterminology was proposed in the previous part to take advantage of unrecognized terminological resources and expand the coverage by finding latent terminology and bridging the terminological gap.

This part gives more details about preterminology, its structure, and its technical aspects. Building correspondences (terminological or translational) is one of the main reasons of increasing the difficulty of producing a term base, because it requires domain-specific and linguistic knowledge. This part introduces the utilization of a graph structure that handles the correspondences in an easy way. It also discusses methods to construct such graphs.

This part is organized as follows. Chapter IV introduces multilingual preterminological graphs, chapter V shows the techniques and approaches in developing these graphs, and chapter VI describes the design of a system to develop and maintain preterminology (SEpT).

Chapter IV Multilingual Preterminological Graphs

Introduction

Structuring preterminology is an important step to make it useful. A terminology, in general, consists of terms structured by confirmed logical relations. Adopting this structure is expensive and requires a heavy involvement of domain experts (Claveau and L'Homme 2005). On the other hand, developing a flat terminological structure will decrease the actual correspondence between the terms and the concepts.

This chapter investigates possible choices for the structure of preterminology, taking into consideration its objectives and resources.

The first section presents various choices for structure, the second proposes a graph structure for preterminology, and the third discusses the operations on such a graph.

IV.1 Structure of Preterminology

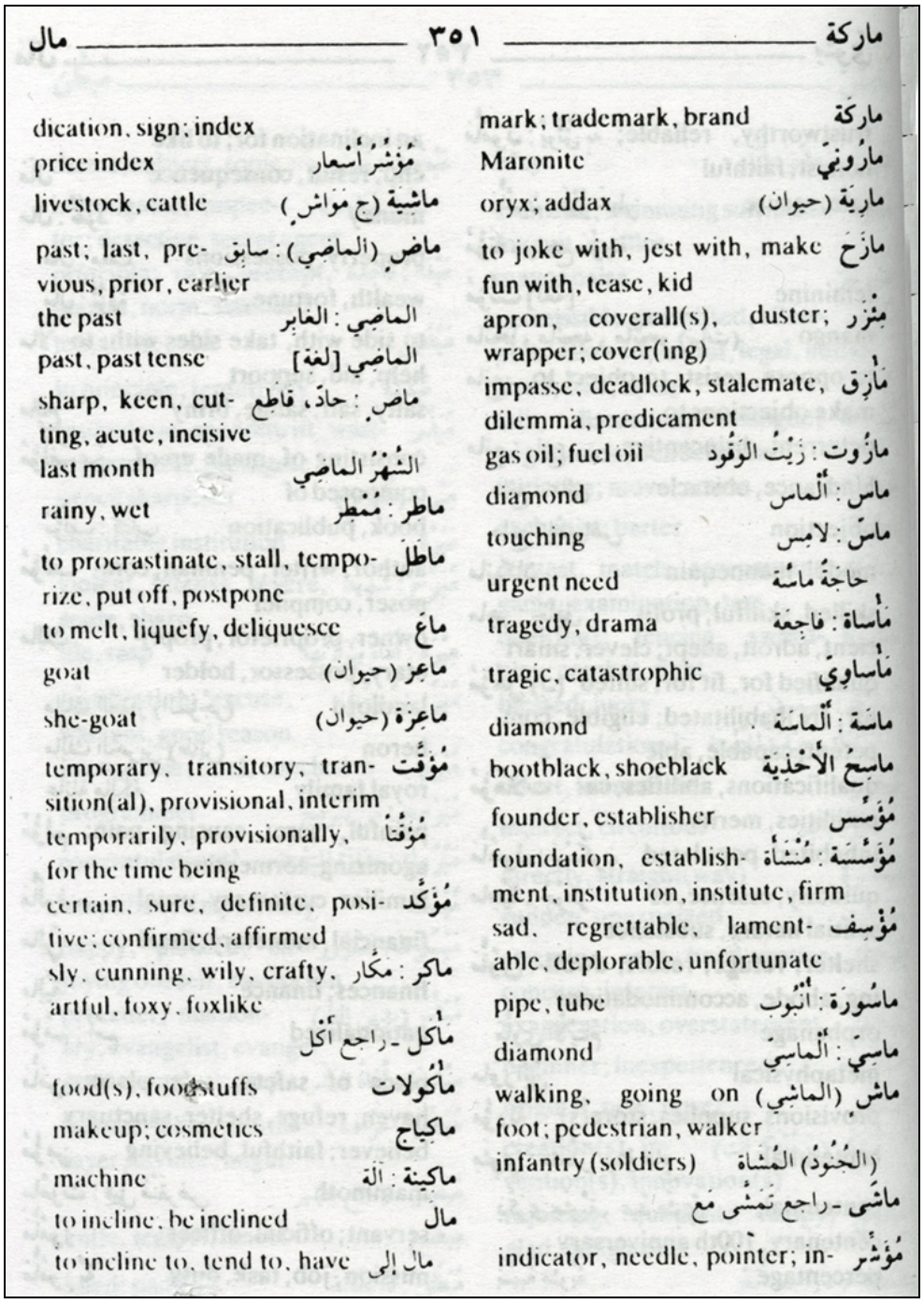
We first examine possible choices for structuring a "preterminological entry", then the possible relations between them, and finally the necessary requirements such as confidence level.

IV.1.1 Possible Choices

IV.1.1.1 Flat Lexical-Based Structure (Gibbon 2002)

a. *Semasiological Structure (Orthography-Based)*

The semasiological orthographic structure deals with terminology of a domain as a list of lexical units, and orders them (alphabetically) according to their orthography (ISO-6156 1987), (ISO-1087-1 2000). This is the most common way of organizing dictionaries and encyclopedias. Figure 28 shows an example of an Arabic-English dictionary that uses this flat structure (Baalabaki and Baalabaki 2007). Here, the Arabic words are ordered alphabetically according to their triconsonantal roots.



مال ٣٥١ ماركة

dication, sign; index	
price index	مُؤشِّرُ أسْعار
livestock, cattle	ماشِيَة (ج مَواشٍ)
past; last, previous, prior, earlier	ماضٍ (الماضي) : سابِق
the past	الماضِي : الغابِر
past, past tense	الماضِي [لغة]
sharp, keen, cutting, acute, incisive	ماضٍ : حادّ، قاطِع
last month	الشَّهْرُ الماضِي
rainy, wet	ماطِر : مُمْطِر
to procrastinate, stall, temporize, put off, postpone	ماطَلَ
to melt, liquefy, deliquesce	ماعَ
goat	ماعِز (حيوان)
she-goat	ماعِزَة (حيوان)
temporary, transitory, transition(al), provisional, interim	مُؤَقَّت
temporarily, provisionally, for the time being	مُؤَقَّتاً
certain, sure, definite, positive; confirmed, affirmed	مُؤَكَّد
sly, cunning, wily, crafty, artful, foxy, foxlike	ماكِر : مكّار
	مأكل - راجع أكل
food(s), foodstuffs	مأكولات
makeup; cosmetics	ماكِياج
machine	ماكِينة : آلة
to incline, be inclined	مال
to incline to, tend to, have	مال إلى
mark; trademark, brand	ماركة
Maronite	مارُونِيّ
oryx; addax	مارِيَة (حيوان)
to joke with, jest with, make fun with, tease, kid	مازَحَ
apron, coverall(s), duster; wrapper; cover(ing)	مِئْزَر
impasse, deadlock, stalemate, dilemma, predicament	مأزِق
gas oil; fuel oil	مازُوت : زَيْتُ الوَقود
diamond	ماس : ألماس
touching	ماسّ : لامِس
urgent need	حاجَةٌ ماسَّة
tragedy, drama	مأساة : فاجِعَة
tragic, catastrophic	مأساوِيّ
diamond	ماسَة : ألماسَة
bootblack, shoeblack	ماسِحُ الأحْذِيَة
founder, establisher	مُؤَسِّس
foundation, establishment, institution, institute, firm	مُؤَسَّسَة : مُنْشأة
sad, regrettable, lamentable, deplorable, unfortunate	مُؤْسِف
pipe, tube	ماسُورَة : أنْبُوب
diamond	ماسِيّ : ألماسِيّ
walking, going on foot; pedestrian, walker	ماشٍ (الماشِي)
infantry (soldiers)	(الجُنُودُ) المُشاة
	ماشَى - راجع تَمشَّى مَع
indicator, needle, pointer; in-	مُؤَشِّر

Fig. 28: Physical dictionary

b. *The Case of Preterminology*

A terminology is a concept-driven lexical resource and so is a preterminology. Therefore, an orthographic representation will not capture the conceptual relations between preterms. This way preterms have to be identified not only by their orthographic representation (which is not necessarily common). But also by their location on the conceptual sphere, accessing them is more difficult. Some complementary access methods must be found, but as a preterminology incorporates more than lexical data, we should benefit from that.

For example, “object oriented programming language” is a generalization for “java”. An orthographic structure will ignore that and both lexical units will be presented separately without a logical relation, but, in a preterminological graph, one will put labels on 2 nodes in a graph and link them by an arc marked “generalized”.

IV.1.1.2 Hierarchical Structure

a. Onomasiological Structure

An onomasiological structure (Gruber 2009) is a concept-based hierarchical structure that does not follow the orthographic arrangement. A term is identified according to its semantic fields and its location in the conceptual sphere (Mariam, Gillam et al. 2005) (Cruse 1986; Aussenac-Gilles and Sörgel 2005).

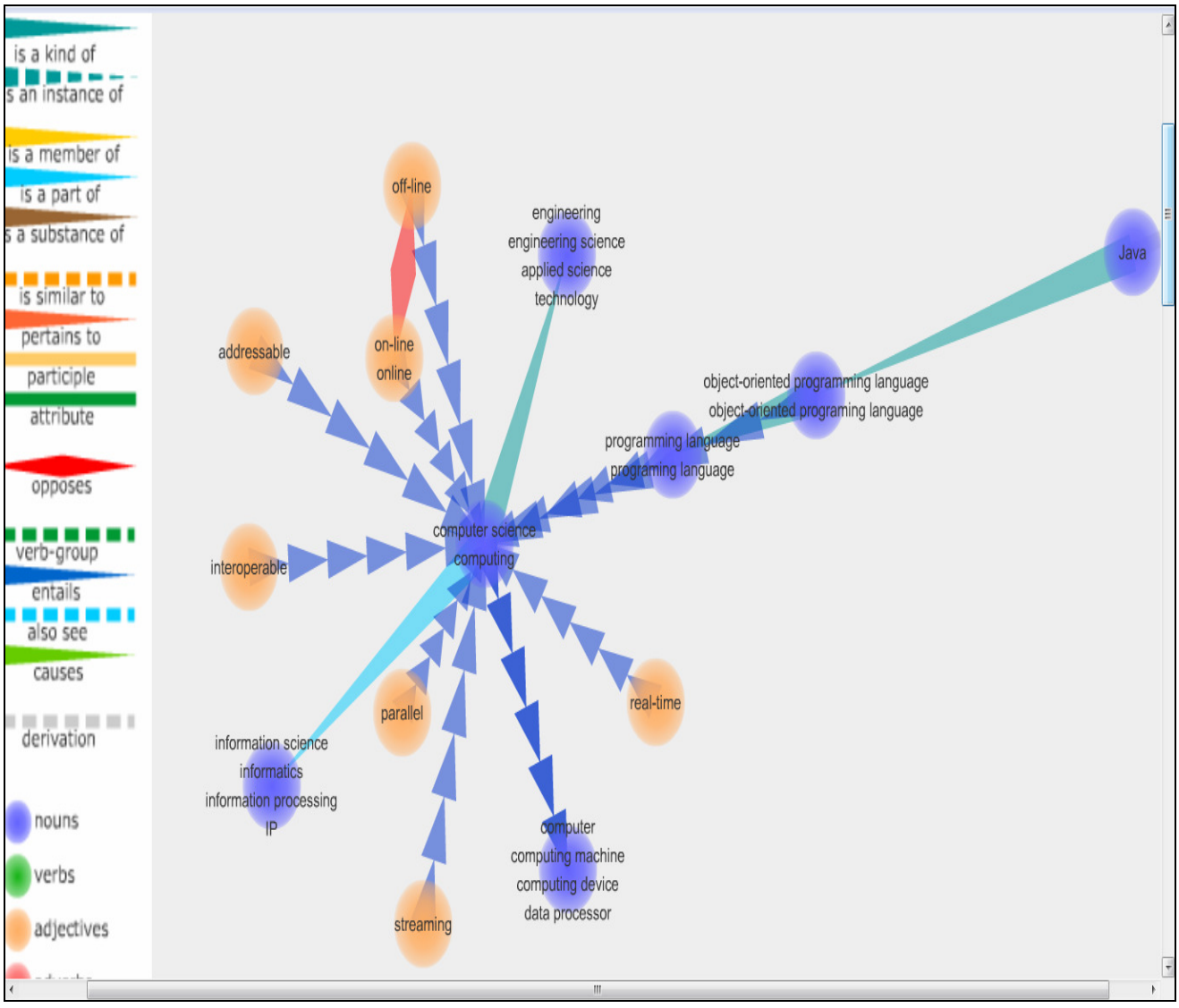


Fig. 29: WordNet as a lexical graph

Figure 29 shows the logical relation between “java”, “object oriented programming language”, “programming language” and “computer science”. This way, even if one does not know the term “java”, one can identify it according to its relations with other more familiar terms.

b. The Case of Preterminology

Although preterminology is concept-driven, adopting an onomasiological structure is not technically realistic for a huge term base with high lexical coverage and dynamic domain, because it requires the involvement of specialists. Beside, at the macrostructure, level the problem becomes more complicated, as matching monolingual onomasiological resources with different amounts of data is likely to be very difficult.

As we are targeting a high lexical and linguistic coverage using unconventional methods and resources, we will not adopt this structure.

IV.1.1.3 Hybrid Structure (Graph)

For the sake of increasing the coverage and associating unconventional resources, a preterminology can tolerate the presence of un-validated correspondences between preterms and concepts, as well as un-validated relations between preterms. A stricter approach will disallow many possible relations while the other extreme will introduce inaccuracies.

The structure of a preterminology should preserve the confirmed relations extracted from standard terminological resources, and should allow possible relations to be confirmed later. That is why we will keep a record of terms and their relations, and associate a confidence level to the extracted relations (see figure 30).

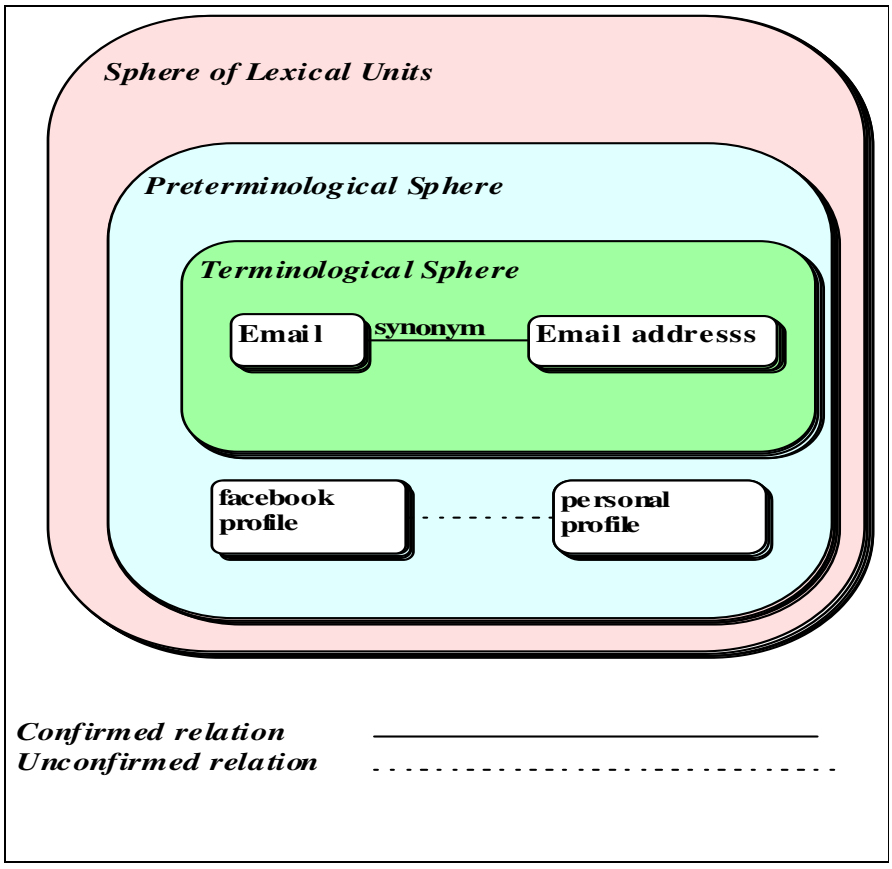


Fig. 30: Relations in a preterminology

IV.1.2 **Relations between Preterms**

To understand the real requirement for the structure of a preterminology, this subsection presents the relations that a multilingual preterminological resource should preserve in order to make good use of what is available in terminological resources.

We are interested in two kinds of relations:

(1) multilingual synonymy: deals with translational correspondences between preterms

(2) Lexical correspondences: applying general unconfirmed links between preterms to represent terminological relations, such as synonymy, hyponymy…

IV.1.2.1 Multilingual Correspondence between Preterms

a. *Multilingual Synonymy*

A preterm T_{La} is a translation of T_{Lb} if both represent the same concept in languages La and Lb; figure 31 illustrates this.

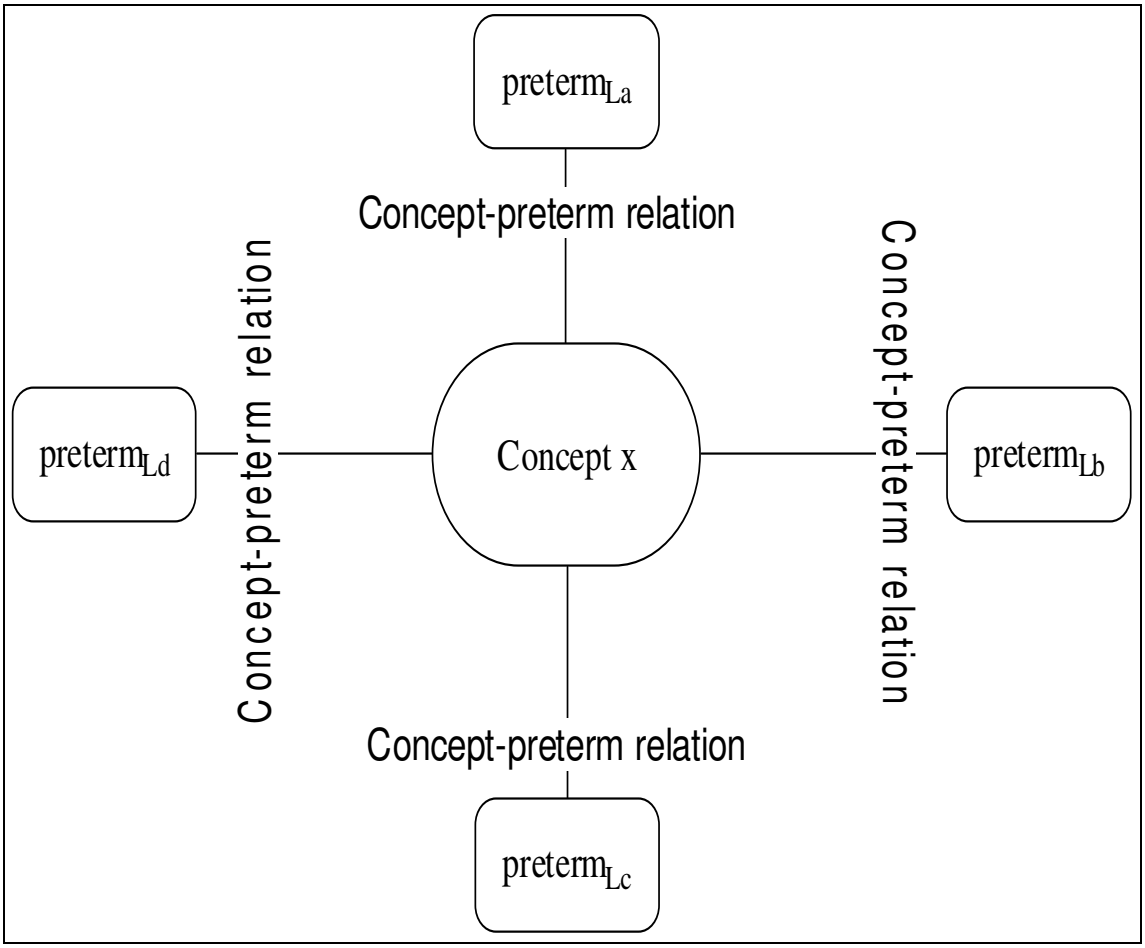


Fig. 31: Concept and multilingual synonymy between preterms

Developing multilingual preterms by connecting concepts with terms is unrealistic because there is no unified concept representation amongst the monolingual sets. However, a concept is usually represented in the form of a foreign term in a different language and it can go through a

bilingual translation process which is a relatively easier task, considering the availability of bilingual lexical resources and translators (see figure 32).

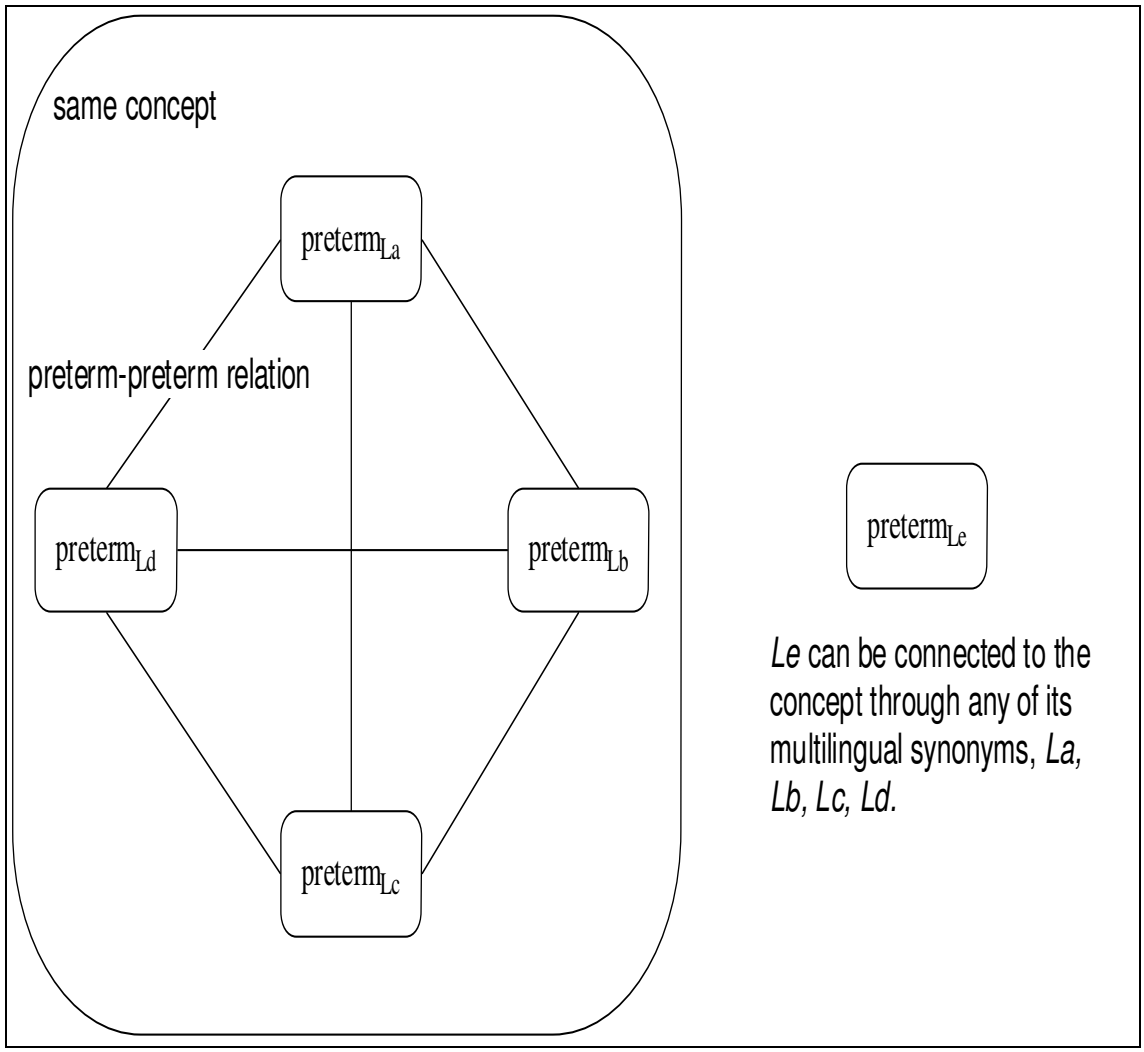


Fig. 32: Synonymy-based preterminology

b. Translation Candidates

A preterm-preterm translation relation can be built from various resources. All the possibilities should be considered as candidates. However, a preterminology support tool should implement a confirmation technique that would distinguish the more accurate translation candidates based on multiple resources validation.

IV.1.2.2 Lexical Correspondences between Preterms

As shown in chapter III, preterms depend on diverse resources with various levels of formality and completeness. There are many cases where a resource gives an indication of a possible relation between two terms without knowing the nature of that relation. We do not want to lose that information, but such relations are not specific and not well defined. Hence, beside the usual terminological relations, we will use a default relation between preterms that states that there is a relation between a pair of preterms, without defining it exactly. This information can be accumulated and processed later to transform it into a confirmed sensible terminological correspondence.

IV.1.2.3 The Needed Capabilities to Represent Preterminology

The capabilities required to represent a multilingual preterminology, are summarized below.

a. *Diversity of Resources*

The structure should handle diverse resources, from digital content to human-based contributions.

b. *Confidence Technique*

The desired structure should implement a confidence technique to differentiate between reliable and unreliable resources, and confirmed and unconfirmed relations.

c. *Elicitation Capabilities*

The desired structure should record all the information available in the resources, and it should provide the possibility of exploiting such information for the sake of eliciting and expanding terminology.

d. *Comprehensibility*

The structure should be comprehensible by average contributors, and make human contribution possible.

e. *Usability and Automatic Maintainability*

The structure should produce preterminological resources that can be used and maintained by others in a systematic way.

IV.2 Graph Structure and Components

IV.2.1 Overview: the Choice of a Graph Structure

A graph is capable of satisfying the requirements of structuring a multilingual preterminology because of the following:

- graphs are easily understandable by humans;
- a graph is easy to process and maintain automatically;
- one can handle diverse resources through labeling its components (nodes and edges);
- it is flexible enough to deal with diversity in the quality of the resources;
- graphs have a solid theoretical background and there are many tools to manipulate them.

Therefore a graph structure can handle preterminology and it can be easily constructed, manipulated and adopted by the people in the domain. This subsection presents our model of Multilingual Preterminological Graphs (MPGs) and its components, which can represent the preterminological sphere of a particular domain.

IV.2.2 Graph Theory Overview

IV.2.2.1 Formal Definitions

An undirected graph G(N,E) is a set of nodes N (also known as vertices) and a set E of unordered pairs e=(n1,n2) of elements of N called edges or arcs (also known as arcs). Figure 33 shows an example of undirected graph (Loerch 2000). In such a graph, (n1,n2) and (n2,n1) denotes the same edge.

An arc e=(n1,n1) is a self-loop (buckle) of the node n1.

Two nodes n1 and n2 are adjacent in G if (n1,n2) ∈ E.

An arc e=(n1,n2) is adjacent to a node n if n1=n or n2=n.

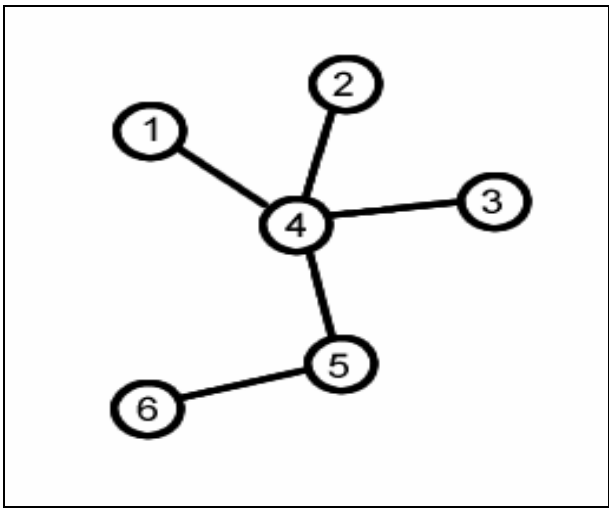


Fig. 33: An Example of an undirected graph

In figure 33, N=[1,2,3,4,5,6] and E=[(1,4), (2,4), (3,4), (4,5), (5,6)]

A path P is a sequence of consecutive edges in a graph and the length of the path is the number of edges traversed by P, for example, 1, 4 and 3 are in the same path.

The degree of a node is the number of edges at that node, for example the degree of the node 5 equals 2.

A directed (oriented) graph (or digraph) DG(N,E) is a set of nodes N and a set E of ordered pairs e=(n1,n2) of distinct elements of N called edges .

Two nodes n1 and n2 are adjacent in DG if (n1,n2) ∈ E or (n2,n1) ∈ E .

IV.2.2.2 Applications

Graph structures have been used to solve many problems by matching the components of a graph with certain elements in real life. For example, computers connected to each other through a network can be represented as a graph, where the computers are nodes and the connections between them are edges. Such representations ease the process of analyzing such a network. For instance, a computer network can be analyzed to find the fastest routing path between two machines in a network.

Another example: representing a social network as a graph helps identifying key actors in society.

In linguistics, many researchers have used graph structures to tackle problems in natural language processing. For example, UNL (Universal Networking Language) (Uchida and Zhu 2003) is an artificial language that can be used as a pivot graph representation in interlingua-based applications. In the UNL approach, information represented by natural language is encoded, sentence by sentence, as hypergraphs composed of directed relations between nodes or hypernodes (containing Universal Words and semantic attributes). For example, the sentence “the sky was blue?!” can be represented by the UNL graph shown in figure 34.

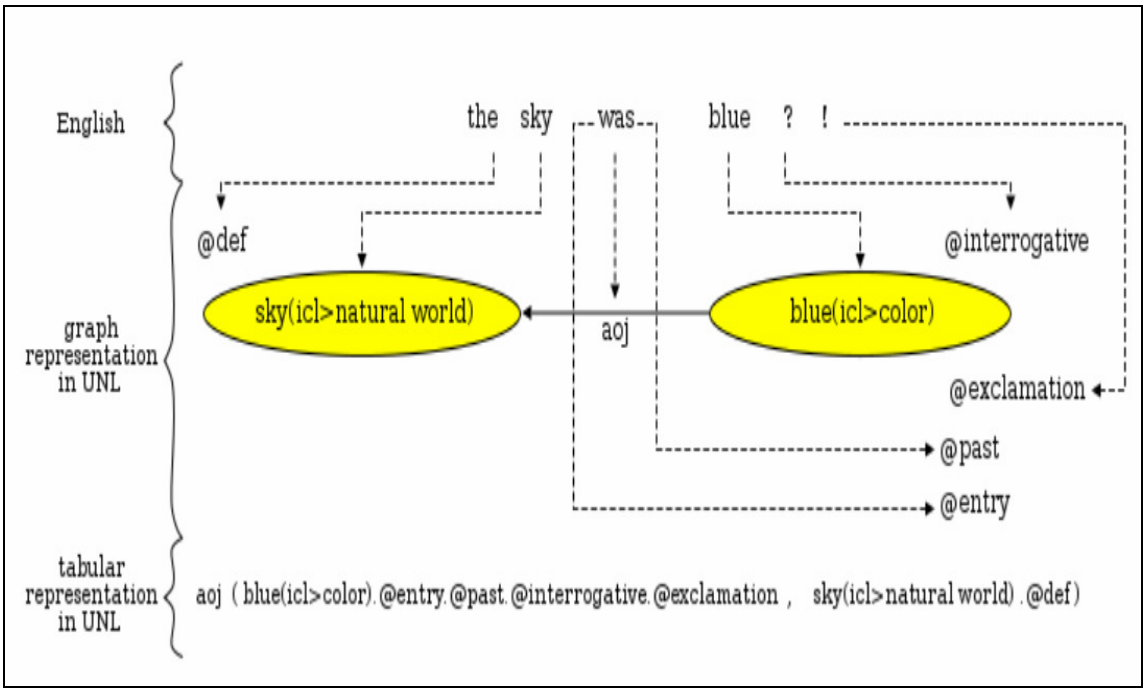


Fig. 34: UNL example

IV.2.3 **Multilingual Preterminological Graphs**

IV.2.3.1 Concept

As a multilingual preterminology is a concept-driven lexical resource with the possibility of including unconfirmed content and relations, the proposed multilingual graphs will feature both confirmed and unconfirmed.

The graph will have a large number of lexical units from different languages, and relations between pairs of lexical units. These relations have been introduced above.

IV.2.3.2 Formal Definition

a. Multilingual Preterminological Graph

An MPG is a graph contains both directed and undirected arcs. Its nodes bear quadruples of the form <preterm, language code, source, counter>. Its undirected arcs bear triples of the form <symmetrical relation, relation type, weight>. Its directed arcs bear triples of the form <asymmetrical relation, relation type, weight>.

The symmetrical relations are monolingual synonymy (SR), interlingual synonymy (translational synonymy, SR) and the "R" relation which is an undefined relation.

The asymmetrical relations are, hyponymy (HR), hyperonym (HPR), and "AR" which is an undefined asymmetrical relations between 2 preterms. [1]

This definition is based on the general definition of a digraph at the following references (Even 1979) (Loerch 2000).

[1] Only symmetrical relations will be tackled in this thesis, due to the nature of the resources. Thus, the main focus will be on TR and SR. while “R” will represent the remaining undefined relations.

MPG of domain *X* contains possible multilingual preterms related to that domain, connected to each other with relations.

b. *Preterminological Sphere*

A preterminological sphere of a particular language L is a subgraph of MPG, where SN is the set of nodes that represent preterms in language L, and SE is the set of all arcs relating nodes of SN.

Hence, an MPG represents several monolingual preterminological spheres, all relative to the same domain.

IV.2.4 **Nodes**

IV.2.4.1 Definition

a. *Nodes*

In an MPG, a node of *N* consists of *p, l, s, occ,* where *p* is the string of the preterm, *l* is the language, *s* is the code of the first source of the preterm, and *occ* is its number of occurrences. Note that *l, s,* and *occ* can be undefined. In that case they are denoted by "–".

Example:

```
N=
      {
            [silk road, en, log, 194],
            [Great Wall of China, en, wikipedia, 5],
            [الصين, ar, contributorx,6]
      }
```

In the above example, we have three nodes; two are English and one is Arabic, and the three come from different sources.

Note that the English terms and the Arabic term belong to the same *N,* thus, to the same MPG.

b. *Preterms as Nodes*

A preterm is a lexical unit (a string of characters). When it is represented as a node, the node should be identified uniquely by the lexical unit and language. Hence, if a new preterm has to be added to the MPG, we should check if the same preterm exists with the same language. MPG does not allow redundancy; if the preterm already exists in the same language, we should increase its count (occ).

IV.2.4.2 Sample Preterm Representation

As an example, figure 35 features an MPG of for nodes,

```
N=
      {
            [Java, en, –,1],
            [object-oriented programming, en, –,1],
            [برمجة كائنية التوجه, ar, –,1],
            [برمجة غرضية المنحى , ar, –,1]
      }
```

The next chapter explains the relations (edges) between such preterms.

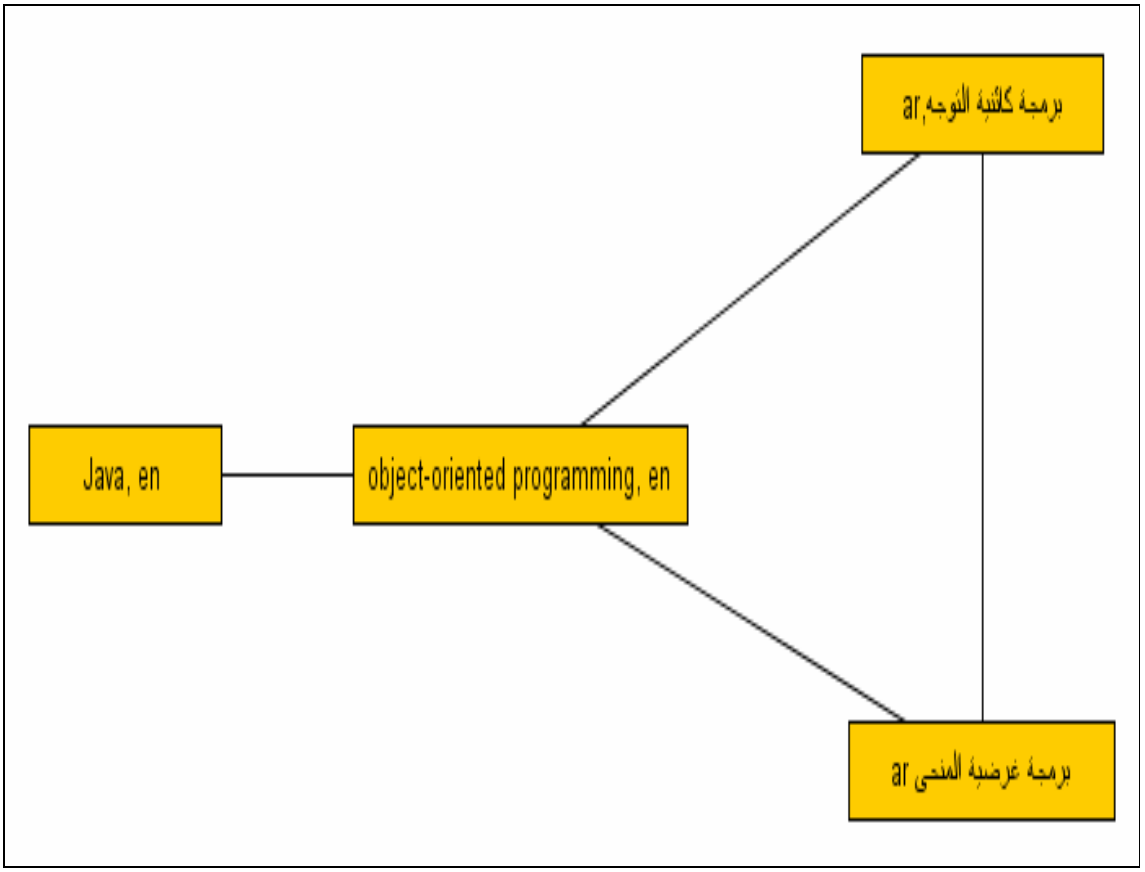


Fig. 35: Preterminological nodes

IV.2.5 **Edges**

IV.2.5.1 Definition

An edge $e=\{n_1, n_2\}$ is an ordered labeled pair of nodes adjacent in an MPG. An edge represents a relation between two preterms represented by their nodes. An edge e will bear a vector [rw, tw, sw], where the value of any component is a real number or undefined (denoted by "-". Only one of the 2 components tw, sw can be defined on the same arc.

The nature of relations varies. However, edges are weighted with several weights (described in the next section) to indicate the possible nature of this relation.

IV.2.5.2 Example

In figure 35, the edges E in the example MPG is as follows:

```
E=
      {
            ([Java, en], [object-oriented programming]),
            ([object-oriented programming, en], [برمجة كائنية التوجه,ar])
            ([برمجة كائنية التوجه,ar], [برمجة غرضية المنحى , ar])
            ([object-oriented programming, en],[برمجة غرضية المنحى , ar])
      }
```

IV.2.6 **Relations and Weights**

IV.2.6.1 Relational Weight

a. Concept

As mentioned earlier, preterminology tolerates unconfirmed lexical content and relations. Although it is difficult to build validated and confirmed coordinating relations, many resources indicate explicitly or implicitly that there is a relation between two preterms.

We will associate a "relational weight" to a preterminological relation as a default relation between two preterms, to specify its strength. This weight can and will be used for further calculations.

b. Definition

Relational Weight. *rw*: For an edge *e={[p1,l1,s1,occ1], [p2,l2,s2,occ2]}, rw(e)* is a positive real number that indicates that there is a relation between the preterm p1 and the preterm p2. The nature of the relation is not yet implied by the value *rw*.

IV.2.6.2 Translational Weight

a. Concept

We will also associate a weight to a preterminological translational relation, which will express the reliability (or confidence level) of the stored translation (in the domain at hand).

b. Definition

Translation Weight *tw*: For an edge *={[p1,l1,s1,occ1], [p2,l2,s2,occ2]}, tw(e)* is a real positive number indicating that p1 in language l1 is a translation of p2 in language l2, and vice versa.

IV.2.6.3 Synonymy Weight

a. Concept

One of the most important relations between terms is synonymy (Lafourcade and Prince 2001), and as it can be realistically deduced from other relations, and it is useful in variety of applications, it is suggested to have a weight to measure the possibility that two terms from the same language represent the same object.

b. Definition

Synonym Weight *sw*: For an edge *={[p1,l1,s1,occ1], [p2,l2,s2,occ2]}, sw(e)* is a real positive number indicating that p1 and *p2* are synonyms.

IV.2.7 **Sample Graph**

Figure 36 shows a sample MPG on the domain of the history of the Silk Road with 11 nodes (preterms) and 13 edges, it has 2 nodes in Arabic, 7 in English and 2 in French. Note the rw and tw between the nodes.

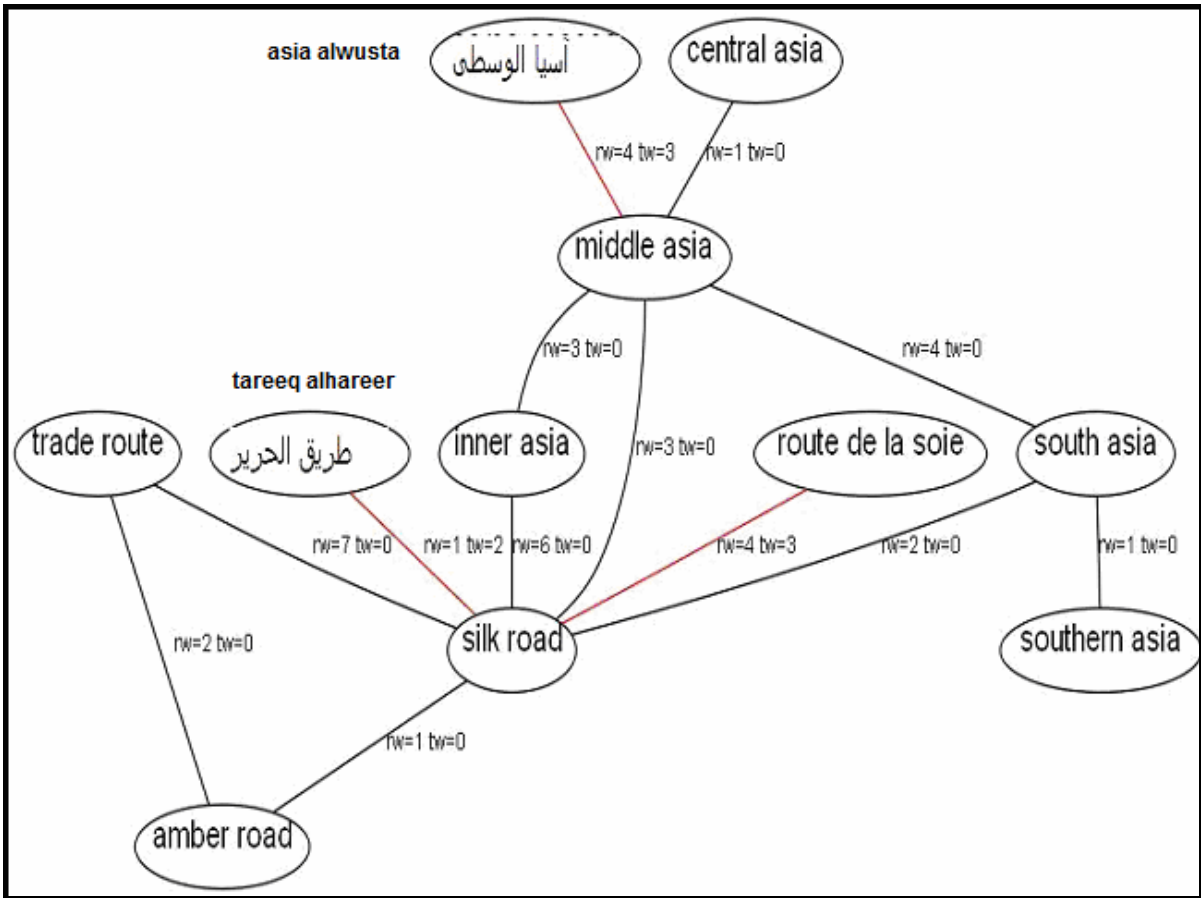


Fig. 36: Sample MPG

IV.3 Operations on the MPG

After showing the main components and structure of MPGs, this section presents the basic operations and algorithms on the Multilingual Preterminological Graphs for the purpose of representing multilingual preterminology.

The graph is implemented as a list of nodes N and a list of edges E. Each node has an ID "id" that uniquely identifies the node (Chapter 7 will give a detailed description of the implementation of MPGs and its data structures).

IV.3.1 Add Node

IV.3.1.1 Concept

Adding a node to an MPG is the process of introducing a new preterm in a particular language. The arguments of this function are the preterm (as a string), its language code, and its source list. If it is already in the MPG, its number of occurrences will increase. If not, a new node will be added to the MPG.

IV.3.1.2 Algorithm

```
addnode(preterm1, language1,source1);{
      if preterm1 in language1 exists already then {
            retrieve id;
            N[id].occ++;
```

```
      }else{
            N[number of current nodes+1].p=preterm1;
            N[number of current nodes+1].l=language1;
            N[number of current nodes+1].s=source1;
            N[number of current nodes+1].occ=1;
            N[number of current nodes+1].id= number of current nodes+1;
      }
}
```

IV.3.2 **Delete Node**

IV.3.2.1 Concept

Deleting a node is the process of removing a preterm from the graph, and any arc that is adjacent to that node.

IV.3.2.2 Algorithm

```
deletenode(preterm1, language1){
      retrieve N where N.p=preterm1 and N.l=language1;
      ID= N.id;
      for each edge e (n1, n2){
            if n1 = N[ID] then{
                  remove e from E;}
            if n2=N[ID] then{
                  remove e from E;}
            remove N[ID] from N;
      }
}
```

IV.3.3 **Add Edge**

IV.3.3.1 Concept

This algorithm constructs a new edge between two nodes, n1 and n2, given the two nodes, a type of weight t, and a weight w.

IV.3.3.2 Algorithm

```
addedge(n1, n2, t, w){
//t is the type of the weight (1 for rw, 2 for tw, 3 for sw)
      if there is no edge between n1 and n2 then {
            add a new e(n1,n2) to E;
            if t=1 then e(n1,n2).rw=w;
            if t=2 then e(n1,n2).tw=w;
            if t=3 then e(n1,n2).sw=w;
      }else{
            if t=1 then e(n1,n2).rw= e(n1,n2).rw +w;
            if t=2 then e(n1,n2).tw= e(n1,n2).tw +w;
            if t=3 then e(n1,n2).sw= e(n1,n2).sw +w;
      }
}
```

Conclusion

A terminology represents the conceptual sphere of a domain. As concepts have relations between them, it is important to capture such relations to build a preterminology. A fully hierarchical onomasiological structure is very difficult to maintain and build. On the other hand a lexical semasiological structure is much easier to construct and maintain. However, it does not preserve the conceptual structure, and it renders preterminology as a list of lexical units.

This chapter suggested the utilization of a graph structure to represent a preterminology, where preterminological relations will be represented as graph edges. If onomasiological relations are

presented and confirmed, they will be shown on the graph. If they are not confirmed, they still can be shown with a weight indication.

Multilingual Preterminological Graphs (MPGs) have been defined and described. An MPG is a container of multilingual preterminology of a domain. MPGs are made of nodes and weighted edges, where a node represents preterms and an edge represents a relation between two terms. The weight indicates the nature of this relation and its reliability. The basic operations on the graphs have been shown. The next chapter describes in detail how these operations will be used to construct an MPG representing a given preterminology.

Chapter V MPG Construction and Preterminology Elicitation

Process Overview and Introduction

The process of constructing a multilingual preterminological graph for a domain involves initializing the graph with preterms extracted from conventional and unconventional resources, multilingualizing the graph using lexical translation techniques, and then analyzing it to exploit the graph structure to find more interesting lexical information. Figure 37 shows the whole process.

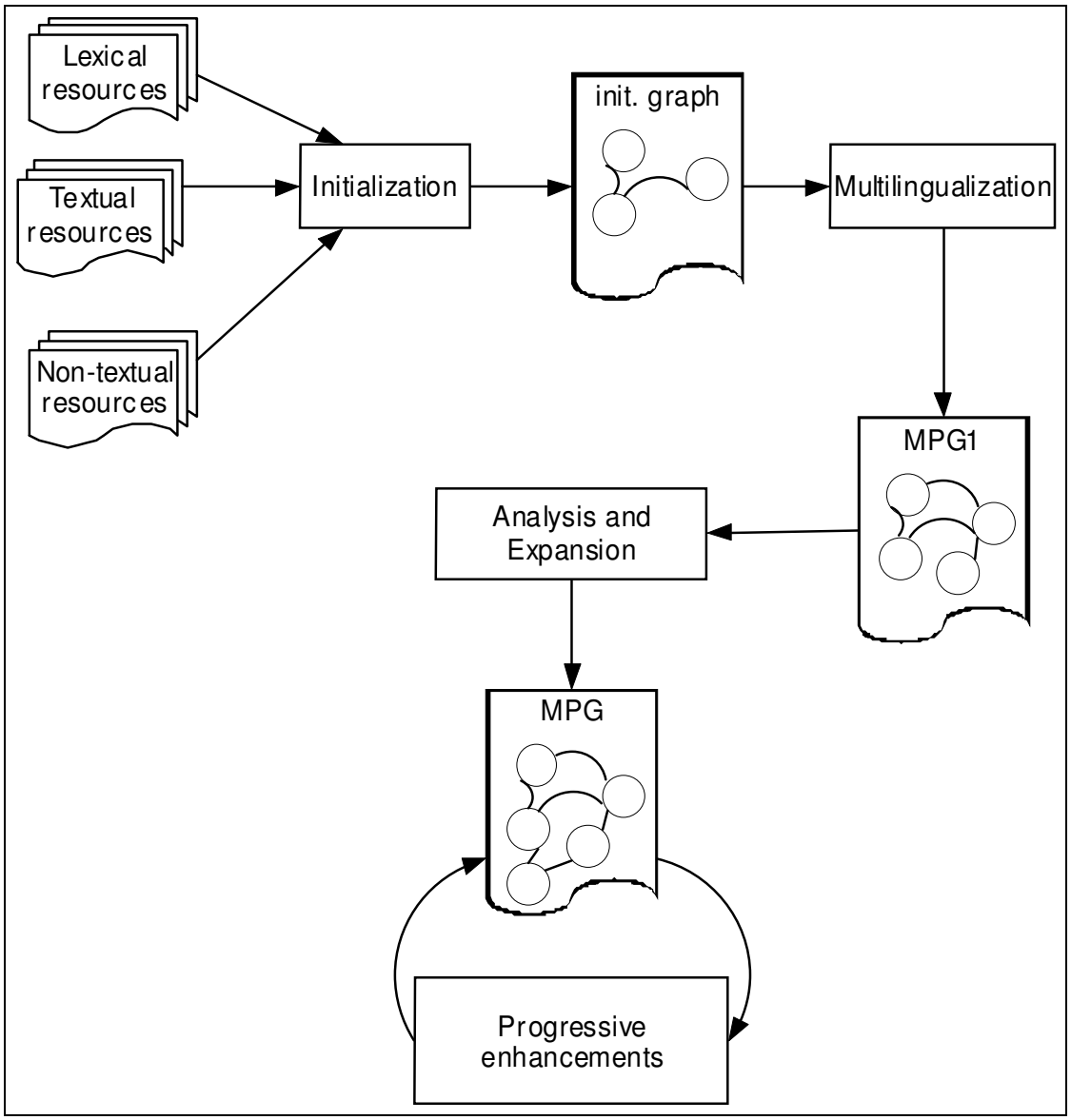


Fig. 37: MPG construction

After the graph has been constructed, it can be enhanced progressively by human contributions; this chapter will describe the details of this process.

V.1 Elicitation Methods and Approaches in Relation to a Community

V.1.1 Overview and Introduction to Terminological Elicitation

Terminology Elicitation is the process of drawing out latent and absent terminology through the following:

- compiling what is available in lexical and non lexical resources;
- analyzing what is available;
- exploiting the human factor, in the process, as a source of knowledge and "terminological wisdom".

This section explains the elicitation approaches using human knowledge for preterminology and its multilingual graphs. Passive and Active approaches will be defined and discussed.

V.1.2 Passive Approaches

In the passive approaches, contributors' actions do not involve any intention to provide contribution to lexical data; however their actions are processed automatically to find lexical data.

Passive approaches are not only concerned with compiling and extracting terminology from digital resources, but also with the analysis of the trends in using terminology for a specific domain.

V.1.2.1 Lexical Resources Compilation

Lexical resources are essential for preterminology as they provide the initial step in developing an MPG and multilingualizing it. Even though they often contain general purpose material, lexical resources are easy to utilize and have good coverage. Monolingual lexicons, multilingual dictionaries, terminological databases, glossaries… can enrich the preterminology and its corresponding graph.

V.1.2.2 Textual Approaches

Automatic analysis of textual material to find preterminology is considered as passive contribution. The provided preterms are included in the text, but are not provided as proper Contribution Units (*CUs)*.

This means the contributor who published the text in digital format did not intend necessarily to contribute to a lexical or terminological resource. However, his text might be rich in terminology.

Textual resources can help the development of preterminology as follows:

- extracting monolingual terms from domain dedicated textual material using standard term recognition techniques and tools.
- extracting (bi/multi)lingual terms from parallel or comparable texts related to the considered domain.
- utilizing structured textual resources for preterminology, like monolingual or multilingual encyclopedic texts.

V.1.2.3 Non-Textual Approaches

Not all digital information is available in textual human readable format, but it might be available in a way only machines can understand. Such resources are built by computers to be analyzed by computers, and can contain interesting information for preterminology.

Server access log files register the interactions between the human user and an online server including what the user visited, when, and from where.

For a website, the access logs can be further analyzed to find what interests the visitors, what the main characteristics of the visitors are, and how they arrived at the website. Software like AWStats and Google Analytics have been developed to analyze such log files and present the website administrators with neat reports.

Recently, many visitors depend on search engines to find what they need. Usually they search for a content related to a certain concept, they formulate it using a keyword (term), and they ask the search engine to provide them with relevant websites. If they are not satisfied, they often send a second request including another keyword for the same concept or for a similar one.

Access log files record all the keywords. It would benefit a preterminological resource to exploit these keywords as they carry the exact representation of concepts from Internet users who might be domain experts.

V.1.3 **Active Approaches**

Contributors in these approaches are actively involved in providing lexical data as a *CUs*, either directly or indirectly. This approach is often called participative, collaborative or volunteering.

Written languages (textual material) change more slowly than corresponding spoken languages. At the language level, this causes a "diaglossia" which is the gap between the language used by the community and the language available in textual media, At the lexical level, this causes a terminological gap.

As preterminology tries to bridge that gap, depending on digital material is not enough. Hence, the community of the domain should be involved either directly or indirectly in the development of the corresponding preterminology, by contributing to its MPG.

V.1.3.1 Direct Approaches

In this case, the volunteer is working as an amateur terminologist. It is more like the approach of Wiktionary, where the contributor formulates the contribution unit CU directly as a dictionary entry.

Although this approach is ideal, it is very difficult to motivate volunteers to do a non-spontaneous contribution, especially to a lexical resource such as a preterminology, because a lexical resource, by its nature, is not easy to construct, it requires special knowledge, and the contribution is poorly acknowledged to the volunteer. Furthermore, the structure of the contribution unit does not allow the user to add his opinion or personal flavor, unlike other contribution units in Wikipedia for example. All that makes it difficult to get direct non-spontaneous contributions.

V.1.3.2 Indirect Approaches

Much of human knowledge gathering systems use these approaches as they are more appealing to crowd of non professionals. The approach depends on gathering the CU from the volunteer through a non-contributive activity, for example, through a game. As the process of providing lexical data is difficult for average people, we are seeking an indirect approach by attaching the

process of providing multilingual terminological data to a neither linguistic, nor terminological activity, but to a domain-related activity, which will encourage more domain experts to participate.

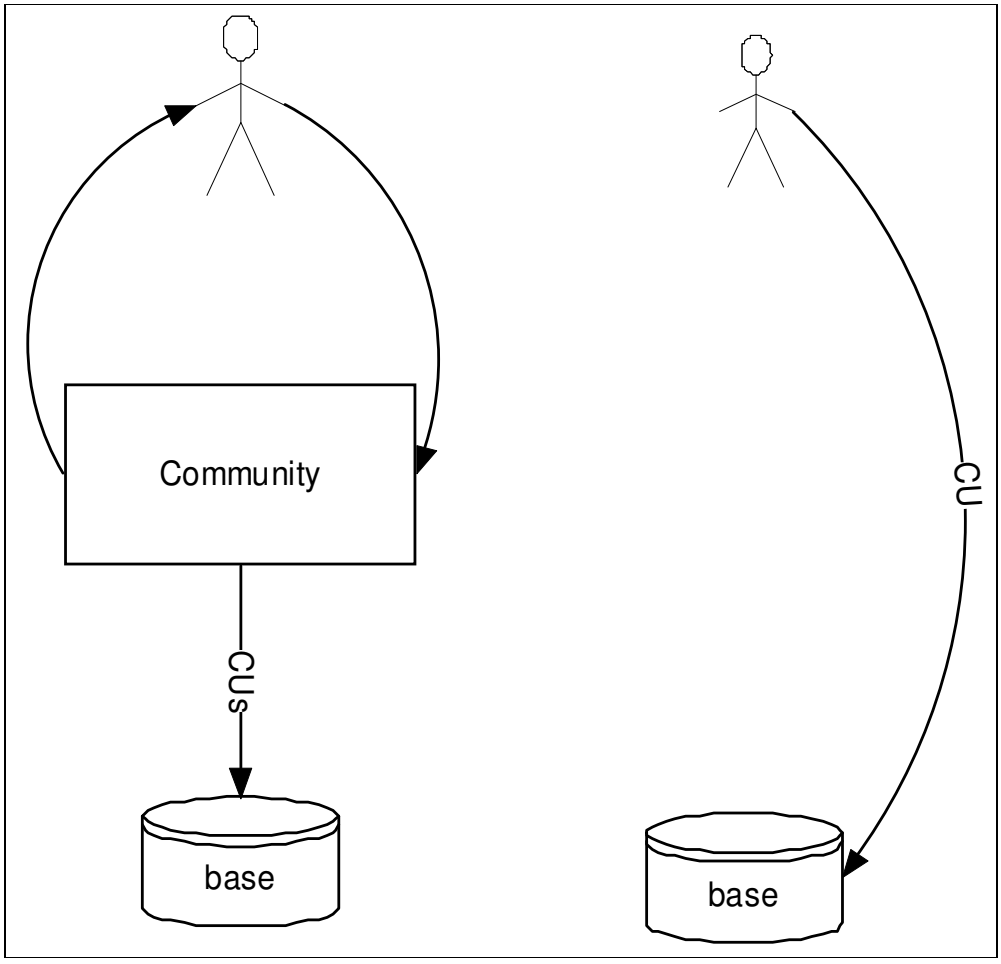


Fig. 38: Direct and indirect contribution

Figure 38 shows the main difference between direct and indirect contribution. In the absence of a community activity, the direct contribution will be a linguistic volunteer approach that does not have a real motivation.

V.1.3.3 Serious Games

Serious games (or "game with a purpose", a GWAP) constitutes an indirect spontaneous approach for contribution to big knowledge bases. A GWAP depends on massive contributions from volunteers through an online game. There is a variety of games in different domains now, like Open Mind Common Sense (Singh, Lin et al. 2002), ESP games, Learner (Chklovski and Gil 2005-b), FACTory games (FACTory 2010), Mindpixel (Mindpixel 2009), and JeuxDeMots (Joubert and Lafourcade 2008). The games motivate players to contribute through ranking and counting player's scores. An important distinction between such games is the settings and the approach in validating and counting the scores.

a. *Approach A: Reward Based on Volume of Contribution*

Here, there is no actual validation of the contributed data; the scoring is based on the amount of contribution. An example is Open Mind Word Expert (Mihalcea and Chklovski 2003), which is aiming at building large semantically annotated data from volunteers.

b. *Approach B: Simultaneous Agreement on One Contribution*

In Google Image Labeler and ESP Games, the scoring and validation are based on the agreements between two players on a contribution set, at the same time.

c. *Approach C: Previously known answers*

Similar to reCAPTCHA, where we have a word that is known and another that is not, Learner uses this approach to collect paraphrases, see figure 39.

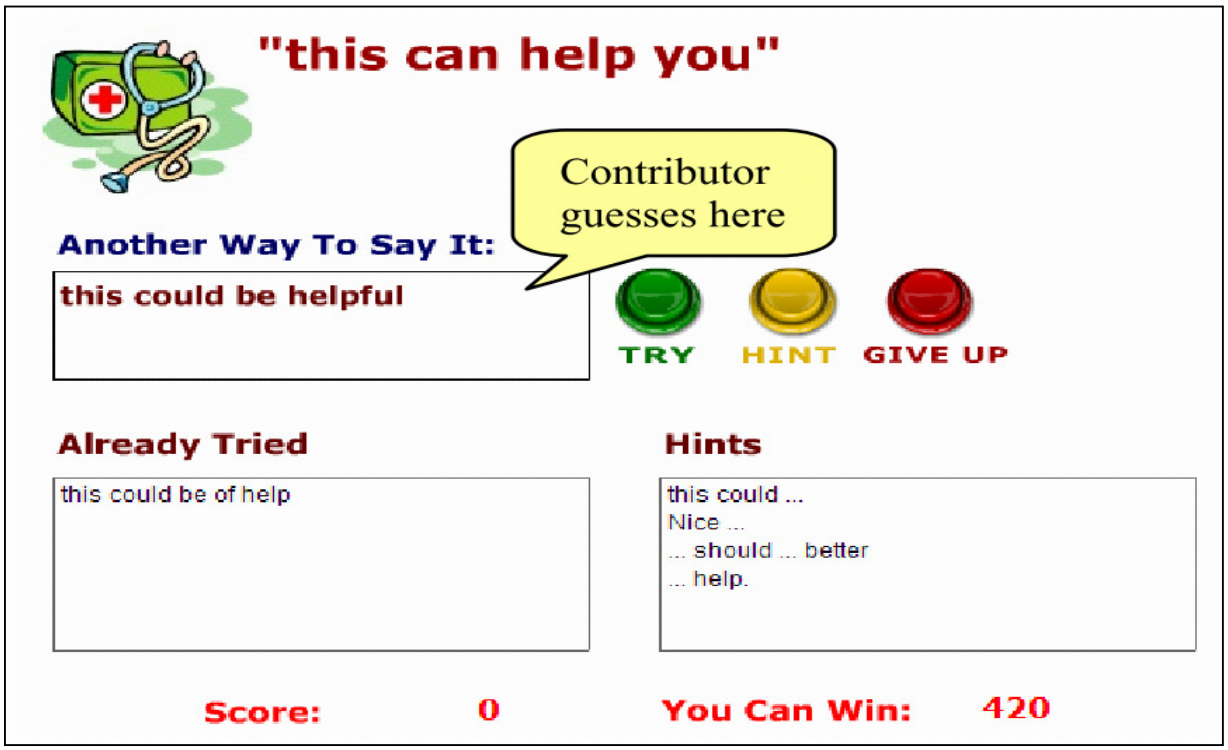


Fig. 39: Learner

The score and validation of the contribution is based on the set of questions with pre-known answers, see figure 10. This approach does not need 2 players with the same interest at the same time, which reduces the bottleneck problem.

V.1.3.4 Applications to Preterminology

Applications for preterminology (such as GWAP) need initial preterminological data, hence there is a need to use passive approaches for initialization. Applications such as GWAPS and domain-related applications can achieve “progressive enhancements”, as the bottom of figure 37 shows.

V.2 MPG Development

The previous chapter suggests that a multilingual preterminology should be developed and maintained using an MPG, which should be constructed and enlarged using various conventional and unconventional resources in order to collect latent and absent terminology.

The previous section showed that active contributions cannot be obtained without initialization using passive resources. This section explains the method for initializing and constructing an MPG, according to the steps shown in figure 37:

- initialization (Passive Contribution);
- multilingualization (Passive Contribution);
- expansion (Passive Contribution);
- progressive enhancement (Active Contribution).

V.2.1 Initializing Preterminology

Building an MPG is not only concerned with compiling data, but also with structuring it and establishing relations between the preterms. To initialize the graph, one compiles preterms from digital resources without making confirmed translational or relational correspondences between the nodes. The initialization process relies on two passive approaches, textual and non-textual.

V.2.1.1 Terminology Extraction

Although terminology extraction is limited by the available corpus and the computational approach, this method is used for seeding many lexical resources with initial material extracted from relevant text.

A typical term extractor works as follows: it receives text, analyses it and produces possible terms. Some of these term extractors are available as online services (LCL 2010) (Yahoo 2008).

The MPG initializer receives the domain-related text and sends it to a term extraction service, to receive a list of candidate terms. Each term on this list will be added to the initial graph as a new node, as figure 40 shows. At this point, no edges have been developed yet.

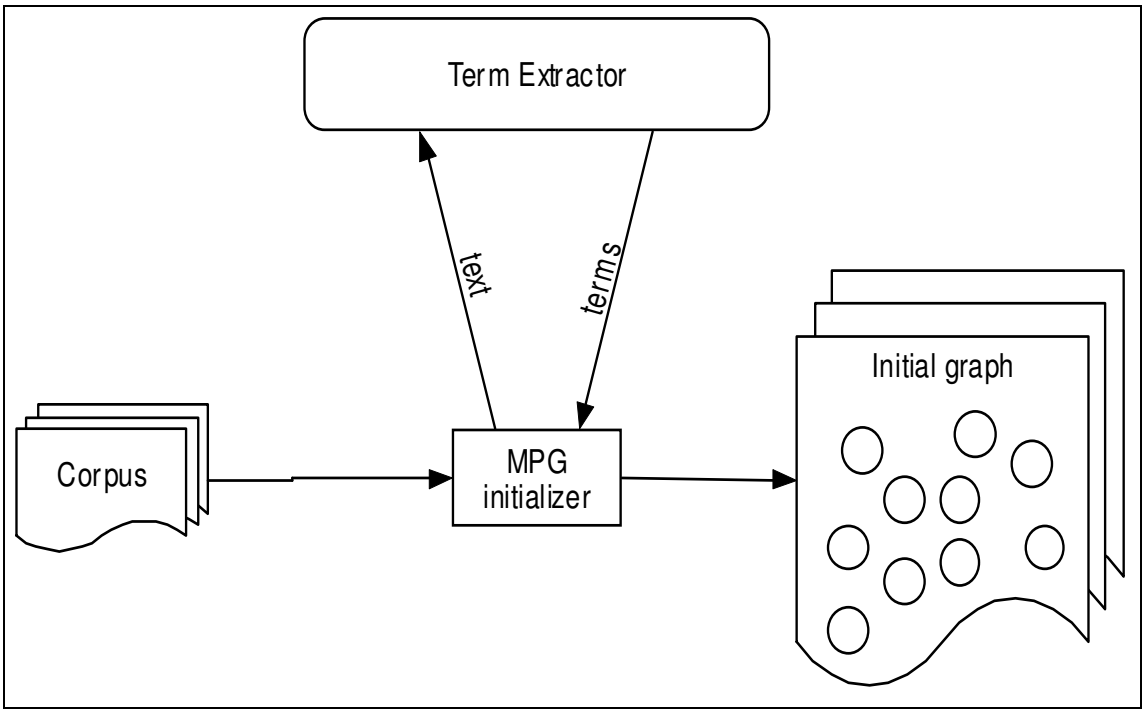


Fig. 40: Calling term extractor

V.2.1.2 Implicit Human Contribution: Analyzing Access Log Files

Access log files constitute a very useful resource that is related to a specific domain, as they register the interactions between a domain-related online community on one side and users (who might include domain experts) on the other side.

A server access log file keeps track of the requests that have been made to it, along with other information like request time, IP address, referred page. Such a resource is not considered textual because it is not comprehensible by humans, but it is easy to analyze by a computer.

We analyze two kinds of requests that can provide us with information to enrich the MPG:

- Requests made to a local search engine devoted to a website and its documents and local pages.
- Requests with reference from a web-based search engine (like Google, Bing, Yahoo!, Altavista…).

From these requests, we can obtain the search terms that visitors have used to access the website. Moreover, we can understand the way users interpret a concept into lexical units. Finding a pattern in their requests may lead to find a relation containing the preterms used in requests.

For example, if we find that five different users have sent two consecutive search requests, containing respectively the terms t1 and t2, then there is a possibility that t1 and t2 have a lexical relation.

The preterms extracted from access log files (non-textual resource) can be merged with the initial graph extracted from the text.

The graph constructor analyzes the requests to expand the initial graph by creating edges between preterms found within the same search session.

The relation weight between x and y, *rw(x,y)*, is set to the number of sessions containing x and y within the log file.

For example, *rw(x,y)* = 10 means that 10 persons thought about x and y within the same search session.

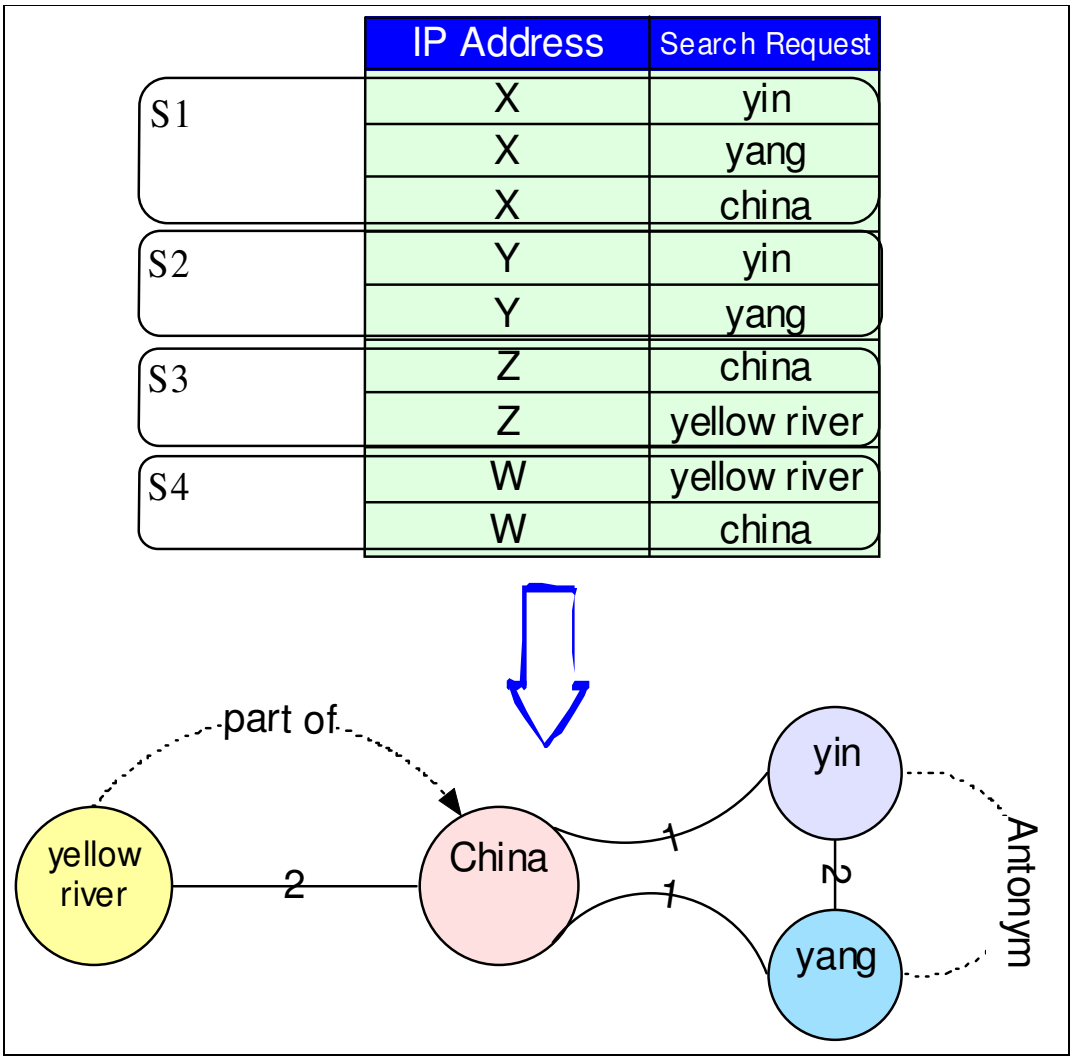


Fig. 41: Example of constructing a MPG from an access log file

Figure 41 shows an example of a log file and the produced graph. The proposed method does not discover the kind of relation between the terms. However, it discovers that there is a relation, for example, three users requested results for "yang" followed by "yin" within the same session. Hence, an edge with weight 2 was constructed between them.

Search requests are rich of preterminology, as they contain terms that are in use by the community. Extracting these terms and structuring them will serve the purpose of covering the preterminological sphere of a domain.

V.2.2 **Graph Multilingualization**

We assume that the initial graph has been constructed. This section describes the process of multilingualizing the graph by translating each preterm of the initial graph using online resources.

V.2.2.1 Node Multilingualization

a. Algorithm

```
For each node n in the original MPG{
      For each target language tg{
            For each available source u{
                  ntg=Translate n into tg using source u;
                  if ntg does not exist in thecurrent MPG then{
                        addnode(ntg);
                        addedge(n, ntg, tw, 1);
                  }
                  if ntg does exist in current MPG then
                        {add edge(n, ntg, tw, 1);}
            }
      }
}
```

b. Basic Concept

As figure 42 shows, every preterm in the initial graph will be sent to various online resources to be translated into different languages. Each translation will be added to the graph as a new node, with an edge between the preterm and its translation. If the same translation is provided by different resources, the translation weight *tw* will be increased, meaning that the translation has gained more confidence.

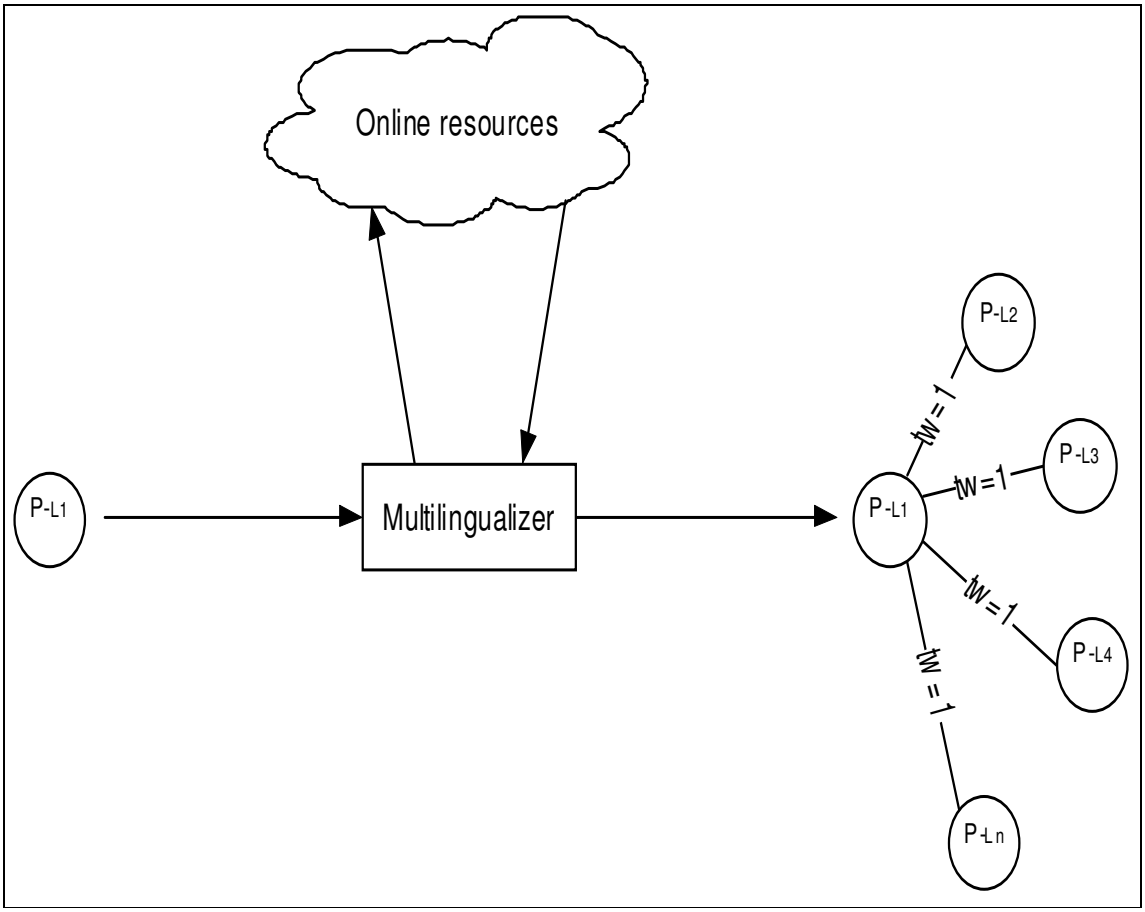


Fig. 42: Preterminology multilingualization

V.2.2.2 Automatic Lexical Translation Using Online Resources

a. Wikipedia

Wikipedia (Wikipedia-A 2008) is a rich source of preterminology, as explained in chapter 3. Wikipedia is being exploited as follows:

- We select a root set of preterms;
- For each root term, we construct its corresponding Wikipedia article's URL, and then we use the language roll of the article, as figure 43 shows (Daoud, Kageura et al. 2009);
- We extract key terms from the article for further expansion.
- For example:
  - Ex. Root term: Cuneiform_script
  - En URL: http://en.wikipedia.org/wiki/Cuneiform_script
  - Fr URL: http://fr.wikipedia.org/wiki/Cunéiforme
  - Ar URL: http://ar.wikipedia.org/wiki/كتابة_مسمارية

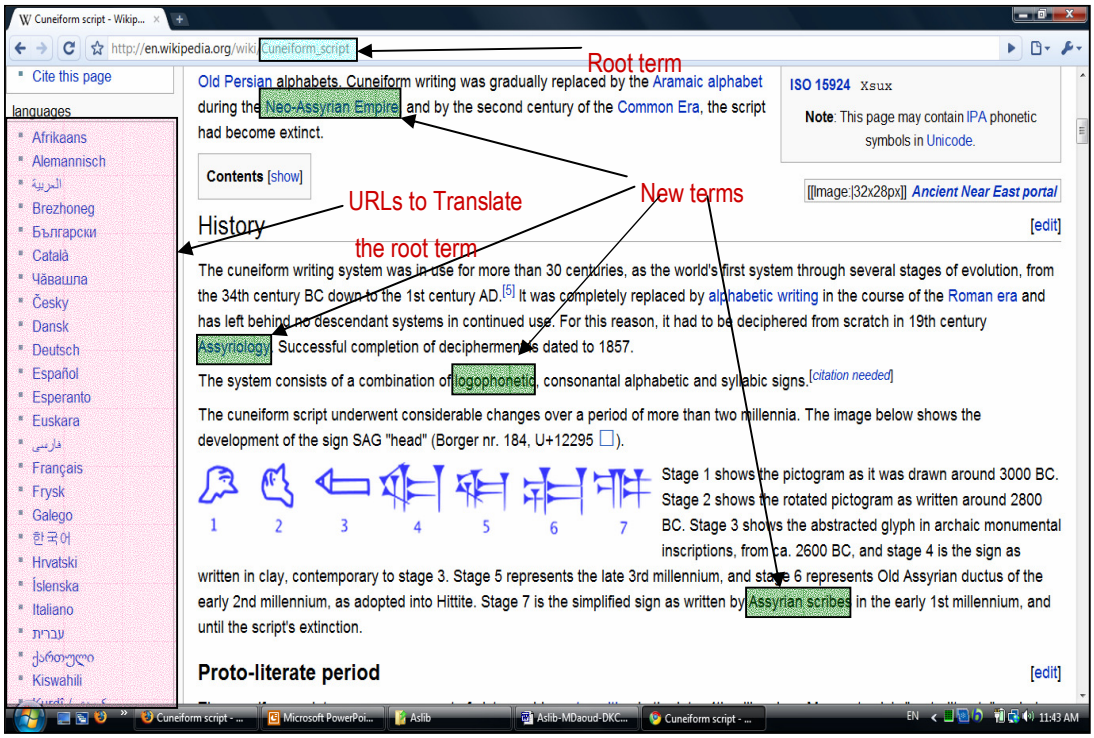


Fig. 43: Extracting multilingual preterminology from Wikipedia

b. Online MT

Machine translation systems can be used as general-purpose MRDs (Vo-Trung 2004). One of the main advantages of MT systems is the good coverage even for multiword terms. The agreement of some MT systems with other resources on the translation of one term enhance the confidence of the translation. Another positive point is that the results of MT provide a first draft to be post-edited later.

Example of MT systems handling many languages:

- Google Translate (Google 2008) (50 languages);
- Systran (Systran 2009) (14 languages);

- Babylon (Babylon 2009) (26 languages).

Figure 44 shows an example of translating the term "great wall of China" into Arabic using these three MT systems.

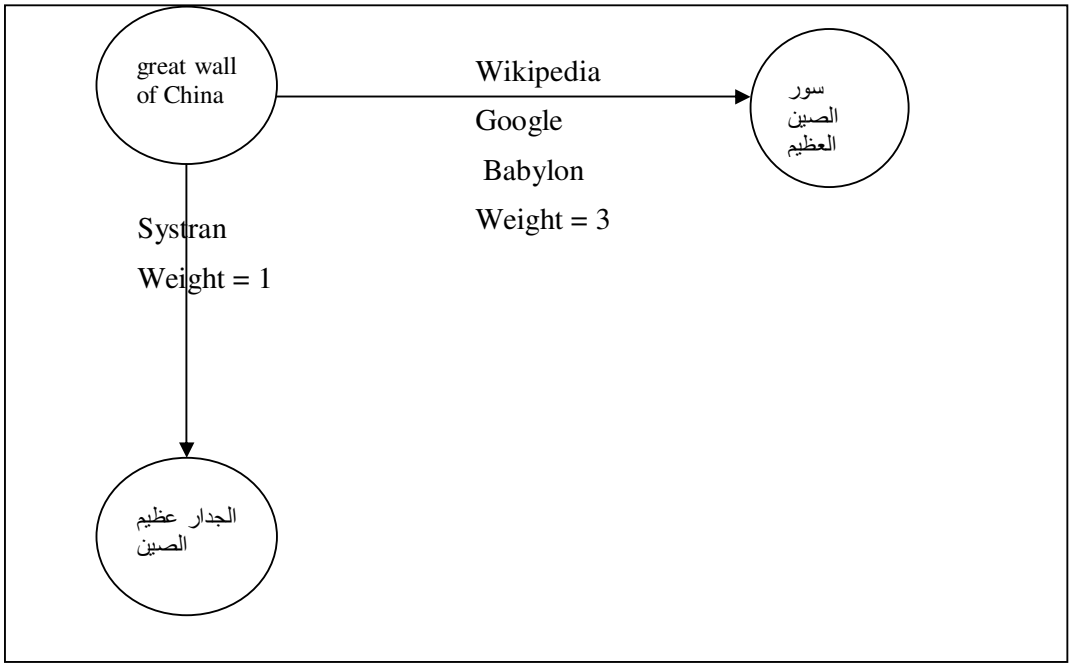


Fig. 44: Example on preterm multilingualization

c. *Online Dictionaries*

Online dictionaries and term bases are also a valid and important source of lexical translation and should be used for preterminology multilingualization.

V.3 MPG Expansion

In the first two steps, the graph compiles a large amount of multilingual data with unconfirmed relations. The structure of the graph itself and these relations can be used to "elicit" hidden correspondences. This section explains the computations on the graph to find hidden possible relations between the preterms already compiled in the first two steps.

V.3.1 Basic Principle

Relational Weights *rw* on the graph are calculated based on the following hypothesis:

Hypothesis 1: there is a possibility that two preterms available in the same search session are related.

This relation so far is not known, however we can utilize it, and the translational weights on the graph to confirm the synonym relations, based on the second hypothesis:

Hypothesis 2: if two preterms share the same translations, the possibility that they are synonyms increases.

Figure 45 shows an example, where the preterms X1 and X2 share three translations in French, Italian, and Arabic. We suggest that there is a possibility that the two preterms are synonyms.

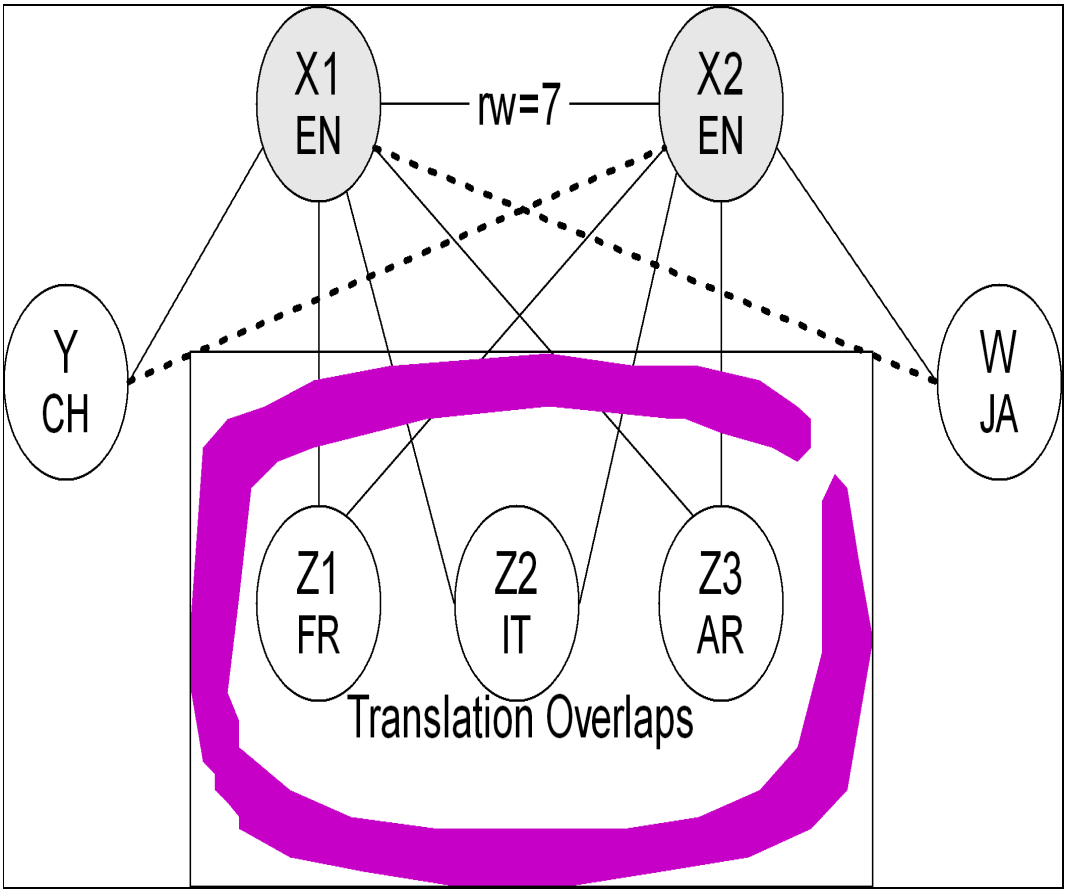


Fig. 45: MPG expansion

V.3.2 **Expansion Formula**

Based on the hypotheses in the previous section, one can formulate a procedure to analyze the graph in order to calculate synonymy weights between nodes.

Based on the hypotheses we have two kinds of information that can validate the *sw*:

- *rw*: based on hypothesis 1, the possibility that two preterms are synonyms increases if they have a higher *rw*. However, the importance of *rw* decreases if the preterms have high rdegree (degree of edges at a preterm where *rw* is higher than 0), meaning that the term has many relations with many terms, and there is less significance to the relation between these two particular preterms.
- Translation overlaps: based on hypothesis 2, the possibility that two preterms are synonym increases if they have a higher translation overlap. However, the importance of the overlaps decreases if the preterms have high tdegree (degree of edges at a preterm where tw is higher than 0), meaning that the two preterms have many translation relations and they only share few translations.

Based on that, we propose the following formula for finding the *sw* between two terms:

$$synonym\,weight(X1, X2) = \frac{(rw(X1, X2))}{(min(rdegree(X1), rdegree(X2)))} + \frac{(\#\,translation\,overlaps)}{(min(tdegree(X1), tdegree(X2)))} \quad (1)$$

sw between X1 and X2 is the sum of the significance concluded from rw, and translation overlaps.

For example, in figure 45, *sw(X1, X2)=(1/1+3/4)=7/4*.

V.3.3 **Indirect Translations**

V.3.3.1 *Basic concept*

If the graph has a preterm *t1* and its synonym *t2*, then edges with high *tw* can connect *t1* and *t2*. In other words, *t1* and *t2* correspond to the same concept and it is possible that they have the same translations. We call this *indirect translation* because there is no direct edge between the term and translationally equivalent terms.

Therefore, x is an indirect translation of y, if *x* connects to *x1*, and there is a path *p* from *x1* to *y*, where *tw(x1,x)* > *k*, *k* being a confidence threshold, and where all edges of *p* have *sw* > *k1*, *k1* being a synonymy confidence threshold.

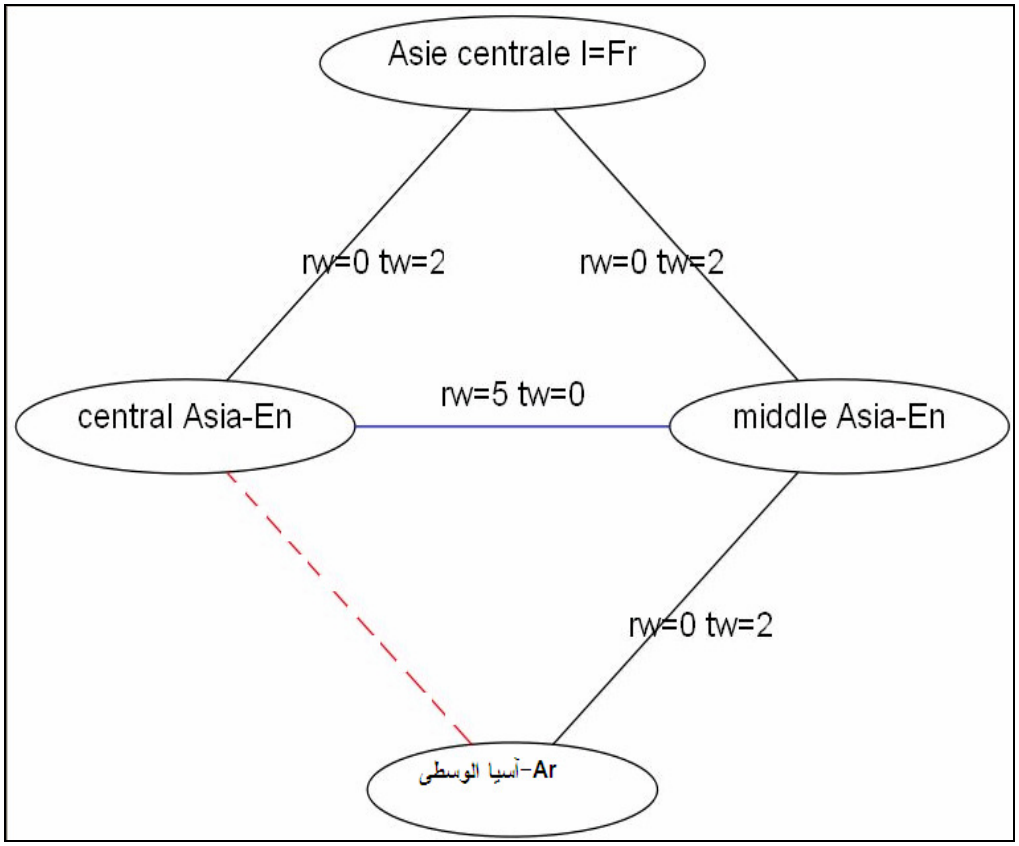


Fig. 46: SW calculation

For example, in figure 46, "آسيا الوسطى" is considered as an Arabic translation of "central Asia" if *k=1*, and *k1=1*. This is because *tw("آسيا الوسطى", middle Asia)=2*, *and sw(central Asia, middle Asia)=5/1+1=6*, based on Formula 1.

V.3.3.2 Algorithm

This algorithm adds indirect translations based on the relations between two nodes of the graph. It receives its information as preterms, and in addition to that it receives 2 confidence threshold k and k1.

```
IndirectEdgeConstruction (node1, node2, k, k1)
      If node1 and node2 have a path p between them where all the edges on p
      have sw>k1 then
            For all neighbors x1 of node1
                  If tw >k
                        addedge(node2, x1,tw,1)
            For all neighbors x2 of node2
                  If tw >k
                        addedge(node1, x2,tw,1)
```

V.4 Lexical Knowledge Extraction from MPG

There is a need to extract lexical preterminological knowledge from an MPG in order to function as a useful repository. This section explains the methods of retrieving monolingual and multilingual preterminology from an MPG in a systematic and standard way. We also define a possible data format to exchange the multilingual preterminological knowledge in order to make it useful for other applications and systems.

V.4.1 Extracting Multilingual and Monolingual Preterminology from MPG

V.4.1.1 Extracting Sub-Graphs from MPG

To be useful in applications as a proper lexical resource, a preterminology should be transferred into a common lexical format that is easy to understand, import and export, and be integrated with other systems and applications.

a. Example

Figure 47 shows an example of a sub-MPG extracted from an MPG, based on the language; the subgraph consists of all the nodes with English preterms.

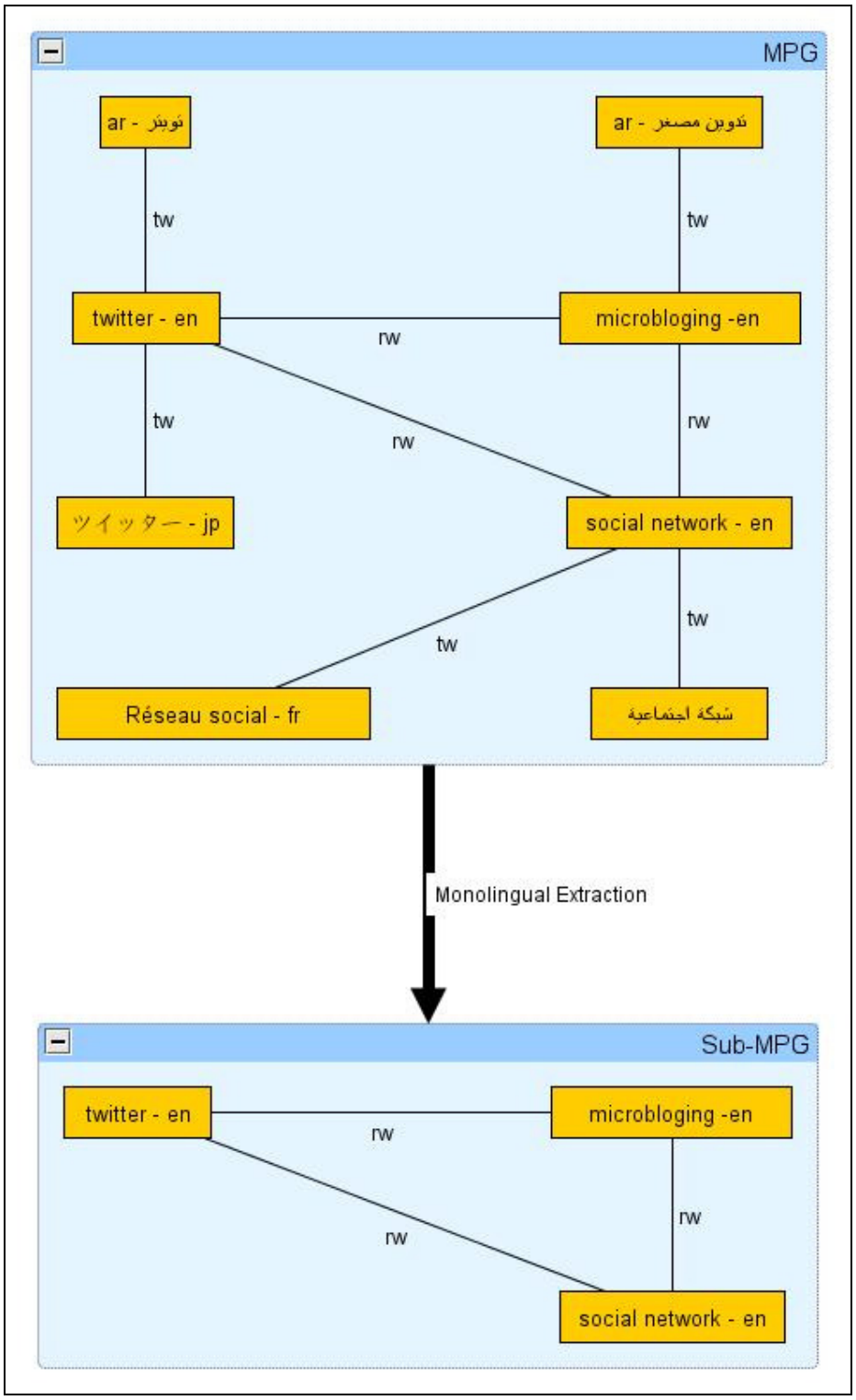


Fig. 47: Subgraph extraction

b. *Parameters*

To extract a sub-MPG one needs to specify the criteria of selecting the components of the original graph. The following parameters can be tuned to extract the desired sub-MPG, including extracting monolingual graphs:

- Selected node:
  - L: language of the nodes.
  - occ: number of occurrences of the preterms.
- Selected edges:
  - k_{SW}: to filter all edges where sw is less than k_{SW}.
  - k_{RW}: to filter all edges where sw is less than k_{RW}.
  - k_{TW}: to filter all edges where sw is less than k_{TW}.

c. General Algorithm

```
ExtractSubMPG(l,occ, k_SW, k_RW, k_TW){
      for each node n of MPG{
            if n.language=l then{
                  if n.occurences=occ then{
                        submpg.addnode(n);
                  }
            }
      }
      for each edge on MPG{
            if e is between n1 and n2 but if n1 & n2 are nodes of submpg
            then{
                  if sw> k_SW and tw> k_tW and rw> k_rW then{
                        submpg.addegde(e)
                  }
            }
      }
      return submpg
}
```

V.4.1.2 Storing MPG

Storing the graph in a digital format involves transforming it from its internal computer representation into a textual representation (in xml, text…). The following algorithm prints the content of the graph into a file:

```
writeMPG(k){
      for each node ni
            if ni is not marked as printed
                  start new record
                  print ni
                  mark ni as printed
                  currn=adjacent nodes of ni
                  for each currnj
                        if tw(ni,currnj)>k
                              print currj
                              mark currj as printed
                  end record
}
```

V.4.2 **Preterminology in XML format**

V.4.2.1 Possible Structure

An MPG can be textually represented in many ways, as any graph. This subsection shows some possibilities.

a. Minimal Text

MPG can be a list of numbered nodes followed by a list of arcs, for example:

Nodes

1: term1, lg1, occ1, loc1

2: terms2, lg2, occ2, loc2

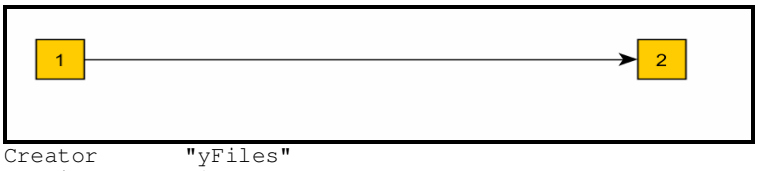
Edges

1, 2, sw12, tw12, rw12

Although that kind of format is universal, it is overwhelmingly difficult to implement automatic functions on graphs by manipulating their textual descriptions. For example, it is difficult to delete a node while keeping track of its relations.

b. GML

GML, the **G**raph **M**odelling **L**anguage (Himsolt 2010), is an ASCII-based portable file format for describing graphs. GML's key features are portability, simple syntax, extensibility and flexibility. It has been used by several applications like (yed (yWorks 2010), tulip, cytoscape…). The graph is printed as nodes with their labels followed by edges. Here is an example of a simple graph that consists of 2 nodes and 1 edge between them:




```
Creator       "yFiles"
Version       "2.7"
graph[
       hierarchic    1
       label  ""
       directed      1
       node[
              id     0
              label  "1"
              graphics      [
                     x      276.0
                     y      230.0
                     w      30.0
                     h      30.0
                     type   "rectangle"
                     fill   "#FFCC00"
                     outline       "#000000"]
              LabelGraphics[
                     text   "1"
                     fontSize      13
                     fontName      "Dialog"
                     anchor "c"]]
       node[
              id     1
              label  "2"
              graphics[
                     x      652.0
                     y      230.0
                     w      30.0
                     h      30.0
                     type   "rectangle"
                     fill   "#FFCC00"
                     outline       "#000000"]
              LabelGraphics[
                     text   "2"
                     fontSize      13
                     fontName      "Dialog"
                     anchor "c"]]
       edge[
              source 0
              target 1
              graphics[
                     fill   "#000000"
                     targetArrow   "standard"]]
]
```

However, this syntax is more suitable for viewing purposes and can not be transformed into other formats especially with the absence of robust XML→GML analyzers.

c. GraphML

GraphML (GraphML 2010) is an XML-based file format for graphs. It consists of a language core to describe the structural properties of a graph and a flexible extension mechanism to add application-specific data. Here is how a graph of two nodes and one edge, can be written in GraphML:

```
<?xml version="1.0" encoding="UTF-8"?>
<graphml>
  <graph id="G" edgedefault="undirected">
    <node id="n0"/>
    <node id="n1"/>
    <edge id="e1" source="n0" target="n1"/>
  </graph>
</graphml>
```

V. Archer (Archer 2009) used the GraphML format to represent his multilevel linguistic graphs. He argues that GraphML is supported by several APIs and graph editing systems such as Boost Graph Library and yed. We think it is suitable for MPG, because it allows defining MPG-specific data at the same time it can be transformed and processed by other applications and systems.

V.4.2.2 GraphML File Format

The following is a sample MPG in GraphML format; it represents the MPG in figure 48.

```
————XML STARTS————————————————
<?xml version=”1.0″ encoding=”UTF-8″?>
<GraphML>
        <key id=”d0″ for=”node” attr.name=”preterm” attr.type=”string”></key>
        <key id=”d1″ for=”node” attr.name=”language_code” attr.type=”string”>
        <default>eng</default></key>
        <key id=”d2″ for=”node” attr.name=”source” attr.type=”string”>
        <default>unknown</default></key>
        <key id=”d3″ for=”node” attr.name=”occ” attr.type=”string”>
        <default>1</default></key>
        <key id=”d4″ for=”edge” attr.name=”rw” attr.type=”double”>
        <default>0</default></key>
        <key id=”d5″ for=”edge” attr.name=”sw” attr.type=”double”>
        <default>0</default></key>
        <key id=”d6″ for=”edge” attr.name=”tw” attr.type=”double”>
        <default>0</default></key>

        <graph isAcyclic=”true” id=”dsr” >
                <node id=”2353″ >
                        <data key=”d0″>great wall of China</data>
                        <data key=”d1″>eng</data>
                        <data key=”d2″>dsr_log</data>
                        <data key=”d3″>11</data>
                </node>
                        <node id=”2354″ >
                        <data key=”d0″>سور الصين العظيم</data>
                        <data key=”d1″>ara</data>
                        <data key=”d2″>wikipedia</data>
                        <data key=”d3″>3</data>
                </node>
                        <node id=”2355″ >
                        <data key=”d0″>万里の長城</data>
```

```
                        <data key=”d1″>jpn</data>
                        <data key=”d2″>dsr_log</data>
                        <data key=”d3″>4</data>
                </node>
                        <node id=”2356″ >
                        <data key=”d0″>Grande Muraille de Chine</data>
                        <data key=”d1″>fra</data>
                        <data key=”d2″>wikipedia</data>
                        <data key=”d3″>3</data>
                </node>

                        <edge source=”2353″ target=”2354″>
                        <data key=”d4″>0</data>
                        <data key=”d5″>0</data>
                        <data key=”d6″>3</data>
                </edge>
                        <edge source=”2353″ target=”2355″>
                        <data key=”d4″>1</data>
                        <data key=”d5″>0</data>
                        <data key=”d6″>2</data>
                </edge>
                        <edge source=”2353″ target=”2356″>
                        <data key=”d4″>0</data>
                        <data key=”d5″>0</data>
                        <data key=”d6″>2</data>
                </edge>
        </graph>
</GraphML>
——————————-XML ENDS——————————
```

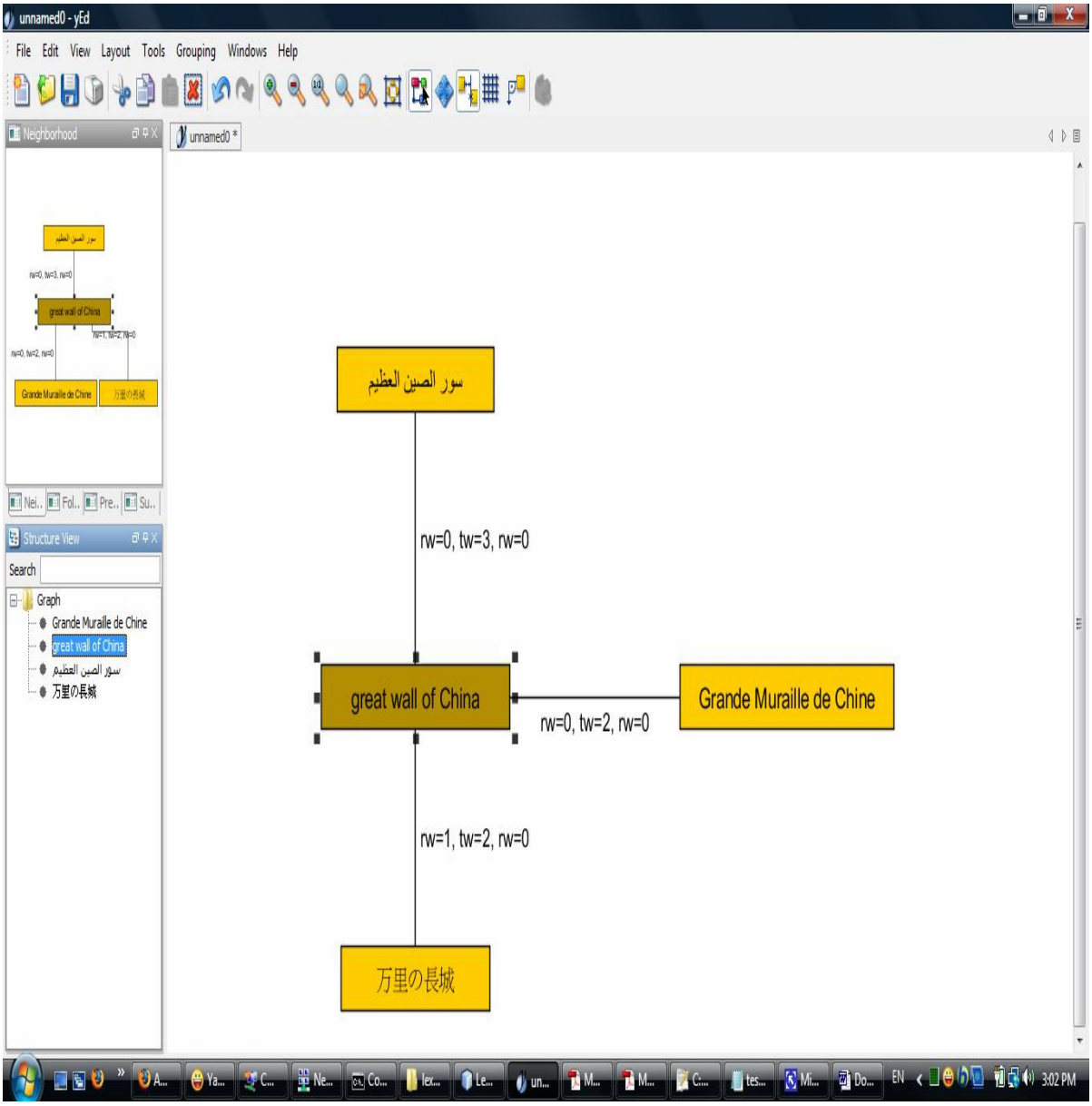


Fig. 48: Sample MPG

V.4.3 MPG ←→ Multilingual Database

It is possible to extract specific information from MPGs based on specific parameters, and after that we can represent information in XML format. An XML file can generate database entries to serve as a lexical repository.

Such a database consists of the following entities:

- Preterms: nodes on the graph, with all of their information;
- Monolingual pairs: preterm and its synonym;
- Multilingual pairs: preterm and its translation.

The next chapter discusses the details of the applications of MPGs.

Conclusion

The details of constructing Multilingual Preterminological graphs have been described. A graph can be developed with a combination of Passive and Active approaches. Passive approaches depend on domain-dedicated digital (textual or non textual) resources, while active approaches depend on human direct or indirect contribution.

Access log files can be a rich resource of hidden terminology, and they can be used to initialize the graph. Then the MPG is multilingualized using online resources, further expanded to benefit from its structure, and finally preterminology can be extracted in various formats.

Chapter VI System for Eliciting Preterminology (SEpT)

Introduction and Objectives

As presented in the previous two chapters, MPG preserves the preterminological sphere of a domain. Its structure and unconventional resources helps "eliciting" terminology by finding hidden correspondences between unrecognized terms.

MPGs can be developed using passive and active approaches, the initialization and multilingualization have been discussed in the previous chapter. They involve both passive approaches. On the other hand, further improvements need active approaches where members of the community are encouraged to contribute in an active way through the availability of initial data collected using passive contribution methods.

This chapter presents the design and the implementation of SEpT (7 in French, System for Eliciting preTerminology), a computer system that functions as a preterminology elicitor through building and maintaining MPGs. This chapter will present SEpT, its requirements, implementation and applications.

Section 1 shows the system design of SEpT and its functional requirements. Section 2 presents an instance of SEpT dedicated to the construction of an MPG for the resources of the Digital Silk Road Project (Ono, Kitamoto et al. 2008) (Ono, Kitamoto et al. 2007). Finally, section 3 describes another instance of SEpT integrated with a serious game (Arabic JeuxDeMots) (JDMAR 2010).

VI.1 Elicitation Design

This section analyses the requirements to build an elicitor for preterminology, and presents the design and technical architecture of SEpT.

VI.1.1 Automatic Elicitation

Automatic Elicitation has been explained in detail in the previous chapter. Basically, it is the construction and expansion of multilingual preterminology using the following:

- *digital (lexical, textual, non-textual) resources*, by compiling preterms and determining undefined possible relations.
- *lexical translation*, by translating preterms into various languages using multiple resources to determine translational relations.
- *MPG structure*, building heuristically edges and relations based on the information available in the graph itself.

VI.1.2 Human-based Elicitation

The human brain is the main source of knowledge in our approach, because the traditional methods of sharing knowledge through books, Internet, media, and alternative media, are difficult to automate.

For us, then, the source of multilingual preterminological knowledge is the multilingual community of a given domain. To build a multilingual preterminological repository we start

from the already shared knowledge (passive), and we engage the community in an activity of sharing multilingual preterminological knowledge.

Passive automatic approaches have been introduced to generate an MPG. This subsection explains the applications for engaging the community in an active process of sharing preterminology by improving the MPG progressively.

VI.1.2.1 Applications Eliciting Explicit Contributions

Some applications have been developed for the purpose of eliciting explicit contributions. Here application could be a method in a way; we will call it a "contributive application".

a. General Design

The general design of a contributive application is illustrated by figure 49. A community member interacts with the MPG through an application service that exchanges information with the member to provide him with a useful service. In return, the member can provide the application with lexical knowledge, that is then reflected in the graph to validate a relation or add a new one.

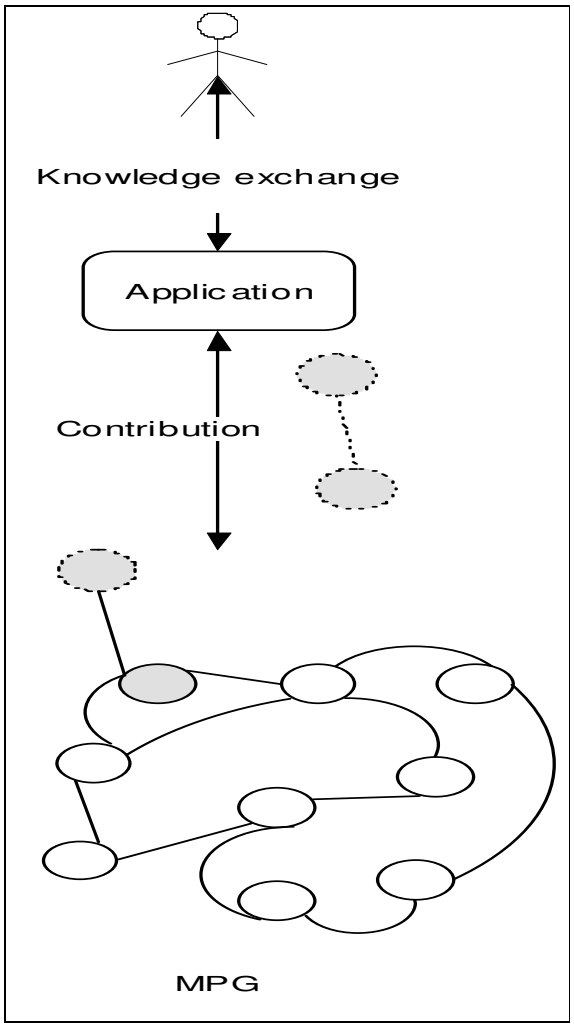


Fig. 49: Architectures of a contributive application

b. *Linguistic (Lexical)*

As we target lexical knowledge, the basic application (into which preterminological elicitation is integrated) is typically associated with a linguistic tasks, such as lexical translation and acquisition. Hence, this approach is more direct and more effective in the presence of professionals. As domain experts are rarely translators or terminologists, a basic linguistic activity may hardly attract them.. Moreover, if a community member who is not involved in the basic linguistic task uses the contributive task, there will be an unfair exchange of information: the community member will provide information and get nothing useful for him in return.

c. *Non-Linguistic (Community Related)*

Another option to involve and attract domain experts is to exploit a common online activity related to the community itself. Studying an online community of a domain to know what are the most common and repetitive activities makes it possible for the application designer to attach contribution elicitation devices to such an activity.

For example, in the case of reCAPTCHA (reCAPTCHA 2009) the common activity was creating an account which is done thousands of times a day. Attaching a contribution task to it guarantees a somewhat successful contribution scenario.

In the case of preterminology, we need vast amounts of contributions, not from general internauts but from persons who are interested in the domain.

VI.1.3 SEpT Functional Requirements

After introducing the key elicitation applications that SEpT should consider, this subsection presents the analysis of the main requirements of SEpT. The major purpose of SEpT is to create and maintain a preterminology through an MPG. Hence, such a system should offer the following key features:

- Building an MPG.
- Maintaining an MPG.
- Offline and online utilization of an MPG.

VI.1.3.1 MPG Construction

a. *MPG Level*

SEpT should offer the following features regarding the construction of an MPG:

1. analyzing access log files to extract preterms and possible relations between the extracted preterms;
2. multilingualizing an MPG using online resources;
3. finding synonyms;
4. analyzing the MPG to create heuristic edges based on the expansion formula.

b. *Nodes Level*

Concerning the nodes, SEpT should offer the following features:

1. identify, add, delete and modify any node on the graph;
2. multilingualize any node;
3. retrieve any node.

VI.1.3.2 MPG Usage

a. Online Usage

SEpT should offer the following online functionalities:

1. a protocol that allows users' applications to retrieve information from the MPG.
2. a protocol that allows users' applications to modify the content of the MPG.

b. Extraction (offline usage)

SEpT should store the MPG in a format that can be used and manipulated using different systems and applications.

VI.1.4 **Operational Architecture**

VI.1.4.1 3tier Architecture

Figure 50 shows a 3-tier view of the computer system that maintains a MPG. SEpT implements all the necessary functions for handling the MPG at the business layer. However, contributors have to interact with another application to get an access to SEpT. Such an application offers users a graphical user interfaces and web forms to exchange information with possible preterminology sources. The importance of this layer is that it allows various kinds of applications to interact with preterminology through SEpT.

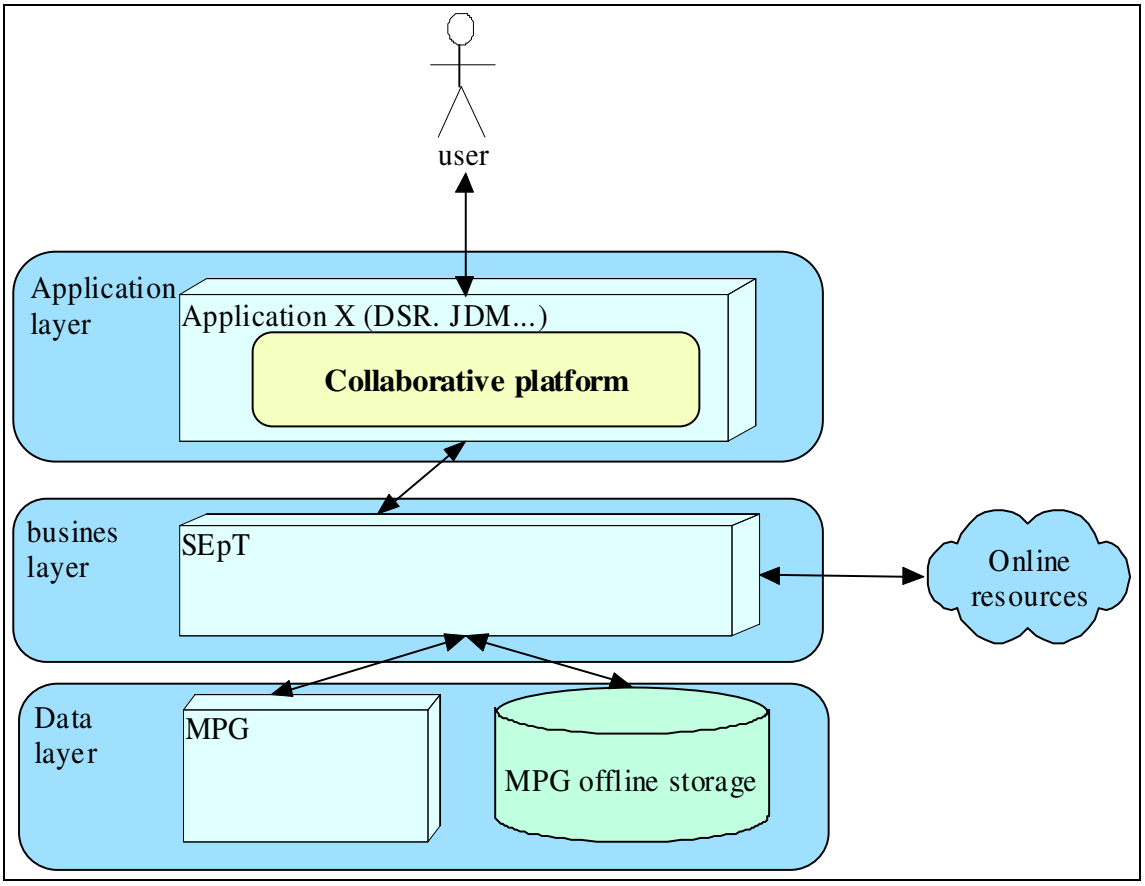


Fig. 50: SEpT's layers

VI.1.4.2 Data Layer

An MPG is manipulated by SEpT while it is in the machine memory, and SEpT exports and imports the content of the MPG into and from XML-based files to preserve the content and render it useful in other settings.

Other than the export/import functionalities, the data layer must offer specific knowledge retrieval features, and functions for adding/modifying preterminological content.

VI.1.4.3 Business Layer

The business layer is the core of SEpT. This layer provides all the construction requirements both at the node and the MPG level. It also interacts with other applications in order to allow them to manipulate and exploit an MPG.

VI.1.4.4 Application Layers

An application is a system that connects the user of an MPG with the MPG by exchanging information with SEpT.

The basic main functionalities that should be included in the application to make it a successful SEpT application are the following:

- Retrieving lexical data from the graph;
- Inserting preterminological data into the graph.

VI.1.4.5 The Fourth Layer in SEpT

The "fourth" layer in SEpT is the community of users who will use and improve the MPG. Although it is not a physical part of the system, it is the key entity in the overall solution.

VI.1.5 **SEpT Context Diagram**

VI.1.5.1 Level 0 Diagram

Figure 51 shows the context-independent diagram of SEpT. The following are the main external entities that interact with SEpT.

- Community users.
- Community resources.
- Online lexical resources.
- Offline storage.

And SEpT consists of the following main internal components.

- Graph Constructor.
- Graph Multilingualizer.
- Graph Expander.
- Graph Extractor.

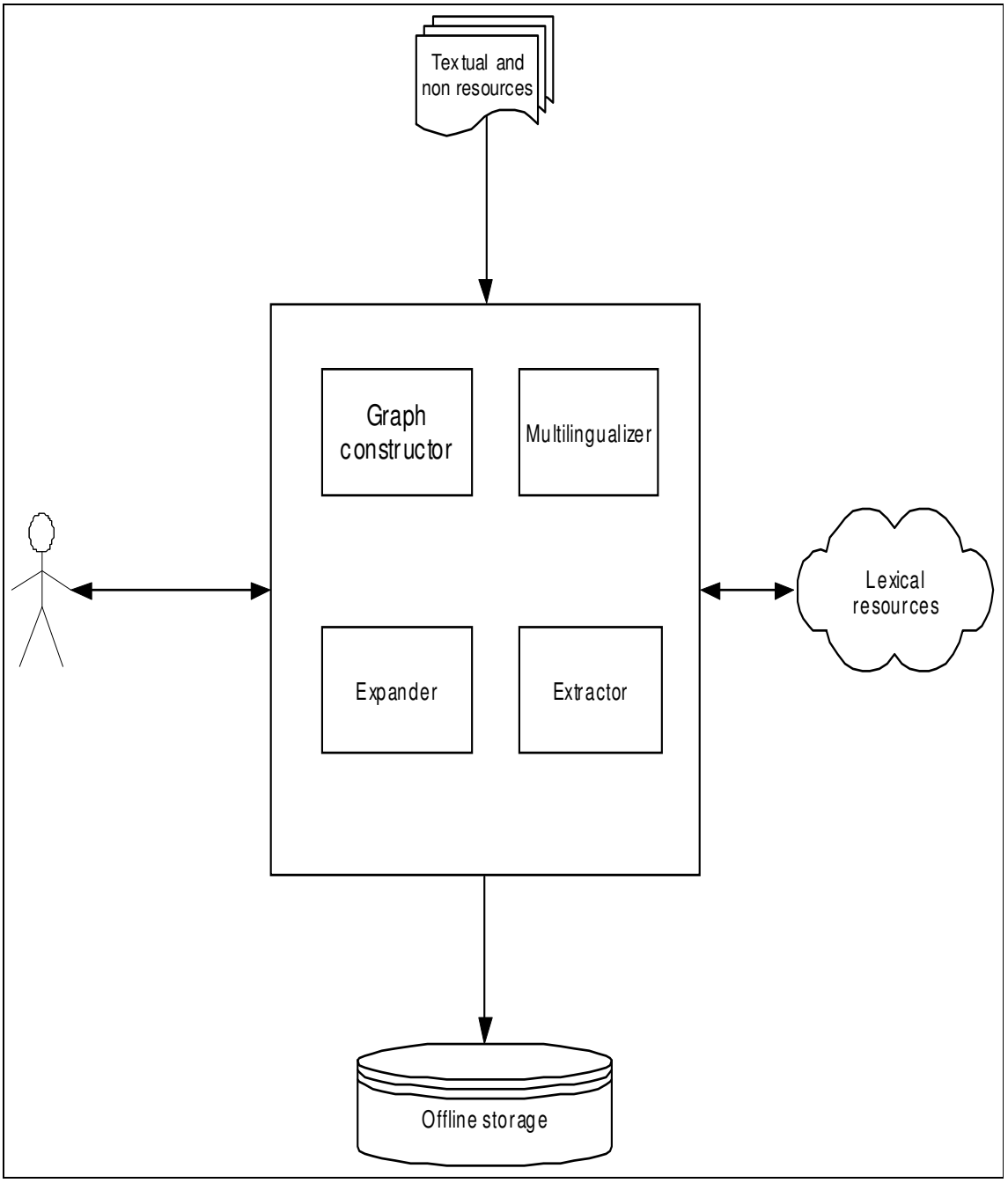


Fig. 51: SEpT's components

VI.1.5.2 External Components

a. Online Multilingual Resources

Online multilingual resources are the main source of monolingual and multilingual lexical data. Systems like Google Translate (Google 2008), Google Dictionary (Google 2010), IATE, Babylon, Systran,… offer online resources "on the cloud".

Such resources are called through http requests, and then results are processed to get the targeted lexical data.

b. Community Resources

SEpT uses textual and non-textual resources to extract terminology and construct and initialize the graph. Textual resources include related corpora, and non textual resources include recent access log files.

c. External Storage

For the graph to be used for other applications it needs to be stored in a common format as files or database (preTerminological Multilingual Database).

VI.1.5.3 Internal Components

a. Graph constructor

The Graph constructor is responsible for analyzing the access log files and extracting preterms and un-validated correspondences between the extracted preterms. It also calls a term extraction engine to extract terms from the internal corpus. The output of this component is an initial graph.

b. Graph Multilingualizer

The graph multilingualizer calls online resources to multilingualize each preterm in the initial graph. Figure 52 shows a typical way to call an online resource such as an MT system.

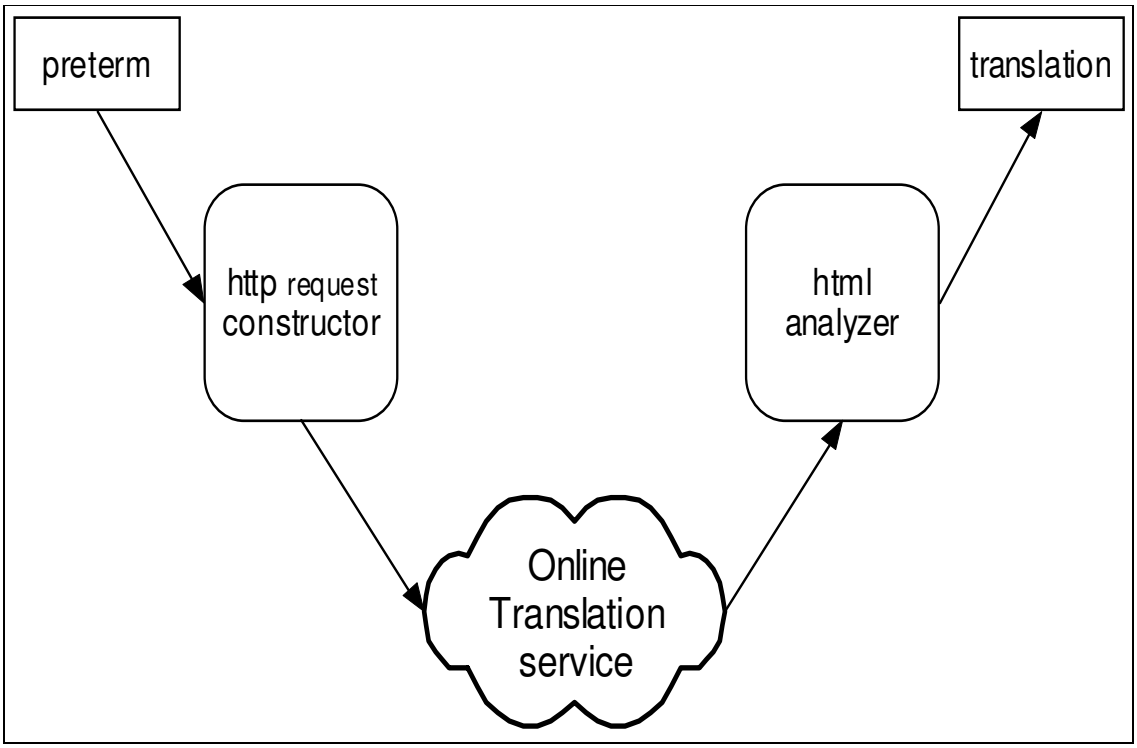


Fig. 52: Calling online multilingual resources

The http request constructor produces the URL that contains the preterm and the source and target languages. After receiving the results, the html analyzer finds the translation from the html page.

The http request constructor and analyzer have been implemented in Java. For every MT system used, there is a special configuration and parameters.

http request constructor configuration:

1. MT service url, ex: http://www.google.com/language_tools
2. Request parameters.
    a. Encoding.
    b. Text.
    c. Language pair.

As for the result analyzer, here are the most important configurations for every MT system to be used:

1. starting point of translated text in the html result template. Usually, it is an html tag with specific attributes;
2. marker of the end of translated text.

c. *Graph Expander*

The graph expander implements the formula for finding synonyms to heuristically create further edges and thus exploits the MPG itself to find new relations.

d. *Graph Extractor*

The extractor extracts sub-graphs, based on specific parameters and it stores the content of a graph into a standard format.

VI.2 SEpT - Digital Silk Road Project

This section presents a specific instance of SEpT dedicated to build a preterminological repository of the domain of the historical resources of the Digital Silk Road (DSR) Project and shows the possible application on that instance of SEpT to archive active contribution from the community of the DSR.

VI.2.1 Overview of the Digital Silk Road Project

The Digital Silk Road project (Ono, Kitamoto et al. 2007) (NII 2003) is an initiative started by the National Institute of Informatics (Tokyo) and the United Nations (UN) in 2002, to archive cultural historical resources along the Silk Road, by digitizing them and making them available and accessible online.

The project includes several sub-projects, each of them is working on a specific type of data to be digitized. For example “Silk Road Maps” is a sub-project where old maps of the historical Silk Road maps are collected and computerized and matched with the current maps.

Another sub-project is the DSR Imaginary Museum, which acts as an online museum of the Silk Road.

One of the most important sub-projects is the Digital Archive of Toyo Bunko Rare Books (NII 2008). Figure 53 shows a screenshot of the main page. In the framework of this project, tens of old rare books available at Toyo Bunko library have been digitized using OCR (Optical Character Recognition) technology. The digitized collection contains books in different

languages (English, French, Russian…), all of them related to the historical Silk Road, like the two volumes of the Ancient Khotan by Marc Aurel Stein.

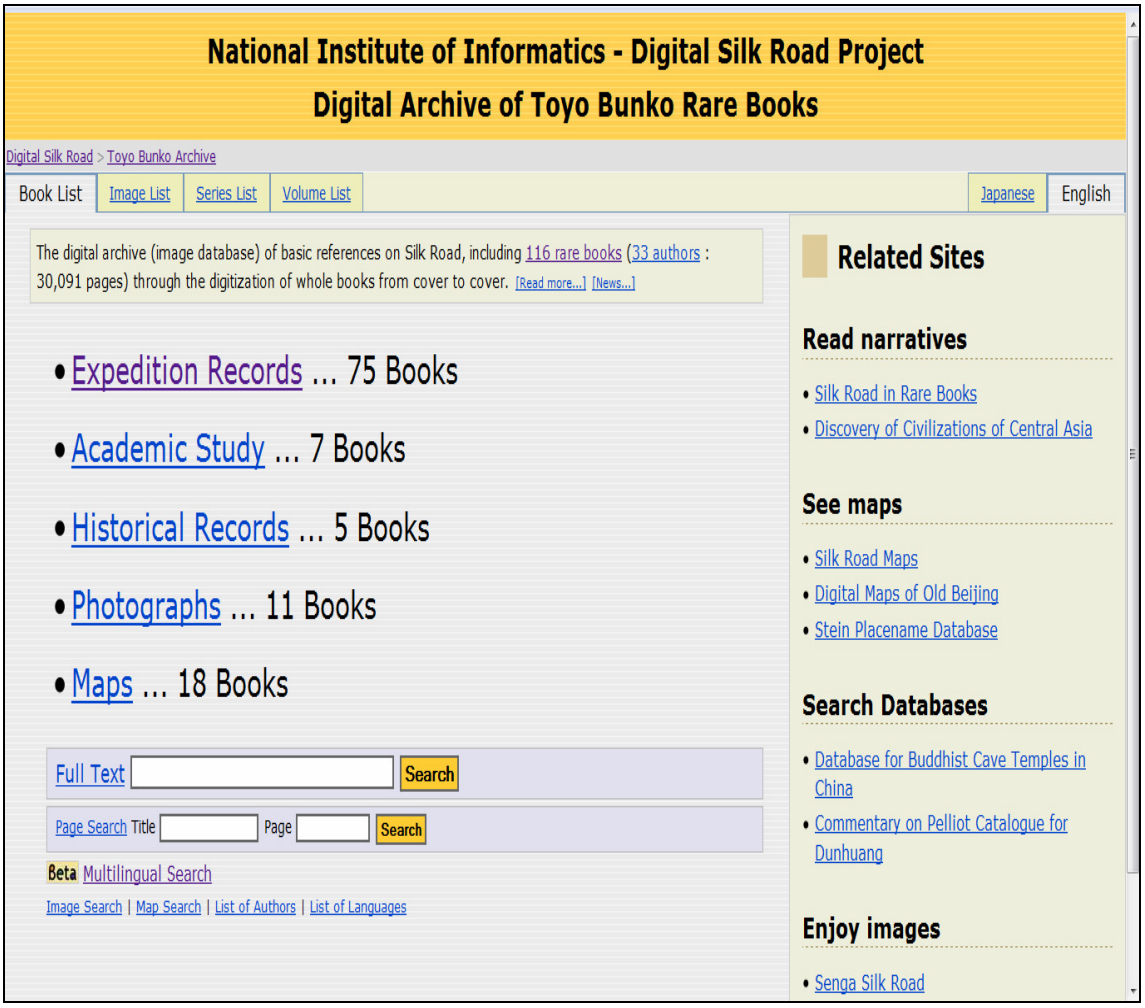


Fig. 53: Digital Archive of Toyo Bunko

VI.2.2 **Available Resources**

The purpose of this digital archive is to make "invisible" precious books visible by everyone, because accessibility to such books is restricted due to their fragility and safety.

The Toyo Bunko library in Tokyo (Oriental Library) is the main collaborator. It has a collection of 880,000 books of historical importance. One of the most interesting collections is the library called "Morrison Library," which consists of 24,000 books about China and Asia written in several European languages. Starting from 2002, DSR has been digitizing the most relevant and important books of this collection. So far 116 books (30,091 pages) have been digitized into a database of images. OCR was applied to obtain the textual material for most of these books; however the quality of the resulting text varies depending on the language and the condition of the physical book.

There is a need to ease the access of this collection and translate some parts of the books into various languages, so it was desired to build a term base for this collection to serve the multilingual community. An initial base of 500 terms was built in 2005, but it was not sufficient for the large terminological sphere of these books.

VI.2.3 **The Need for Multilingual Terminology**

VI.2.3.1 Multilingualism at the DSR

The interfaces of the website of the Digital Silk Road are offered in two languages, English and Japanese.

However, the digitized contents are available in more languages. Table 7 shows the languages and number of books in each one.

Table 7: Languages of the archived books

Language	Number of books	Notes
English	47	
French	14	
German	13	
Russian	11	
Chinese	5	
Swedish	4	
Arabic	4	Bilingual (AR-FR)
Japanese	2	
Italian	1	

Most of the titles and the caption text of the figures of these books have to be available in English and Japanese as well.

VI.2.3.2 Analyzing the Visitors' Linguistic Backgrounds

For the purpose of knowing the linguistic identity of the visitors of the DSR website, the access logs since 2003 to the website have been analyzed.

Table 8 shows the countries from which DSR is accessed; knowing the country of a visitor may indicate his linguistic knowledge. The table shows that around 60% of visitors are from countries other than Japan.

Table 8: Countries of the DSR website visitors (from Jan/2007 to Dec/2008)

Countries	Visitors	Percentage	Assumed languages	Available books
Japan	117782	41.41%	JA	2 books
China	30379	10.68%	ZH	5 books
USA	15626	5.49%	EN	47 books
Germany	8595	3.02%	GE	13 books
Spain	7076	2.49%		
Australia	5239	1.84%	EN	47 books
Italy	4136	1.45%	IT	1 books
France	3875	1.36%	FR	14 books
Poland	2236	0.79%		
Russia	1895	0.67%	RU	11 books
Sweden	1890	0.66%	SV	4 books
UK	1883	0.66%	EN	47 books
India	1878	0.66%		
South Korea	1860	0.65%		
Iran	1821	0.64%		
Taiwan	1809	0.64%		

Pakistan	1780	0.63%		
Thailand	1773	0.62%		
Hong Kong	1765	0.59%		
Brazil	1690	0.62%		
Canada	1643	0.58%	EN	47 books
Indonesia	1595	0.56%		
Switzerland	1589	0.56%		
Norway	1474	0.52%		
Greece	1458	0.51%		
Netherlands	1392	0.49%		
Vietnam	1294	0.45%		
Austria	1278	0.45%		
Finland	1225	0.43%		
Turkey	1197	0.42%		
Kazakhstan	1072	0.38%		
Saudi Arabia	978	0.34%	AR	4 books
New Zealand	965	0.34%	EN	47 books
Egypt	947	0.33%	AR	4 books
Malaysia	919	0.32%		
Uzbekistan	908	0.32%		
Romania	894	0.31%		
Jordan	868	0.31%	AR	4 books
Czech republic	798	0.28%		
Lebanon	735	0.26%	AR	4 books
Kyrgyzstan	720	0.25%		
Portugal	698	0.25%		
Other: Philippines, Cyprus, Syria, Iraq, Bulgaria, South Africa, Tunis…	8089	2.84%		
Unknown	36688	12.90%		
Total	284412	100.00%		

The DSR interface is available in 2 languages (EN, and JP). So visitors should have some knowledge of English or Japanese before accessing the website. Hence, for example, one could assume the importance of an EN ↔ DE repository as the percentage of the visitors from Germany is relatively high; add to that the availability of archived German books online at the DSR website.

VI.2.3.3 MPG for the DSR

DSR needs an MPG to serve its multilingual community as a repository of terminology for multilingual search and general consultation purposes.

At the same time, DSR offers rich passive and active resources, as it has more than 30,000 pages of text and log files since 2003 and a community of loyal multilingual visitors who are interested in the domain and capable of contributing to an MPG.

An instance of SEpT has been assigned to DSR to produce an MPG (MPG-DSR). However, achieving progressive enhancements to the MPG and attracting contributions need useful applications. Then next subsections present two suggested applications for the DSR.

VI.2.4 **Applications: CWS, and CWR**

For a historical archive like the DSR, we find that reading and searching were the most important for users. Log files since 2003 show that 80% of the visitors where interested in reading the historical records. Moreover, around 140,000 search requests have been sent to the internal search engine.

As the visitors are not likely to be linguists or professional terminologists, it is desired to attach the contribution to the MPG-DSR to a community-related activity, such as reading archive and searching. So this section presents two suggested applications for DSR (1) "contribute-while-reading" and (2) "contribute-while-searching" to build and improve the MPG-DSR through SEpT-DSR.

VI.2.4.1 Contribute While Searching (CWS)

As shown in figure 54, historical physical books have been digitized and indexed into a SOLR-based search engine (Apache 2008).

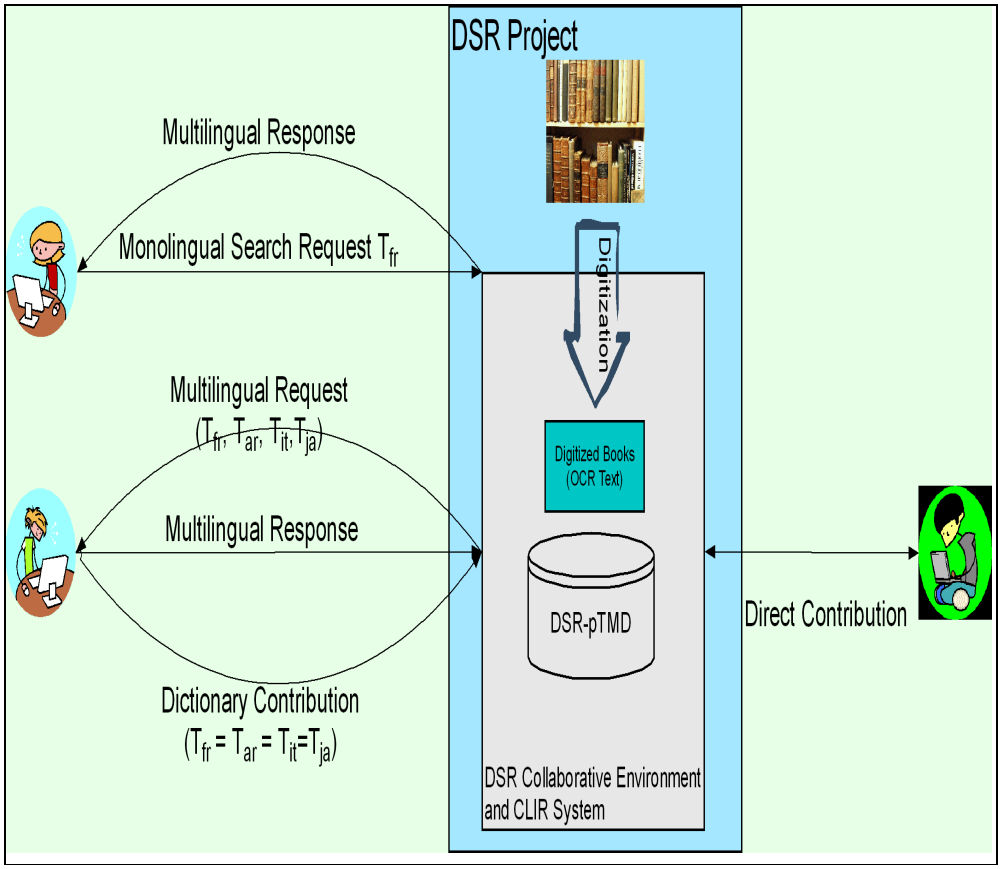


Fig. 54: General architecture of the environment

Users are expected to send monolingual search requests in any language supported by the application to get multilingual answers. Having a term base of multilingual equivalences could achieve this (Oard 1999) (Chen 2002). A bilingual user who could send a bilingual search request could be a valid candidate to contribute. In fact, the same bilingual request could be a valid MPG contribution, and also multilingual requests. This way, users who use the search engine will use the DSR-MPG to translate their requests and will contribute to the MPG spontaneously.

Fig. 55: A Japanese user translating his request

As figure 55 shows, a Japanese user can translate his search request, to get more results, as shown in figure 56. Then s/he sends the request to view the results, see figure 57.

Fig. 56: The term has been translated

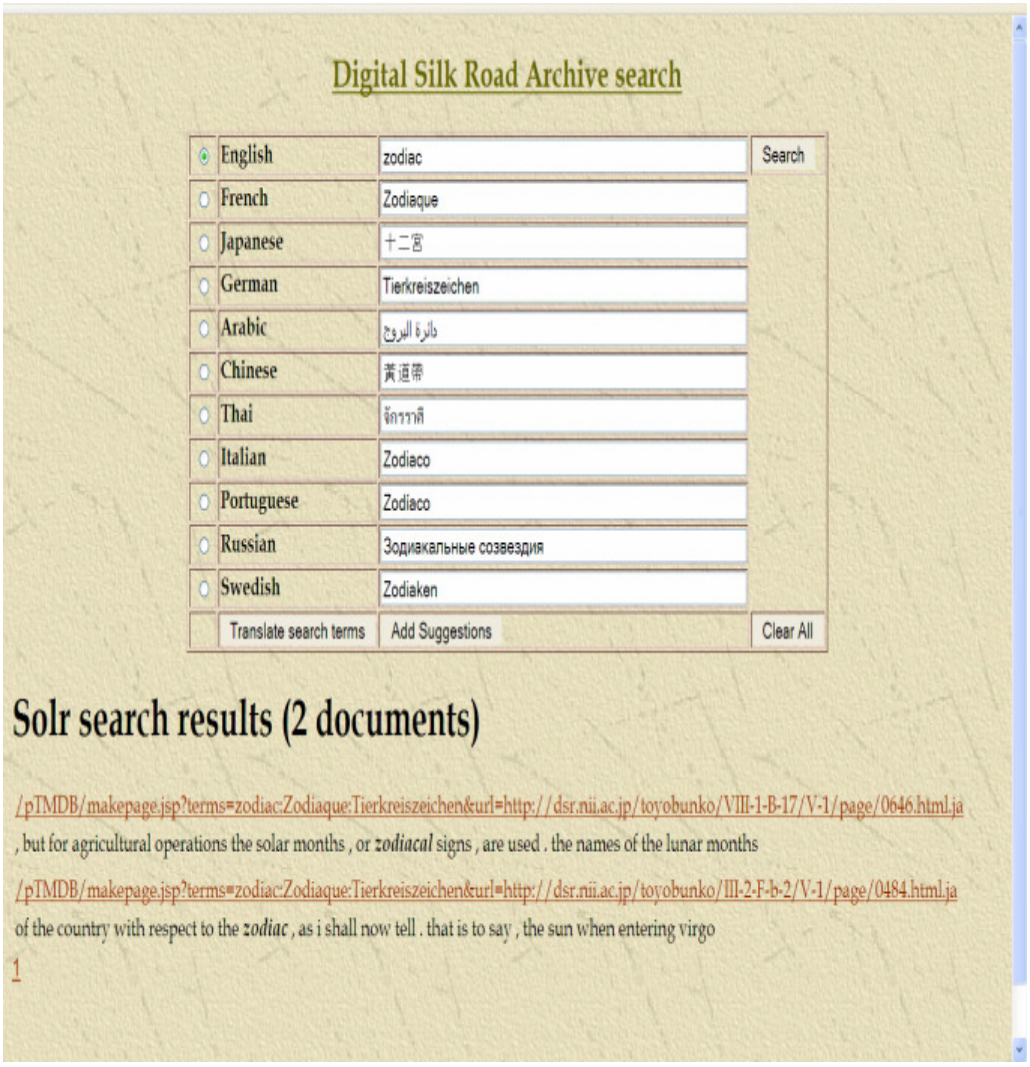


Fig. 57: Search results

During the search process, the user can add a new translation if s/he was not happy with the suggested translation, by clicking on "Add Suggestions". The form showed at figure 58 then appears.

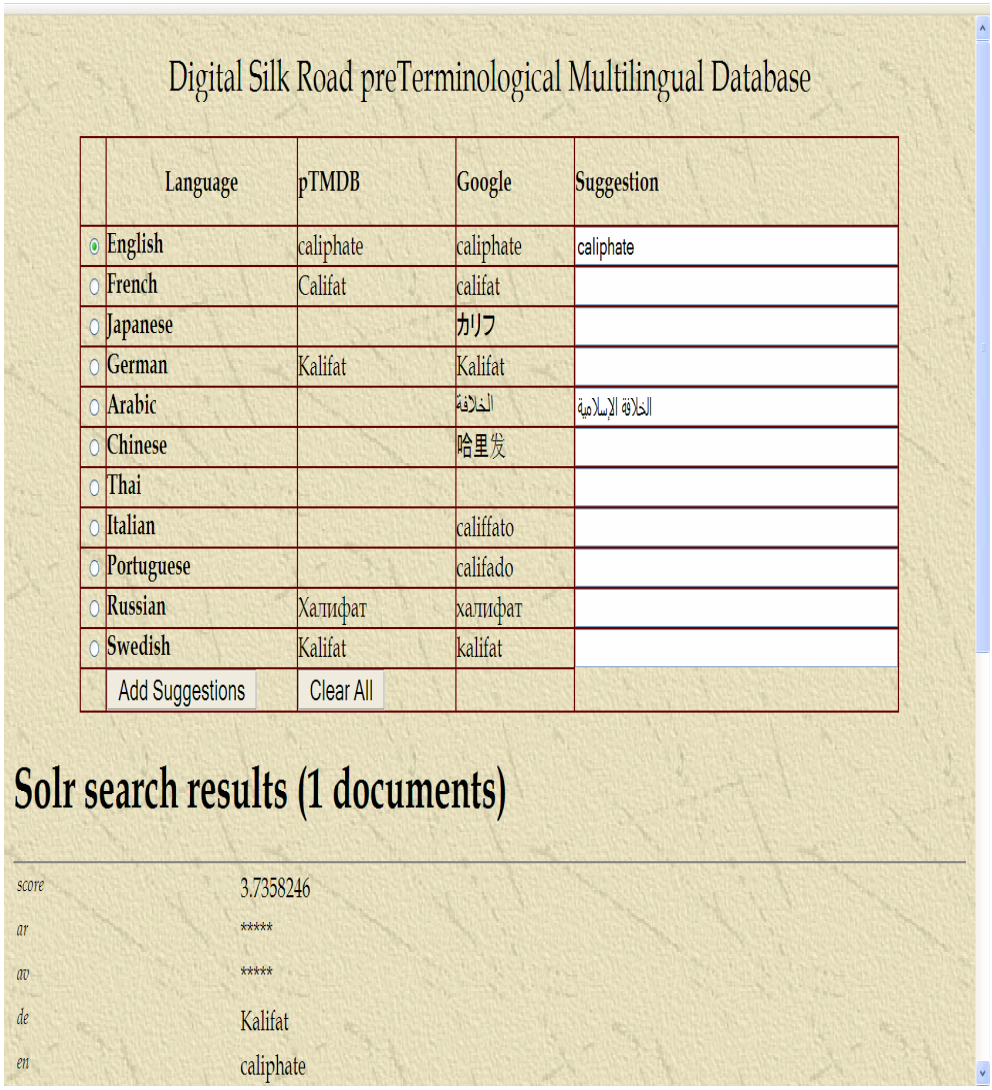


Fig. 58: Contribution page

VI.2.4.2 Contribute While Reading (CWR)

The other interesting application is trying to help users from different linguistic backgrounds to translate some of the difficult terms into their languages while they are reading, simply by selecting a term from the screen.

As shown in figure 59, readers see a page from a book as an image, with its OCR text. Important terms are presented with a yellow background. Once a term is clicked, a small child contribution/lookup window opens, similar to the one in figure 58. A user can also lookup/translate any term from the screen by selecting it.

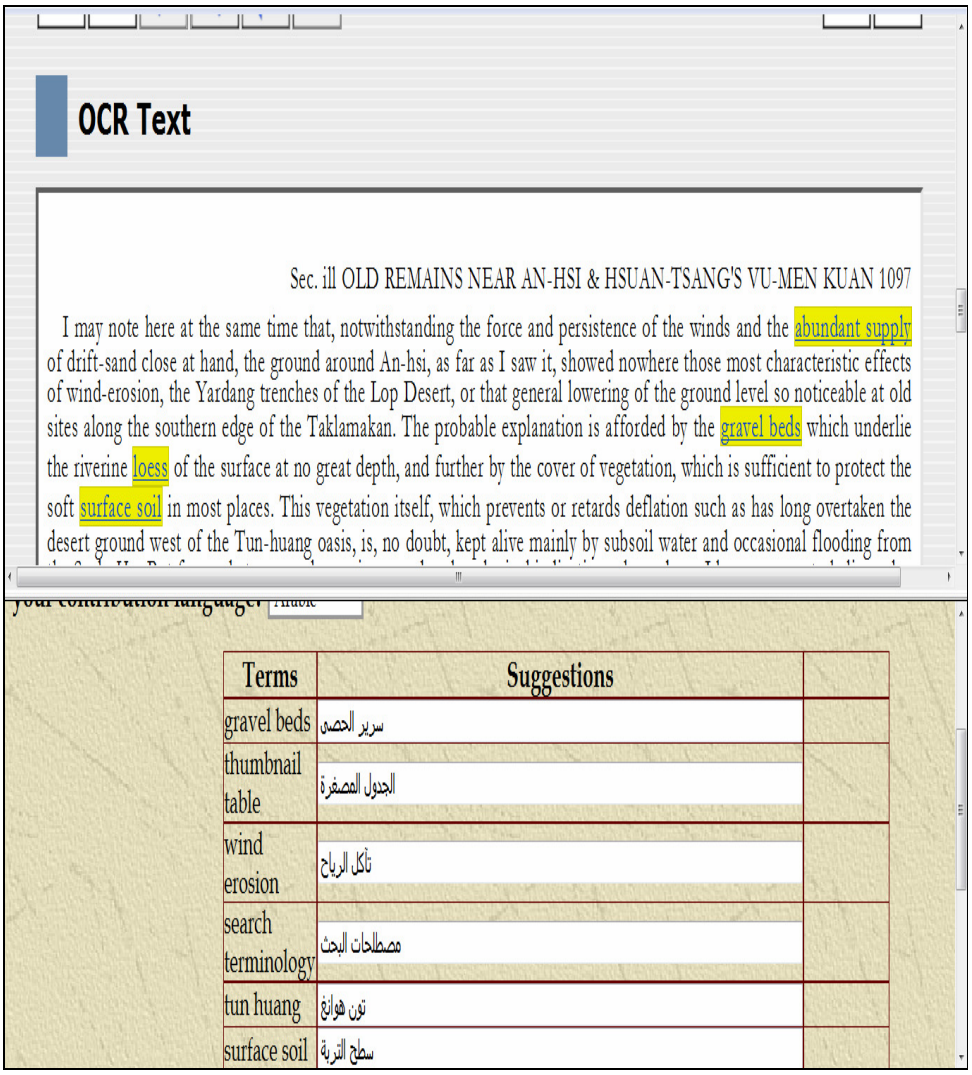


Fig. 59: CWR

VI.3 SEpT ←→ JeuxDeMots

SEpT develops preterminology and offers it for other applications to be expanded and enlarged collaboratively. One interesting active collaborative approach is to rely on serious games to "extract" human knowledge, that is, in the case of SEpT, multilingual preterminological knowledge.

One of the current most relevant games is JeuxDeMots which is an online game that aims at collecting lexical units, and builds lexical functions between them through an entertaining process; the collected data is stored as a graph.

This section presents the utilization of JeuxDeMots as an application of SEpT to attract preterminological contributions through a non-linguistic activity.

VI.3.1 **JeuxDeMots**

VI.3.1.1 Basic Concepts

a. Game Overview

Players create an account to play the game. After that at the first level of the game, the game offers the player a term and within 1 minute he has to input as many lexical units associated with the given term as he can, figure 60. If his answer matched the answers of some previous player at the same term, both of them will have more point and their ranking will go up.

Fig. 60: English JeuxDeMots

At further levels of the game, players are asked to give values of some lexico-semantic functions. For Example, to give values for "Magn" on "fever", JeuxDeMots will show "fever" and ask:

<<give some words to express the intensity of this word>>

b. Lexical Graphs

At level 1, if player1 associates the word "dog" with the word "bark" and player2 does the same, this relation gains more confidence. This is reflected on the created graph by making an edge between dog and bark. This edge is labeled with the type of relation and the weight. For example, the following are the relations associated with dog in the English JeuxDeMots:

- dog `---r_associated#0:60-->` bark
- dog `---r_associated#0:60-->` pet
- dog `---r_associated#0:50-->` bone
- dog `---r_associated#0:50-->` cat

The density of such a graph increases with the amount of players.

VI.3.1.2 Current Versions and Achievements

a. French JeuxDeMots

As the developers of the game are French, the French version was the first to be played. Since 2007, the game has collected more than 1,013,800 relations and 222,034 terms. The game has been played by more than 500 active players.

To achieve such success, the game needs moderation and administration. Furthermore, it needs some kind of community spirit between the players. That is why the game has been equipped with an online forum and social networking options.

b. Other versions

The game is available in other languages. Due to the lack of constant moderation, these versions were not as successful as the French one, but they still showed good potential. The following are the current available versions:

- English: 654 relations, and 35,390 (most of the terms are not contributed).
- Arabic: 1,336 relations, and 2,100 terms.
- Vietnamese: 380 relations, and 97 terms.
- Thai: 108 relations, and 54 terms.
- Japanese: 209 relations, and 116,877 (most of them imported from a dictionary).
- Spanish: 383 relations, and 12,825 (most of the terms are not contributed).

VI.3.2 **Arabic JeuxDeMots**

A SEpT has been assigned to the Arabic-based MPG and as an application Arabic JeuxDeMots was initialized using the Arabic-based MPG, hoping this would expand the graph and enrich its content. This subsection discusses the technical details of integrating SEpT with online serious game.

The initial Arabic JeuxDeMots (JDMAR) (JDMAR 2010) was localized by translating its interfaces in November 2009; in January 2010 it was played by the public.

The initial version was provided with 750 nodes of Arabic preterms to be expanded and developed as a lexical graph. Now there are 1800 preterms, and 1336 relations. The game has been played by more than 50 active players during a period of 7 months.

It needed moderation and advertisement through email, online forums, and online social networks (see figure 61).

Fig. 61: Arabic JeuxDeMots, home page

VI.3.3 MPG ←→ JDM Graph Structure

The players of JDMAR do not interact directly with the MPG, but their contributions are stored in the lexical graph of JDMAR. This graph can be extracted in a GraphML format which is compatible with the structure of MPG, so after receiving contributions, the JDMAR intermediate graph can enrich the content of MPG.

The nodes of the JDMAR graph match the nodes of the MPG, and the same goes for the language. The edges are preserved by converting the weight of the association relation into the rw as both of these relations are not validated. Figure 62 shows an example of a simple graph.

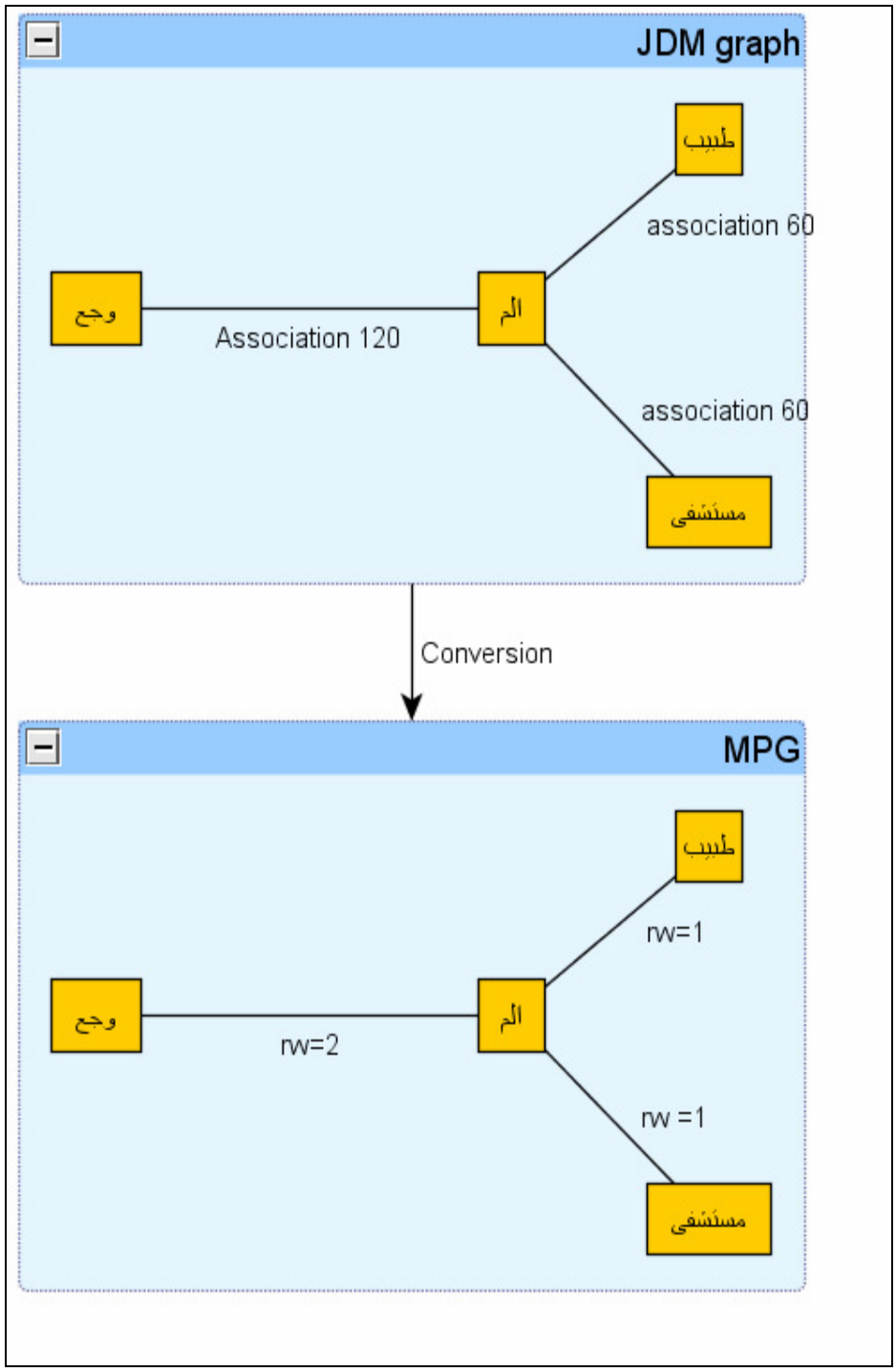


Fig. 62: JeuxDeMots graph to MPG

Conclusion

This chapter presented the design and the implementation of SEpT (7 in French) (System for Eliciting preTerminology), which is a computer system that functions as a preterminology elicitor through building and maintaining MPGs. The basic design and requirements have been laid out.

SEpT has three main operational layers, starting from the Graph which is the data layer where the preterminology and its relations are stored. The main manipulation functionalities are implemented at the second layer (business layer), and finally the application layer interacts with active contributors.

Two main instances of SEpT have been introduced, the Digital Silk Road Project (SEpT-DSR) and the Arabic JeuxDeMots. Both use SEpT to build specialized preterminology, and both act as applications to attract active contributions to their graphs.

Conclusion of Part B

Preterminology must preserve the structure of a conceptual system. A hierarchical ontological based structure needs specialists and is difficult to develop and manage, while a lexical structure can not conserve the structure of the conceptual system. Multilingual Preterminological Graphs have been prepared to make it easy to construct unconfirmed conceptual and translational relations between preterms.

Multilingual Preterminological Graphs (MPGs) have been defined and described as containers of multilingual preterminology of a domain. MPGs have nodes and weighted edges, where the nodes represent preterms and an edge represents a relation between two terms, the weight thereby indicating the nature of this relation and its reliability.

A graph can be developed with a combination of passive and active approaches. Passive approaches depend on domain-dedicated digital (textual or non textual) resources, while active approaches depend on human direct or indirect contribution.

Access log files can be a rich resource of hidden terminology. They can be used to initialize the MPG, which is then multilingualized using online resources, and further expanded to benefit from its structure. Preterminology can be extracted from an MPG in various formats.

A system to develop and maintain MPGs has been introduced, SEpT (7 in French) (System for Eliciting preTerminology), which is a computer system that functions as a preterminology elicitor through building and maintaining MPGs. SEpT interacts with an MPG and with external applications, that must offer facilities to view the graph and receive active contributions.

Part C Experimentation and Evaluation

Introduction à la partie C « Expérimentation et évaluation »

Dans un effort pour résoudre le problème de la terminologie latente et de l'absence de terminologie, la préterminologie a été suggérée comme un moyen facile de développer des ressources lexicales en utilisant des ressources et des approches non conventionnelles. Pour correspondre à la nature dynamique de la préterminologie et son matériau non vérifié, une structure de graphe a été proposée à accumuler des données préterminologiques provenant de diverses ressources. SEpT (Système pour Éliciter la préTerminologiques) est responsable de la construction et de la gestion d'un tel graphe (MPG). Ce graphe est initialisé en utilisant des approches passives, et sa densité comme son matériau est amélioré grâce à des approches à contribution active utilisant un ensemble d'applications qui peuvent motiver des experts du domaine dans le processus d'élaboration du graphe.

Cette partie traite de la mise en œuvre, l'expérimentation et l'évaluation de SEpT et de ses approches. Nous commençons par les détails d'implémentation de SEpT et de ses composants, puis évaluons sa performance en tant que système, et enfin décrivons des expériences sur les approches actives et passives et leur efficacité dans la résolution des problèmes liés à la terminologie latente et absente, et à la mauvaise couverture linguistique et lexicale en général.

Le chapitre 7 explique les aspects techniques de SEpT, son instanciation et son expérimentation. Le chapitre 8 décrit les expériences sur le développement de préterminologie grâce des approches passives. Le chapitre 9 montre le potentiel des approches à contribution active.

Introduction to Part C

In an effort to solve the problem of latent terminology and absent terminology, preterminology has been suggested as an easy way to develop lexical resources using unconventional resources and approaches. To match the dynamic nature of preterminology and its unverified material, a graph structure has been proposed to accumulate preterminological data from various resources. SEpT (System for Eliciting preTerminology) is responsible of building and maintaining such a graph (MPG). It is initialized using passive approaches and its density and material is improved through active contribution approaches using a set of applications that can motivate domain experts in the process of developing the graph.

This part discusses the implementation, experimentation and evaluation of SEpT and its approaches. We start from the implementation details of SEpT and its components, then evaluate its performance as a system, and finally describe experiments of the active and the passive approaches and their effectiveness in resolving latent and absent terminology, and poor linguistic and lexical coverage in general.

Chapter 7 explains the technical aspects of SEpT, its instantiation and experimentation. Chapter 8 describes the experiments on developing preterminology through passive approaches. Chapter 9 shows the potentials of active contribution approaches.

Chapter VII SEpT Usage and Experimentation

Introduction and overview

This chapter provides a brief technical overview and documentation of SEpT and its implementation.

The first section presents some technical implementation details, the second section shows how an instance of SEpT can be used, and the third section evaluates SEpT as a system in terms of its performance and scalability.

VII.1 SEpT Implementation

This section introduces the components of SEpT, their internal and external interactions, and some interesting technical aspects in implementing such components. It also shows how SEpT represents MPGs in the computer memory.

VII.1.1 Data Flow Diagrams

VII.1.1.1 General Operational Diagram (the Case of the DSR)

The following diagram (figure 63) shows how SEpT interacts with the external entities to serve an application like the DSR. The graph is constructed, maintained, and even exploited by a set of applications.

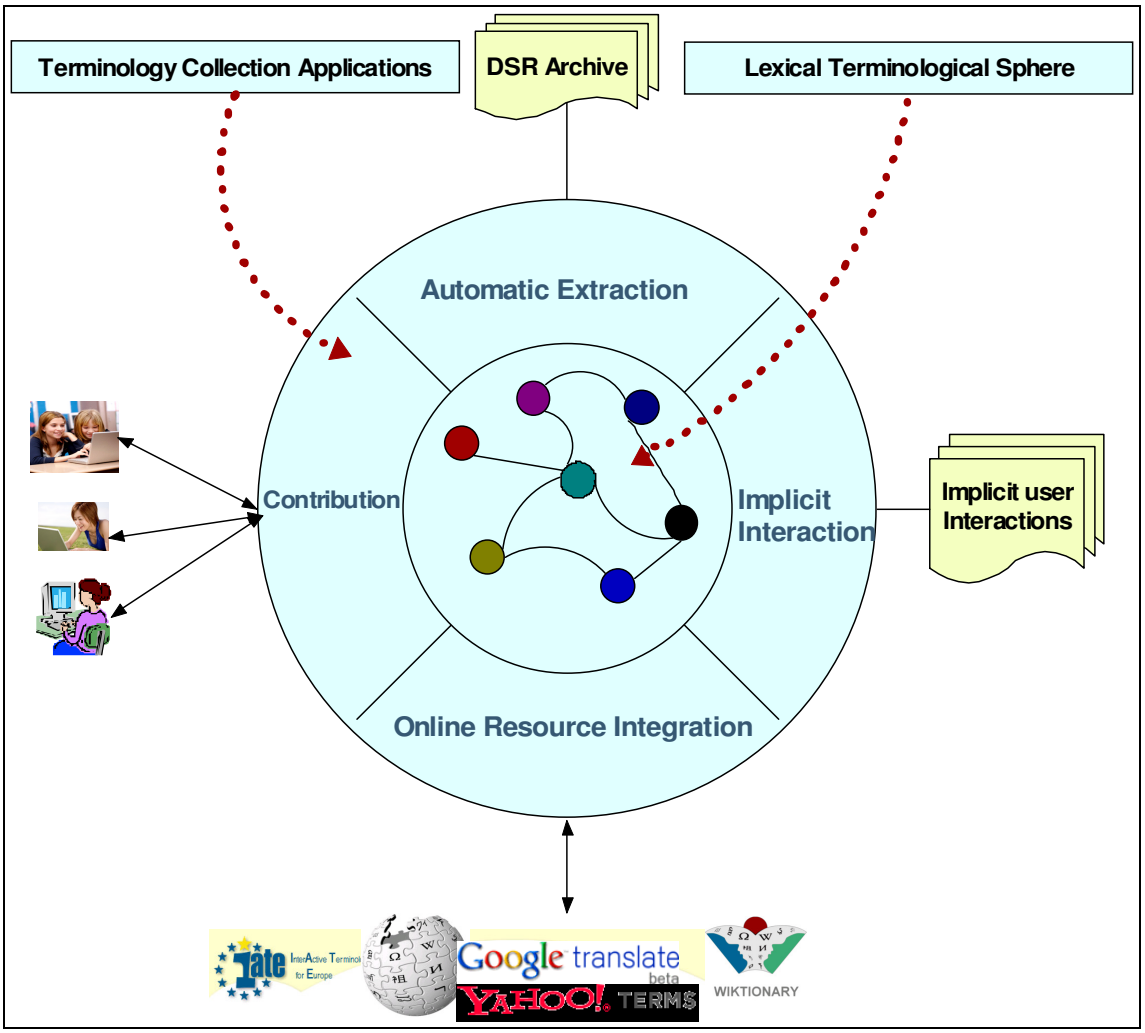


Fig. 63: SEpT-DSR, operational architecture

VII.1.1.2 Data Flow Diagram, Level 1

The following diagram, figure 64, shows the flow of data in SEpT, and presents the external and internal entities of the system. Internal entities (modules and data containers) are part of SEpT. External entities provide the system with raw data and/or exploit the produced MPG.

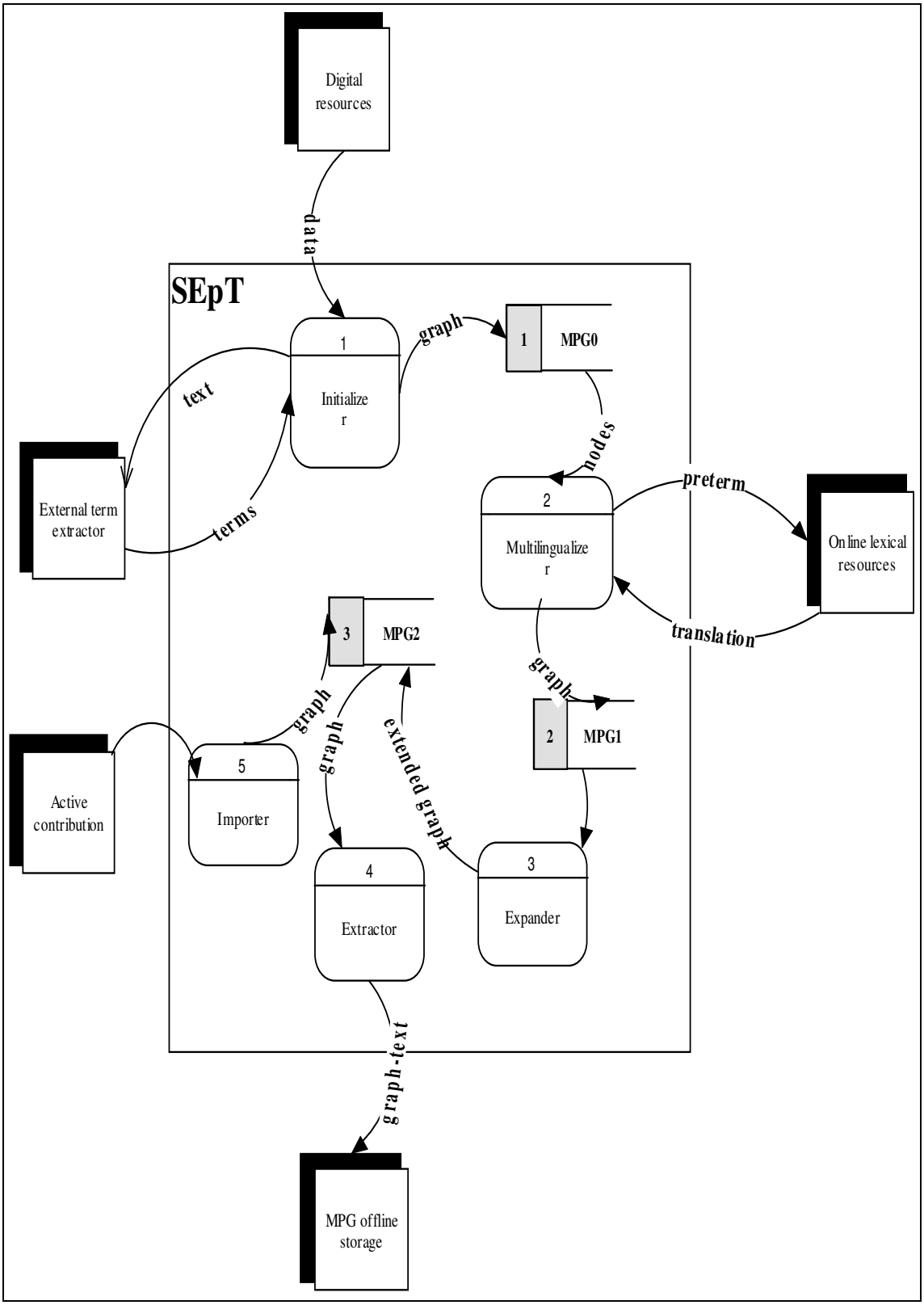


Fig. 64: SEpT, DFD1 (Gane & Sarson style DFD)

VII.1.1.3 SEpT Data Flow Entities

a. Initializer

Sub-components:

- Log file analyzer: this subcomponent analyses access log files sequentially to extract lexical data and create initial relations.
- Term Extractor: it receives textual data and sends it to an external term extraction engine to retrieve a list of candidate preterms.
- Term Merger: in merges the initial graph with the preterminology extracted from textual data.

Input: pre-processed access log files and corpus.

Output: initial MPG.

b. Multilingualizer

Sub-components:

- HTTP request creator: it receives a preterm and its language to formulate an http request to translate it.
- HTML analyzer: it receives the result from the online lexical database and analyzes it to extract the translation.

Input: initial graph.

Output: multilingualized graph.

c. Expander

Sub-components:

- SW calculator: it computes the SW for all the edges.
- Heuristic edge creator: it creates edges based on SW.

Input: MPG1, and parameters.

Output: MPG2.

d. Extractor

Input: MPG2, and parameters.

Output: MPG in GraphML.

e. Importer

Input: preterms and relations in GraphM.

Output: nodes and edges in the memory.

f. Data Stores

MPG0: the initial graph before multilingualization.

MPG1: multilingualized MPG.

MPG2: Expanded MPG.

VII.1.2 **SEpT Classes Diagram**

The following is a UML Class Diagram, figure 65, it shows the classes and main objects that need to be implemented in SEpT.

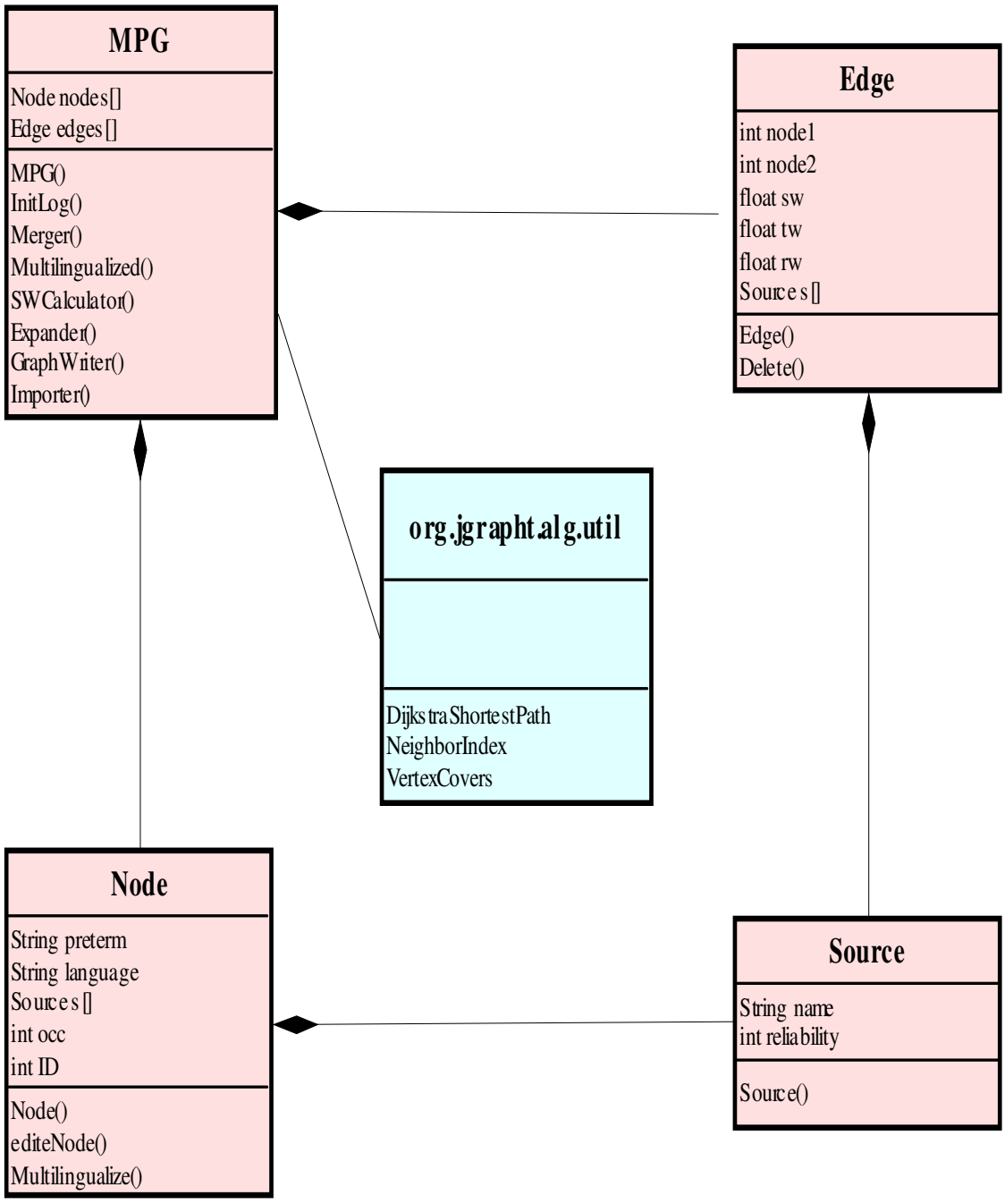


Fig. 65: SEpT Java class library

VII.1.2.1 MPG

As shown in the previous diagram, the main graph consists of two lists objects, one for the nodes and the other for edges. There is also an object encapsulating all the operations described in Part B on MPG, to create, maintain, expand and transport it.

VII.1.2.2 Nodes

The nodes are a list of preterms and their information, as described in Chapter 4.

VII.1.2.3 Edge

An edge object is a list of couples of nodes and the weights to represent the nature and reliability of the relation between these two edges.

VII.1.2.4 Utilities

Basically, the graph is implemented as two (one dimensional) arrays, one for the nodes, and one for the edges. jGraph and JGraphT (jGrapht 2010) are called by MPG for certain methods.

JGraphT is a free Java graph library that implements mathematically well-defined objects and methods for graph-theory. JGraphT supports various types of graphs, including undirected weighted graphs. It offers many utilities that can simplify the manipulation of MPG and SEpT in general.

The source class represents the source of the preterm, so we can track all the confirmation on the term itself and on its translation.

VII.2 SEpT Utilization

This section explains how to use SEpT in an application through a Java library that can be used and maintained easily.

VII.2.1 SEpT Java Library

VII.2.1.1 Why Java?

Java is a "general-purpose, concurrent, class-based, and object-oriented programming language, and is specifically designed to have as few implementation dependencies as possible." (Java 2010)

Java is chosen to implement SEpT because of the following reasons:

- It is an object-oriented programming language, meaning that the necessary objects needed to represent the preterminological graphs and their components can be implemented using Java.
- Many researchers invested on developing reusable Java libraries related to lexical resources, graph structure, and XML files, that can be used by SEpT.
- Java libraries can be used by average developers even for non-Java applications, which is essential for SEpT as it should be used by a variety of applications for active and passive contribution.

VII.2.1.2 Java Classes

a. *node*

One of the most important classes in the library is the node class, which represents a node on the graph; the following is a code snippet of the definition of the class that shows the main variables of the class:

```
public class node {
    private String preterm;
    short  occ;
    private int ID;
    private String lang;
```

```
..
}
```

b. *edge*

The following is the Java code showing the main variables of the edge class:

```
public class edge {
    private int source;
    private int target;
    private float rw;
    private float sw;
    private float tw;
    public int id;
..
}
```

c. *MPG*

The following Java code shows how the nodes and edges are implemented in the MPG class as two lists:

```
public class MPG {
    private String name;
    private static node[] n = new node[10000000];
    private edge[] e = new edge[10000000];
    int ncount;
    int ecount;
..
}
```

VII.2.2 **SEpT Instantiation**

VII.2.2.1 Basic Concept

SEpT as a library is a Java JAR that can be included (imported) in any Java project. Then an instance of the class can be declared and its methods can be utilized in a variety of scenarios.

VII.2.2.2 Sample Instantiation Code

The following code shows an example of how to use SEpT in a Java project. It also shows how to declare an MPG, initialize it, and manipulate it:

```
..
import SEpT.*;
..
public static void main(String[] args) {
MPG mpg=new MPG();
mpg.accessloginit(“logfile.log”);
            mpg.addnode("love","en");
            mpg.addnode("jordan","en");
            mpg.addnode("china","en");
            mpg.multilingualize();
            System.out.print( mpg.printGraphML1());
…
}
```

VII.2.3 **SEpT User Scenario**

VII.2.3.1 Graph Creator

The following UML sequence diagram (figure 66) shows a possible sequence of events to create an MPG using SEpT Java library.

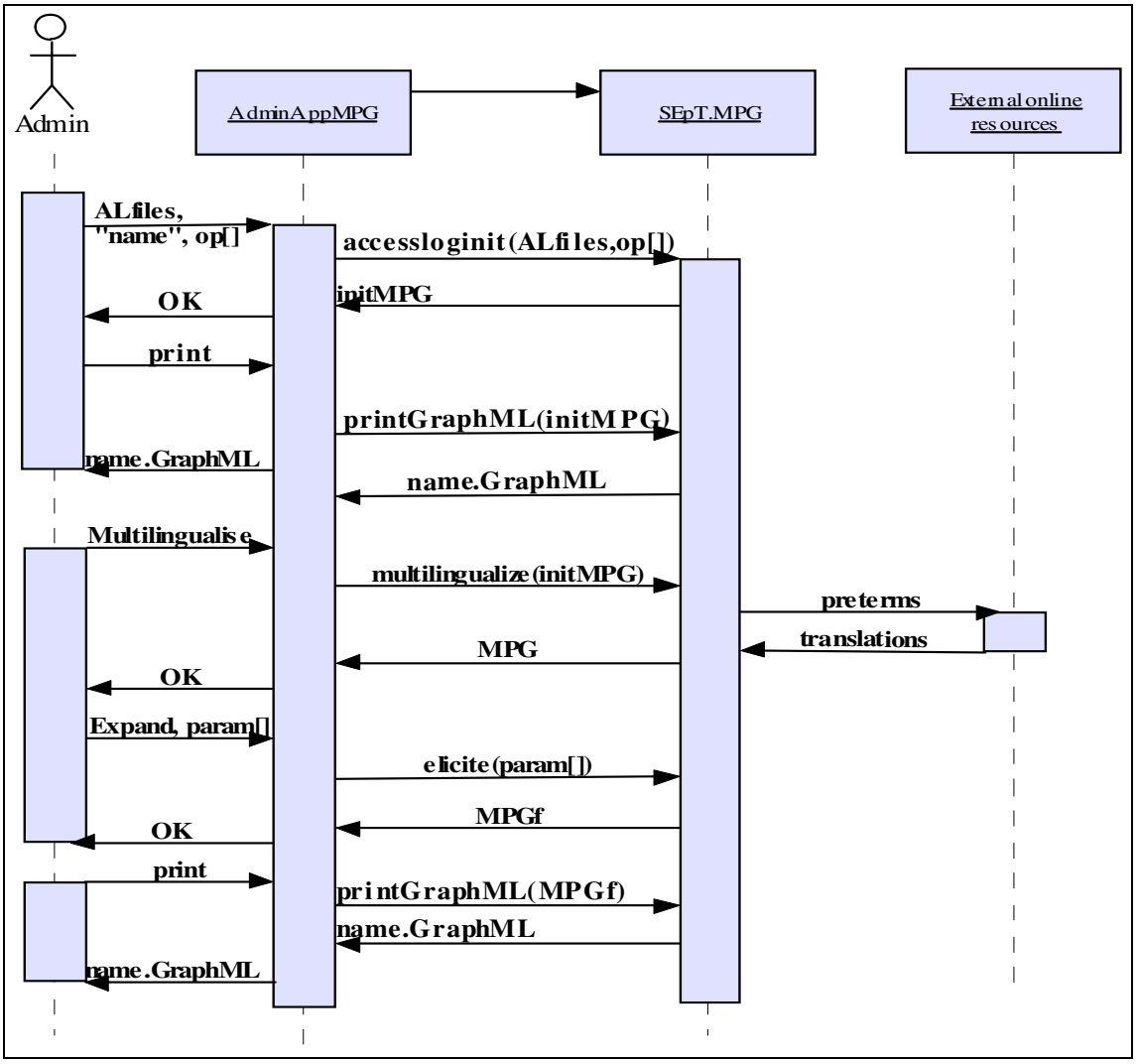


Fig. 66: Construction scenario

VII.2.3.2 Graph User

The following diagram, figure 67, shows a possible scenario of exploiting an MPG through an application.

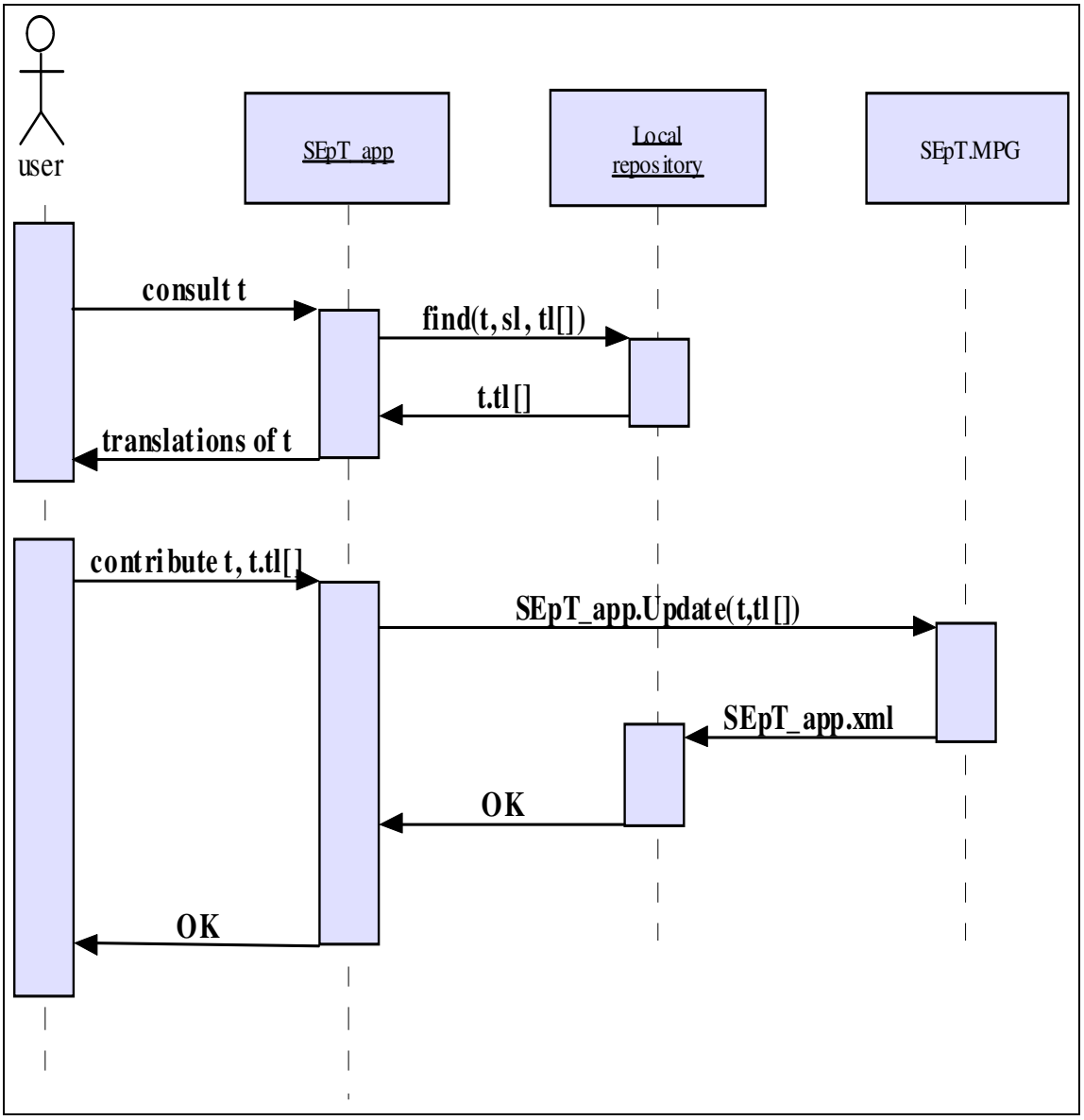


Fig. 67: Consultation and contribution scenario

VII.3 System Evaluation

After showing the design and implementation details of SEpT, this section proposes evaluation criteria to judge the performance and usability of SEpT. As an experiment, we show the results of producing 3 different MPGs to evaluate the performance of SEpT. As for the usability, the SEpT is evaluated on the basis of its functional requirements.

VII.3.1 Evaluation Criteria and Experiment Objective

VII.3.1.1 Performance

We evaluate the performance according to the following criterions:

1. the maximum size that of an MPG representable in SEpT;
2. the time necessary for building and searching an MPG of a given size.

In the case of preterminology, an MPG should be capable of representing a terminological sphere of a domain, which might contain hundreds of thousands of preterms and their relations.

As for processing time, if we have an MPG of P nodes (preterms) and R edges (relations), we require that any elementary operation (adding, modifying…) should be performed in O(log(P+R) time. We also want the waiting time for an elementary operation is less than 0.1 second when $P+R \leq 10^7$ on an average PC (2 GHz, 16Mb, 500Gb).

VII.3.1.2 Modularity,, Usability and Robustness

SEpT should automate the process of graph construction and do all the functional requirements in an easy way to be useful. Mainly, it should satisfy the following:

- Easiness of MPG construction.
- Correctness and usability of MPG manipulation.

It also should be capable of coping with new applications and environments to serve and exploit a variety of collaborative applications.

VII.3.2 **Experiment and Evaluation**

VII.3.2.1 Constructing DSR-MPG

To experiment the usability and performance of SEpT, especially in analyzing access log files, a DSR-MPG has been created using various resources. The purpose of this subsection is to evaluate SEpT as a system, not to show the detailed results of the produced MPG and evaluate them, which will be done in the next two chapters.

a. *Data*

- Access log files from December 2003 until January 2009. 17 GB.
- 50 books of 16,500 pages of digitized books.

b. *Process*

- Step 1: The access log files are filtered to eliminate automatic requests and extract only search requests.
- Step 2: The filtered logs are used by SEpT to initialize the graph.
- Step 3: The pages are used to extract further candidates.
- Step 4: The extracted candidates are merged with the graph.
- Step 5: The graph is multilingualized, using Wikipedia, Google MT and Google dictionary.
- Step 6: The graph is expanded.
- Step 7: The graph is exported into a GraphML format.

VII.3.2.2 Performance

The following table, table 9, shows the resources needed at each step in constructing the graph, the used memory estimated based on the number of nodes and edges on the graph, and based on the class templates. A node is estimated to consume 30bytes of memory, and an edge occupies 12 bytes. The table shows the total amount of memory after each step, and the required run time for each step.

Table 9: SEpT-DSR performance

	Memory (estimated)	Runtime	Associated elementary operation	Needed time for each elementary operation
Step 1	-	Filtering is done by external preprocessing tools (not fully automated)		
Step 2	89,000 * 30bytes + 254,000 * 12bytes = ~ (2.54 + 2.9) = ~5.44 MB	~1 hour, Inter Pentium (R) dual CPU T3400, 2.1GHz	-add nodes from log files	3600 / 89,000 = 0.04 seconds / node
			-add relations from log files	3600 / 25400 = 0.014 seconds / edge
Step 3	27,500 * 30bytes + 5.44 = 0.78 + 5.44 = 6.22 MB	5 hours, Inter Pentium (R) dual CPU T3400, 2.1GHz, fairly good internet connection	- call term extractor	(5 * 3600) / 27,500 = 0.65 second/node
Step 4	6.22 MB	Few minutes	- add nodes (directly)	(5 * 60) / 27500 = 0.011
Step 5	210,781 * 30 + 779,765 * 12 = 6.03 + 8.9 = 14.93 MB	7:30 hours, Inter Pentium (R) dual CPU T3400, 2.1GHz, fairly good Internet connection	- node multilingualization	27000 / 116500 = 0.23 second/node
Step 6	14.94 + 130,900 * 12 = 14.94 + 1.49 = 16.43 MB	2:30 hours, Inter Pentium (R) dual CPU T3400, 2.1GHz	Add edge by expansion	9000 / 130,900 = 0.06 node/edge
Step 7	16.43 MB	30min – 1 hour	Export node	3600 / 210,781 = 0.017 second/node
			Export edge	3600 / 910665 = 0.004 second/edge
Total	16.43 MB	16:30 ~ 17 hours		

The above table also shows an estimation of the required time to perform basic elementary operations on an MPG. Inserting a node to an MPG needs less than 0.011 seconds. The time may increase if the preterm of the node is being extracted from an external (text) or an internal (MPG expansion) resource. Creating an edge takes less that 0.1 second in all the cases.

VII.3.2.3 Usability Remarks

The system constructed the graph as required and it serves two applications: Contribute While Reading, and Contribute While Searching. To evaluate the usability of SEpT and its correctness, we must evaluate the resulting Graph and the applications. Such evaluation will be presented in the next chapters.

Assessment, Observations, and Conclusions

SEpT is implemented as a Java library that can be used with variety of consultation and contribution scenarios.

An instance of SEpT has successfully built a relatively large MPG, such as the one of the DSR, using a huge amount of data to be analyzed: 16,500 pages have been sent to a term extractor, and 17 gigabytes of access log files have been filtered and analyzed to construct the initial graph. All that has been done through SEpT automatically in less than 15 hours, which is reasonable considering the fact that the initialization is needed only once, and the initializer processed log files for more that 6 years. SEpT performs basic operations rapidly. Creating edges, adding nodes, exporting graphs takes less than 0.1 second per operation. Node multilingualization is slightly slower because it depends on external resources accessed by Internet.

The DSR instance of SEpT was robust enough to interact with various external systems (online MT systems, dictionaries and lexical resources) to serve its particular MPG.

The chosen computer representation of the graph as two columns of pointers to small nodes and edges was efficient, so the DSR-MPG occupied around 16 MB of the memory, even though it contained hundreds of thousands of nodes and edges, which is essential in processing large graphs.

Chapter VIII Passive Contribution Experiment

Introduction

This chapter describes two experiments in developing two different graphs for two different domains, where passive approaches were mainly involved. The result is evaluated to validate the concepts of preterminology, and MPG, with focus –in this chapter- on the results of the passive approaches.

We start by presenting the MPG dedicated to the Digital Silk Road, and give the experiment details and an evaluation. Then, in section VIII.2 MPG-Tafseer is presented. It is a graph dedicated to the domain of Arabic Oneirology. The graph has been developed using passive approaches, so the evaluation validates the produced results and the used approach. Finally, the third section provides observations on the experiments and gives assessments.

VIII.1 MPG-DSR

This section describes the experiment of constructing DSR-MPG using passive approaches, and evaluates the produced preterminology as a lexical resource.

VIII.1.1 Evaluation Criteria and Experiment Objective

VIII.1.1.1 Coverage

MPG-DSR will be evaluated according to its, lexical relational and linguistic coverage, and that will give an indication about the effectiveness of preterminology and particularly passive approaches.

a. *Lexical Coverage*

Lexical coverage deals with the amount of concepts that have a linguistic representation in each language L, hence, the lexical coverage ratio from language to language. To estimate the lexical coverage, we will use the "Informational Coverage Formula" proposed in section I.3.3.1. The formula calculates the ratio of entries in each language of a term base to the concepts of the domain. However, it is difficult to find I_s (number of concepts in the terminological sphere).

In this experiment we will use the method of (Etzioni, Reiter et al. 2007) in estimating I_s by selecting a random set of terms and assume that this set represents the ideal terminological sphere and its size it the size of such a sphere. Therefore:

$$IC = \left(\sum_{k} I_{TK}\right) / (N * I_{sr})$$

where IC is the informational coverage, K is a language in the term base, I_{tk}, is the number of entries in K, I_s is the number of entries in the random set, and N is the number of the languages in the term base.

b. *Linguistic Coverage*

Linguistic coverage deals with the amount of languages included in the repository that have enough lexical resources. We will use the formula presented in subsection I.3.2.1.

VIII.1.1.2 Precision and Recall

Coverage as a measure is similar to recall. Precision will be measured to see the inaccuracies that could have been introduced by preterminology and its resources.

In this experiment, precision=C/R, where C is the number of correct translation pairs and R is the number of retrieved translation pairs.

This criterion will also be used to show how the graph structure can be used to increase the recall without diminishing the precision.

VIII.1.1.3 Correctness

In the evaluation, incorrect correspondences will not be counted in the lexical coverage, and they will be reflected in the precision.

VIII.1.2 **Baseline of Terminological Resources for Cultural Heritage**

To evaluate preterminology, certain terminological repositories will be evaluated in convergence to the results of the lexical data extracted from MPG-DSR. This section presents these repositories.

VIII.1.2.1 Previous DSR Terminological Database

The previous DSR Terminological database was developed by professionals as a Japanese-based list of terms. It contains 825 Japanese terms. The following table, table 10, shows how many terms of the 825 are actually translated into different languages.

Table 10: Statistics of the previous term base of the DSR

Language	Number of terms	Percentage
Japanese	825	100%
English	445	54%
Chinese	421	51%
German	165	20%
French	113	14%
Russian	73	9%
Italian	65	8%
Swedish	65	8%
	Total: 2172 terms	

For example, 445 of the Japanese terms have been translated into English, while only 65 terms have been translated into Italian. It is important to compare preterminology to this database as it is very specialized in the domain of DSR.

Informational Coverage for this database is:

(IC) = (825+445+421+165+113+73+65+65)/(8*825) = 2172/6600 = 0.33

This is if we assumed that 825 is the number of concepts of the domain of the DSR, which is not true, as an index of a DSR digitized book may have more than 900 terms. Therefore the current database has low lexical coverage, if we considered the amount if available books. For example there are 14 French books, and the terminological database has only 113 French entries.

VIII.1.2.2 General Purpose Dictionaries

It is also important to compare preterminology to general-purpose dictionaries to test the coverage of MPGs, especially for hidden terminology. Here are the dictionaries we considered.

a. *Google Dictionary*

Google Dictionary is an online service of Google. Initially, it was not a separated service from Google Translate, but now it is a standalone online application.

Google Dictionary is a multilingual dictionary that serves 28 languages: English, French, German, Italian, Korean, Spanish, Russian, Chinese (Traditional), Chinese (Simplified), Portuguese, Hindi, Dutch, Finnish, Hebrew, Czech, Greek, Bulgarian, Croatian, Serbian, Bengali, Malayalam, Telugu, Tamil, Gujarati, Thai, Arabic, Kannada, Marathi.

The Arabic <-> English dictionary of Google is available online, and it can be called automatically. That is why we are using it in our experiment. Beside, it is supported by a Google, so it should have a reasonable size and quality.

b. *Alburaq En-Ar Dictionary*

Alburaq (PalTel 2010) is an English←→Arabic general purpose dictionary available online, it has at least 120,000 English-Arabic senses set. It was selected to evaluate the particular language pair en-ar in convergence to the produced bilingual preterminology, because of its size and popularity amongst other dictionaries.

For this keyword: "(قاموس عربي انجليزي): "qamous arabi injlizi", Arabic-English dictionary" Google ranks Alburaq dictionary first, placing it at the top of the result page, which means it is well-known, and used heavily by Arabic users.

VIII.1.2.3 PanImages

PanImages (Etzioni, Reiter et al. 2007) is a lexical repository compiled automatically from different wiktionaries. It offers lexical translations to images search requests in 100 languages. Visitors can also edit the translation of the search terms. It is important to compare preterminology with other automatic based systems.

VIII.1.3 **Data Used to Build the Graph**

VIII.1.3.1 Textual

50 books (16,500 pages) digitized using OCR technology, each book has been sent to Yahoo! Term to extract the preterminology candidates. The task requires finding the terms in each page, so there is a need to call an efficient term extractor multiple times automatically. Yahoo! Term is convenient as it has an API that allows a program to call it through an http request. We observed that calling Yahoo! Terms did not affect the runtime of the MPG initializer. Besides, it returns results that include multiword terms in an easy to process XML format.

VIII.1.3.2 Non-Textual

As mentioned earlier, access log files of the website of the DSR project from December 2003 until January 2009 were used to initialize the graph. The size of the logs is around 17 GB divided in 1863 files. We are interested in requests that have search terms; here is an example of such requests (these are simplified lines of an access log files that has the IP address and the requested page):

113.213.165.220 GET /cgi-bin/toyobunko/geta_search.pl?sn=all&lang=en&input=天水

116.112.32.132 GET /cgi-bin/toyobunko/geta_search.pl?lang=en&input=**baotou**

116.112.32.132 GET /cgi-bin/toyobunko/geta_search.pl?sn=all&lang=en&input=**包頭**

121.64.205.3 GET /cgi-bin/toyobunko/geta_search.pl?lang=ja&sn=all&input=**ten king**

121.64.205.3 GET /cgi-bin/toyobunko/geta_search.pl?lang=ja&sn=all&input=**七經**

121.64.205.3 GET /cgi-bin/toyobunko/geta_search.pl?lang=ja&sn=all&input=**十王**

VIII.1.3.3 Lexical

The following repositories were used:

- Google translate
- Wikipedia
- IATE

VIII.1.4 **Results**

The initial graph after normalization contained 89,076 nodes. Also 81,204 English preterms candidates have been extracted using Yahoo! Terms. Note that only English books were used, so we assumed they are English preterms. 27,500 of them were not discovered from the access files. Hence, the total number of nodes in the initial graph was 116,576, see figure 68 for sample nodes.

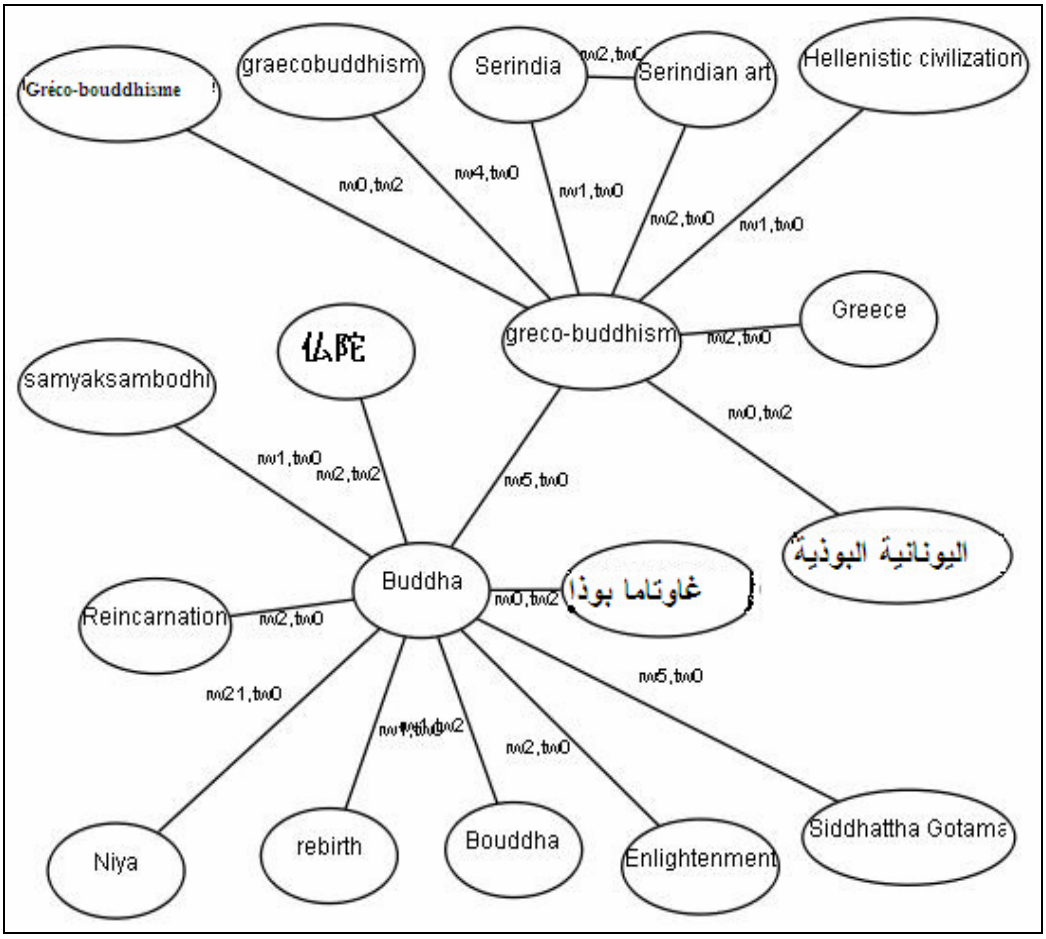


Fig. 68: Sample graph

After multilingualization, the graph had 210,781 nodes containing terms from the most important languages (en, fr, de, jp, zh, pt, it, ru, ar, sv, po, ko…). The graph has now 779,765 edges with $tw > 0$. The important languages are the languages of the majority of the visitors, the languages of the archived books, and representative languages a long the Silk Road (hi, pe, tu, vi, th, pe, uz). DSR-MPG achieved high linguistic coverage compared to the current terminological database as 30 languages have more than 500 nodes on the graph.

If we assumed that the targeted languages are the languages of the DSR books (en, jp, fr, ar, it, du, sv, ru, ch), DSR-MPG would achieve 9/9, assuming 200 is the least number of entries, while the current database achieves 3/9.

Around 55,200 root English terms were used as a seed set of terms; these terms were selected from the initial DSR-MPG. Around 35000 terms were translated from Wikipedia into at least one language, mostly in French and German. Wikipedia increased the number of more confirmed relations in the graph by introducing around 113,000 edges (with *tw*), figure 69. However, MT increased coverage (figure 70).

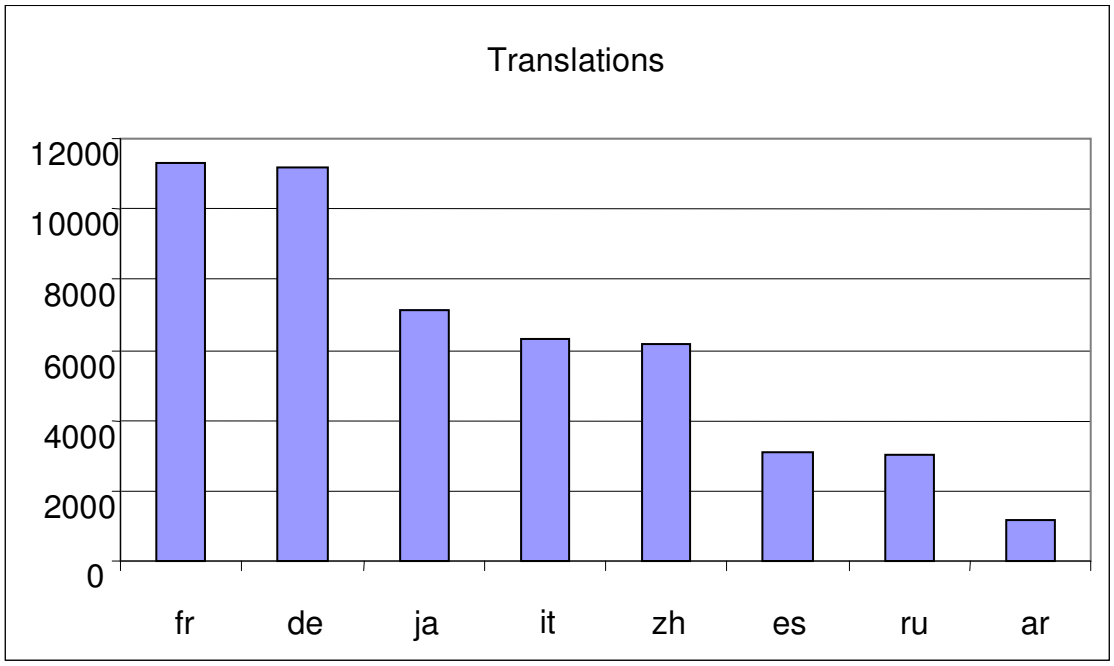


Fig. 69: Number of translated terms in sample languages using Wikipedia

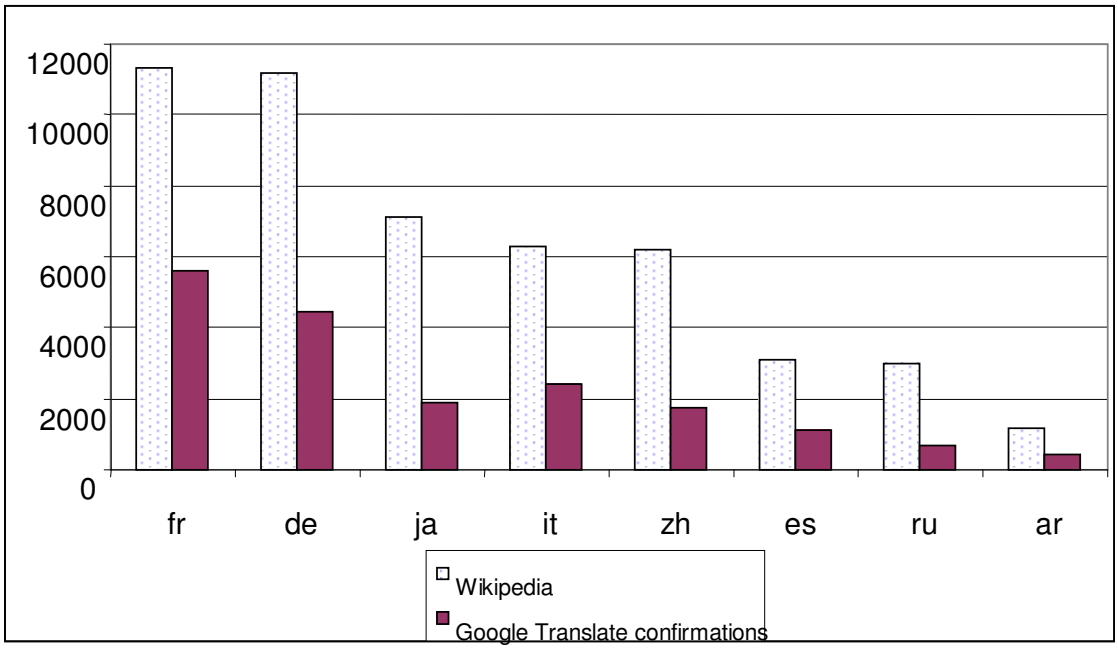


Fig. 70: Terms translated by Google MT and matching the translation of Wikipedia

VIII.1.5 **Evaluation**

VIII.1.5.1 Coverage

a. *Linguistic Coverage*

DSR-MPG achieved high linguistic coverage as 20 languages have more than 1000 nodes on the graph. While the DSR terminological database has only 8 languages, all of them have less than 825 preterms. As we have shown above, if we considered the targeted languages set to be the languages of the archived DSR books, DSR-MPG would achieve 100% linguistic coverage, while the terminological database achieves 33%.

Figure 71 compares between the linguistic coverage (L/N) of DSR-MPG and the current database, assuming that N is 8 (8 languages of the current database). K is a constant, when K=10, means that any language that has entries more than 10 will be counted in L.

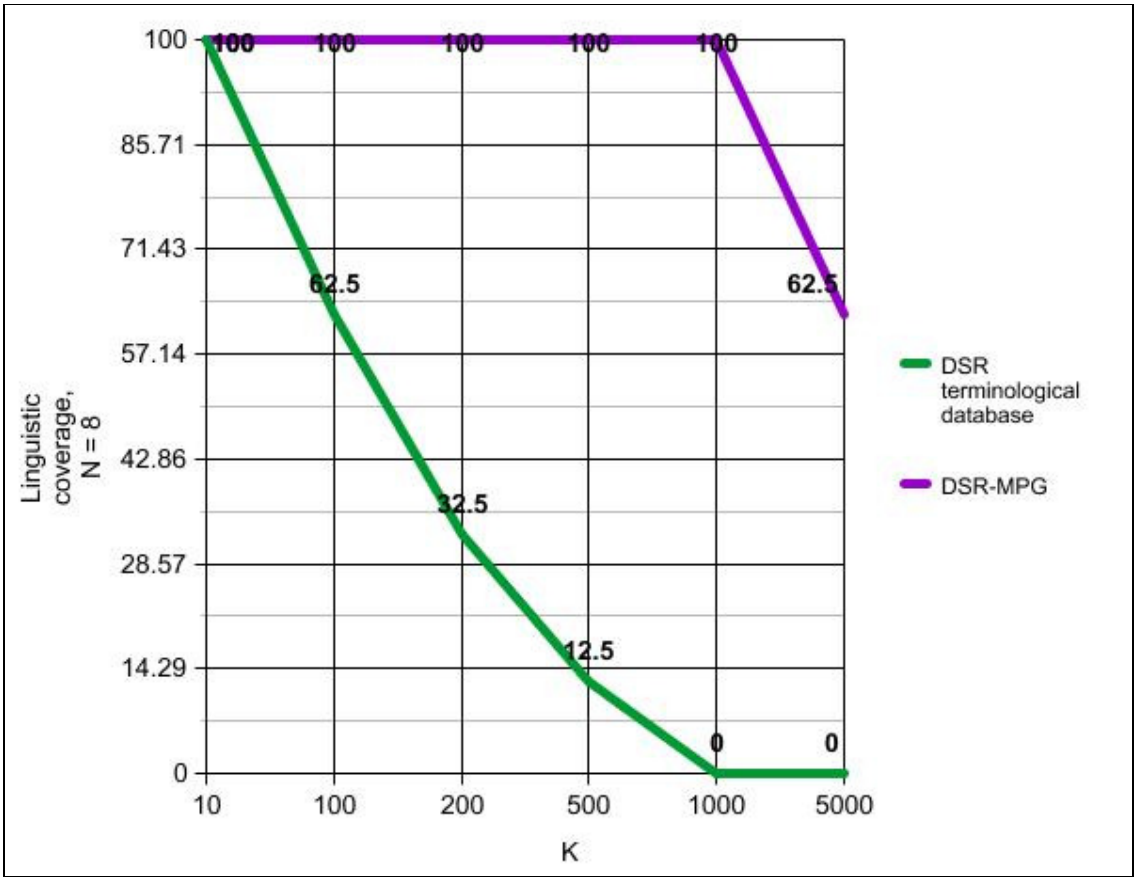


Fig. 71: Comparison between the linguistic coverage (L/N) of DSR-MPG and the current database

DSR-MPG achieved high linguistic coverage even when K=500. While the current terminological database has a linguistic coverage of 12.5% when K=500.

b. *Lexical Coverage*

To evaluate the lexical coverage of the produced graph, 350 English concepts have been extracted manually from the terms at the index pages of the following books (note that such terms are not necessarily included in the original terminological database of the DSR):

- Ancient Khotan, vol.1: http://dsr.nii.ac.jp/toyobunko/VIII-5-B2-7/V-1/
- On Ancient Central-Asian Tracks, vol.1:http://dsr.nii.ac.jp/toyobunko/VIII-5-B2-19/V-1
- Memoir on Maps of Chinese Turkistan and Kansu, vol.1: http://dsr.nii.ac.jp/toyobunko/VIII-5-B2-11/V-1

These terms are related to the DSR and it is essential to cover them. Besides, these sources might contain large amount of unrecognized or latent terminology, as they were digitized recently by OCR and their content are not directly used to build a terminological database.

Hence, we tried to translate them into Arabic and French. Figure 72 compares DSR-MPG and various general purpose dictionaries. Out of the 350 terms, we found 189 correct direct translations into Arabic. However, their number reached 214 using indirect translations. On the other hand, the closest to DSR-MPG was PanImages, which uses Wiktionaries and various dictionaries, with only 83 correct translations.

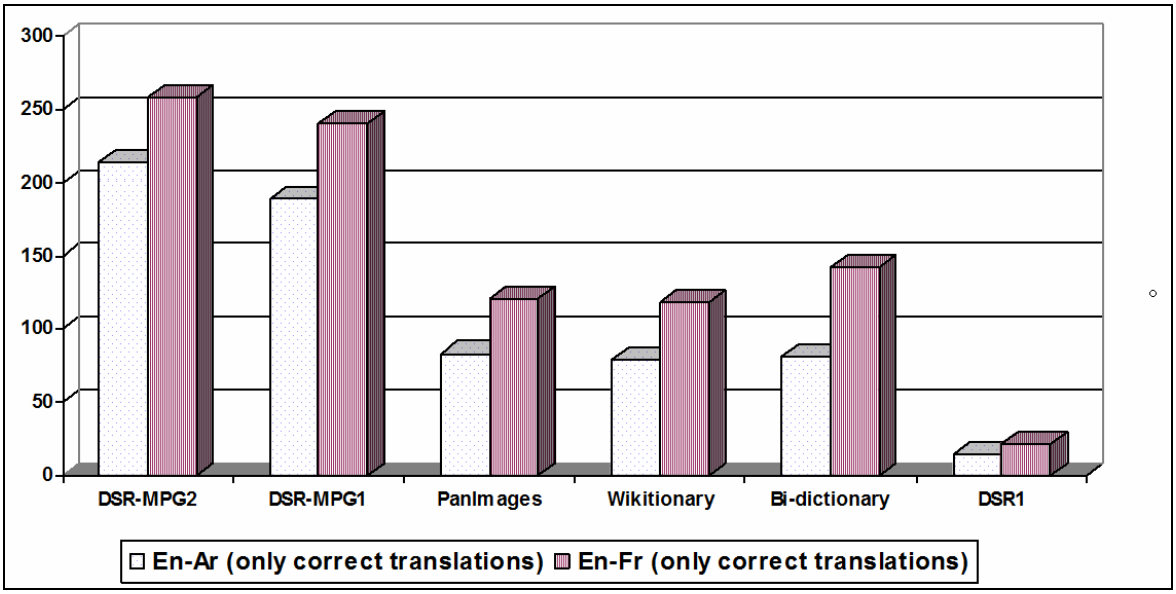


Fig. 72: A comparison between DSR-MPG, and other dictionaries, including (Babylon 2009)

DSR-MPG1 refers to the translations obtained from formula 1, DSR-MPG2 to the translations obtained from indirect translations, which increased the amount of correct translation by 25 terms in the case of En-Ar.

VIII.1.5.2 Correctness (Precision and recall)

MPG2 was built to expand the graph and use its structure to find new relations; this expansion actually increased the coverage (recall) but decreased the precision.

In MPG1, 195 preterms were translated into Arabic, out of 350. 189 out of 195 gave correct translations, and 6 were inaccurate. Hence, the precision was 195/189 = 96.9, and the recall was = 189/350= 54%.

MPG1 was expanded using formula 1 and new translational correspondences were built. The resulting MPG2 achieved better coverage by increasing the recall by 7%. A few inaccurate translations were introduced, which decreased the precision to 92.4%. As the main purpose of preterminology is coverage, we can conclude that MPG2 and its structure successfully increased the coverage, with an insignificant amount of inaccuracies. See table 11.

Table 11: MPG1 vs. MPG2

	MPG1	**MPG2**
Precision	96.9%	92.4%
Recall	54%	61%

VIII.2 MPG-Tafseer

To show the usefulness of SEpT and MPG; for languages with poor lexical resources like Arabic, an instance has been dedicated to construct preterminology for the domain of Arabic oneirology. This section describes the overall service and evaluates the experiment of constructing an MPG for oneirology, called MPG-Tafseer.

VIII.2.1 MPG-Tafseer Based Arabic Dream's Interpretation Service

Through the Islamic history, many scholars studied the interpretation of dreams, believing that an event in a dream reveals a truth or corresponds to a truth (Ostransky 2005). These scholars published many books for interpreting dreams. Muhammad Ibn Sireen (who lived in the 8th century) is a famous Islamic scholar in that domain (Wikipedia-B 2010); he produced many publications (in Arabic) for dream interpretation.

The organization of these books is based on subjects of dreams, or follows the alphabetical order of the main topic of the dream. For example, Ibn Sireen's book (Gouda 1991) (Ostransky 2005) has 60 chapters (doors), where each chapter gives interpretations about a specific subject like "interpreting dreams about trees and their fruits". Each chapter is divided into sections, like "dreams about grape trees". Each section contains several stories and each story describes an interpretation of a situation in a dream. For example: if one dreamed about a "birds nest", one will find the interpretation in the chapter of birds, and according to Ibn Sireen, "dreaming about birds' nest means you have/will have a happy marriage".

Nowadays, these books are available in printed form, and digital form. Aljayyash.net offers the main books online (Aljayyash 2010). Despite the fact that many question the credibility of these interpretations, these books are used by many Arabic speakers from different backgrounds. The process of accessing such interpretations is still an issue, mainly because of the following reasons.

- The typology of dreams is similar to the orthodox theological categories, and that makes it difficult for modern readers to understand.
- The lexical gap between the language of the books (usually written between the 8th till the 13th centuries) and the vocabulary of a modern Arabic speaker (Diab and Habash 2009) is wide.
- The typical challenges of an information retrieval system for Arabic text: morphology, lexical inconsistency, ambiguity… (This problem is not addressed here).

A lexical resource that bridges the gap between the author and the reader (user) is necessary. Concepts have many lexical representations in Arabic; and the old books often do not use the same vocabulary as the one used by modern readers. For example, all of the following, "mal", "nuqoud", "tharwah", "masari", and "fulous" may mean money, but in the books of Ibn Sireen, there are tens of stories for "mal" and "noqoud", and no entries for the last two words. On the other hand, a modern user will use "masari" or "fulous" (both of them are Arabic words that are adopted by the dialectal Arabic). Hence, if someone had a dream about money, it makes sense to

give him results about "mal" and "nuqoud", even though he searched for "masari". A terminological resource that matches the two lexical representations is needed. That is why we built a preterminology of Arabic oneirology. It is a repository of monolingual and bilingual synonyms, helps bridge that gap, and serves in an IR system based on query expansion model. (See figure 73).

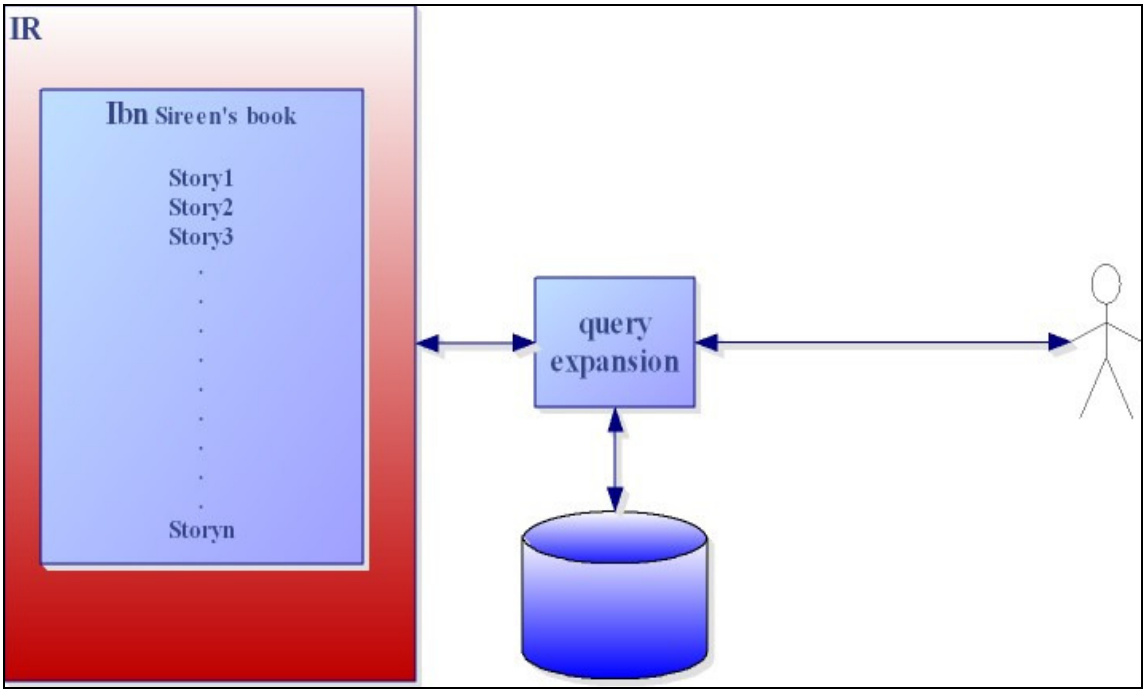


Fig. 73: Query expansion

VIII.2.2 **Building MPG-Tafseer**

a. *Initialization*

The graph was initialized using the set of terms extracted semi-automatically from the book of dream interpretations of Ibn Seerin. No links were established at this step. (See figure 74).

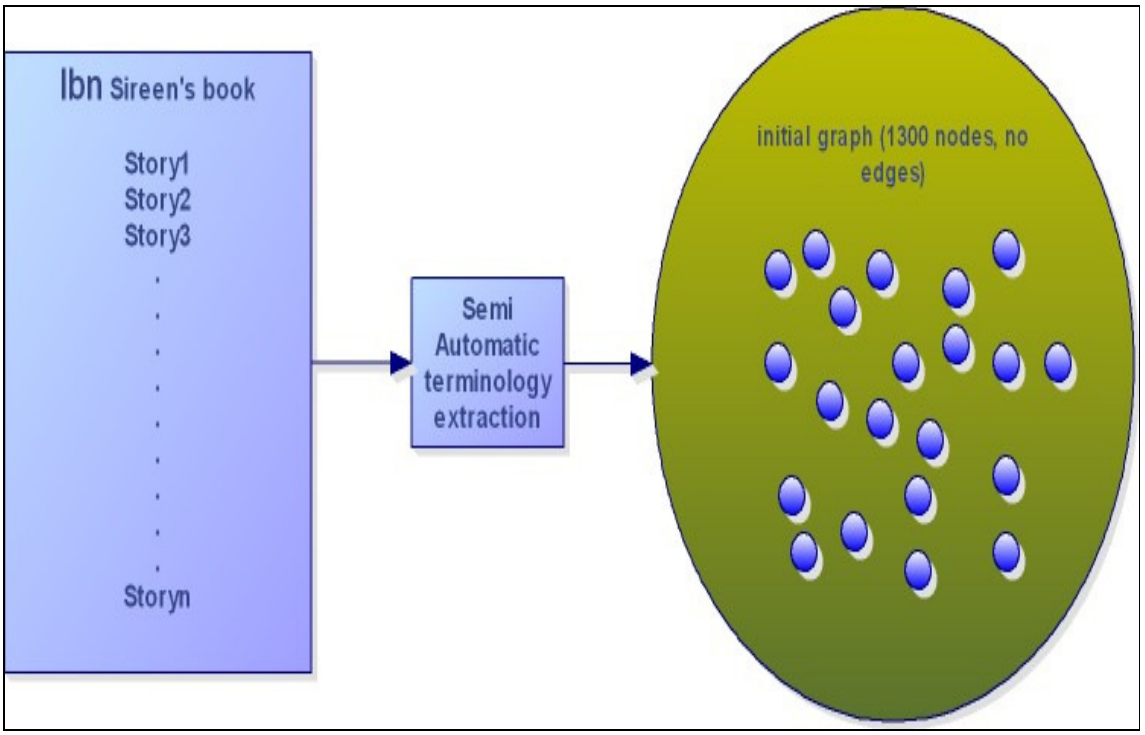


Fig. 74: MPG-Tafseer initialization

b. Graph Multilingualization

The objective here were not to translate the terms only, but to use the Arabic lexical resources available in Arabic - English dictionaries.

We used Alburaq (Arabic←→English dictionary) to translate each term in the initial graph. A translation weight was assigned between the term and its English translation.

Then, the nodes have been translated from English and French terms back into Arabic, in order to find synonyms of the original preterm (through Alburaq, Wikipedia, and Google Translate again, to enrich the Arabic content).

For example, assume that the original preterm is: ("غضنفر"ghadanfar) which means lion. At the beginning, we translate it into English "lion", and then if we translate it into Arabic again from various resources and get more synonyms (recognized by the graph expander as they share the English translation). In our example: "غضنفر ghadanfer" →*ar to en*→ lion → *en to ar*→ "asad, اسد", "haidar, حيدر", "qasham, قشعم", "reabal, رئبال".

At this stage, we have a graph of 40,200 nodes, with only translation weights on edges. Figure 75 shows the multilingualization process.

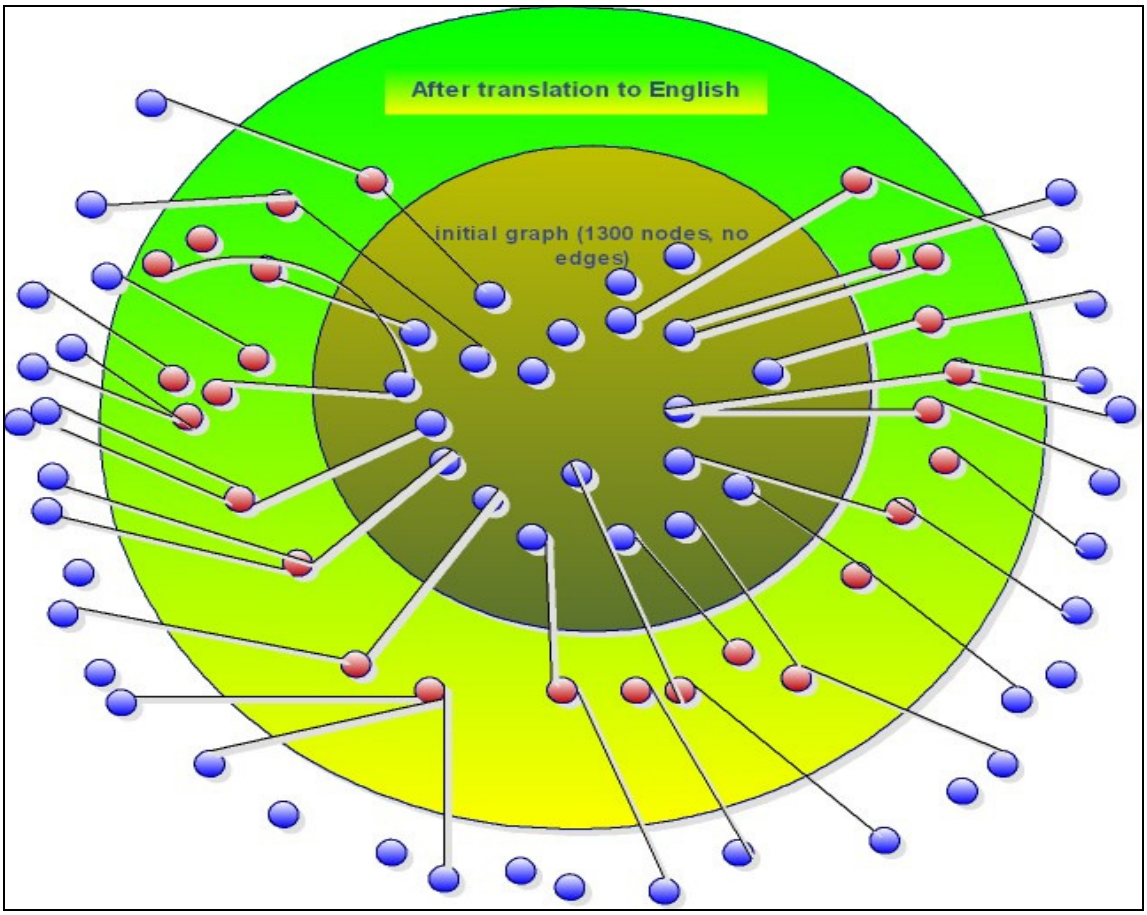


Fig. 75: MPG-Tafseer, Multilingualization

c. *Expansion*

Formula 1 of (Daoud, Boitet et al. 2009) was used to find the synonyms, based on the common translations. For example, if the translation weight between “nuqoud” and “money” is high, and that between “fulous” and “money” is also high, then the possibility that “nuqoud” and “fulous” are synonyms is also high.

Figure 76 shows the process of finding synonyms. Based on that, we construct “heuristic” edges between the nodes and the translations of their synonyms to expand the graph more, as shown in figure 74.

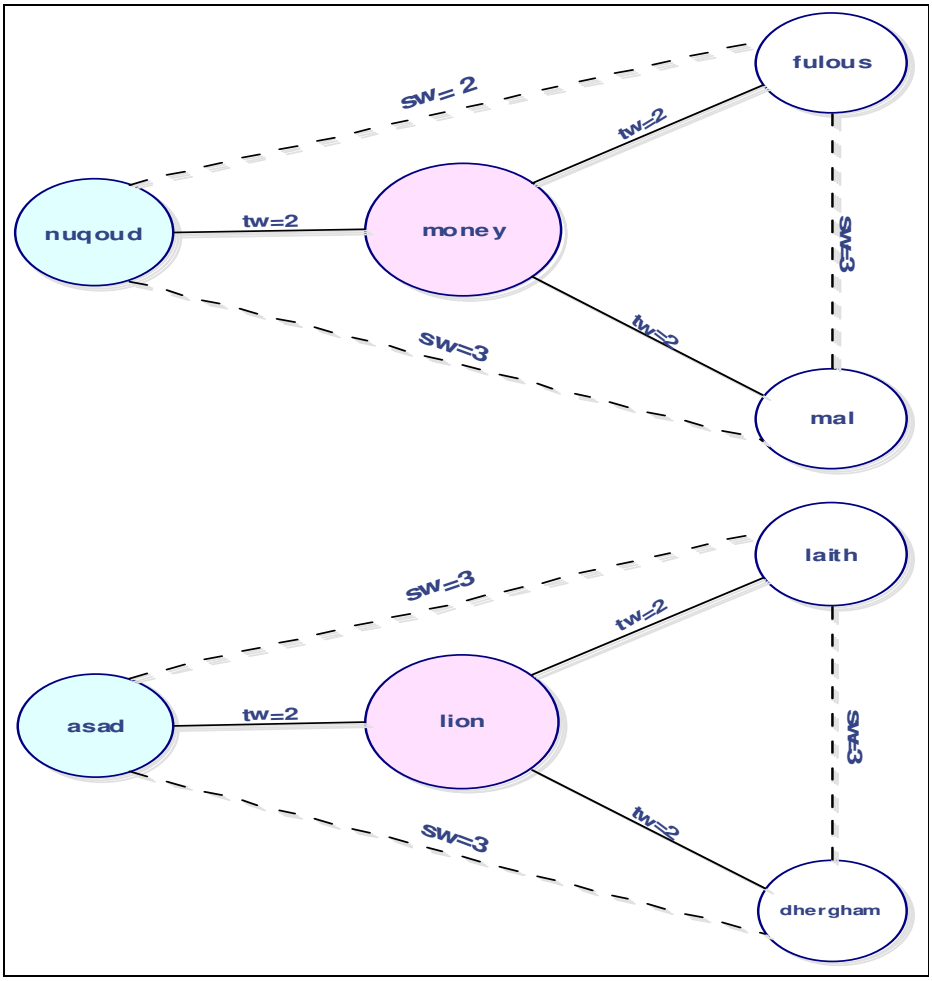


Fig. 76: Expansion

d. *Extraction*

The focus is on the Arabic synonyms of the MPG, hence we extracted the preterms connected to edges with synonym weights and their neighboring Arabic terms, as figure 77 shows.

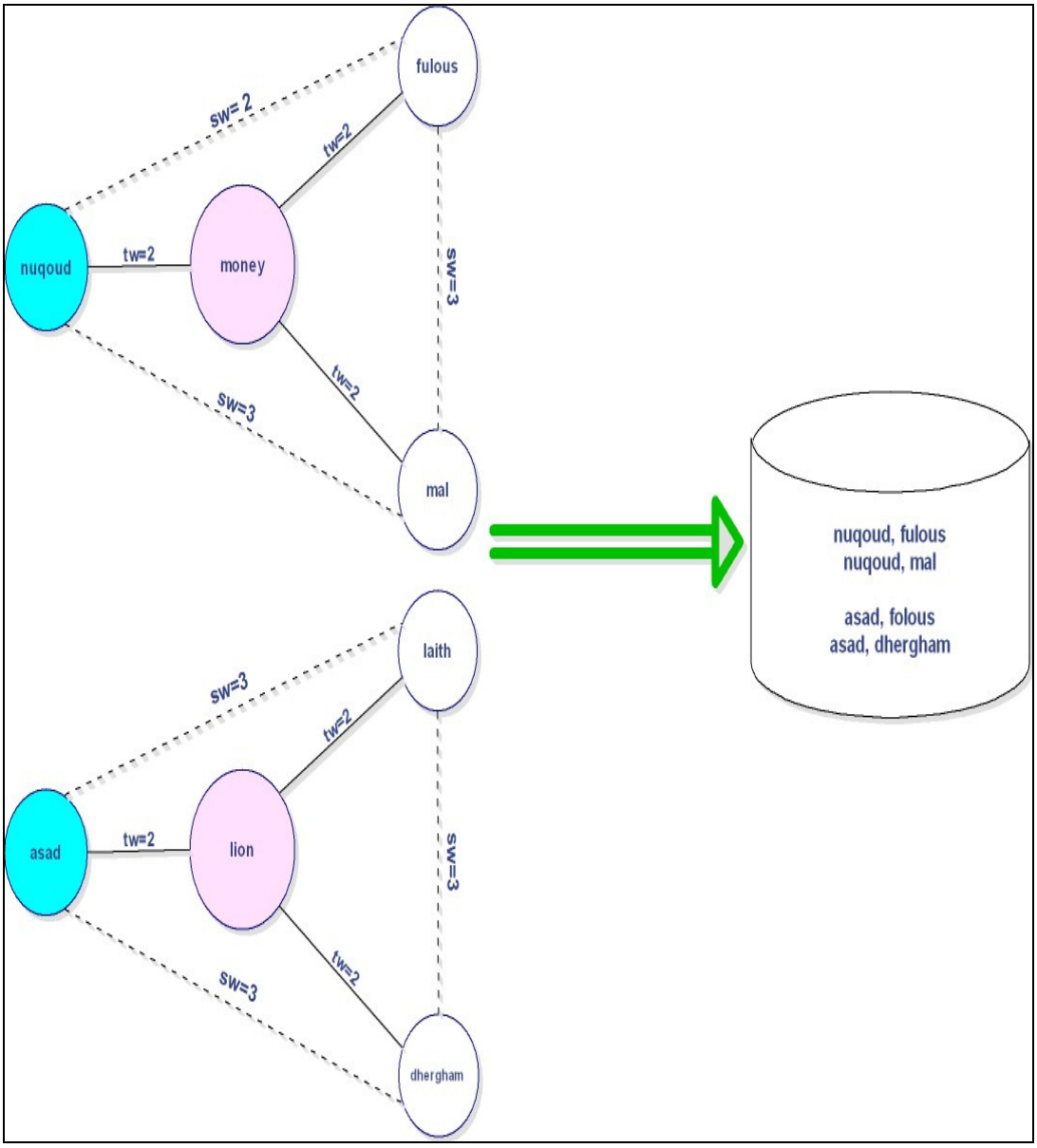


Fig. 77: Lexical knowledge extraction

VIII.2.3 **Results**

We used 1300 Arabic terms to initialize the MPG and an Arabic → English dictionary to translate these terms into English. Then we translated the English terms back into Arabic using the English → Arabic version of the dictionary. This process expanded the graph to 22,600 nodes, half of them having Arabic terms.

To have more variations, we translated the English terms into Arabic and French using Google Dictionary and Wikipedia, which is constructed by online volunteers who may be experts in the domain. This process expanded the graph 40,200 terms.

At the latest stage we used formula (1) to find more synonym edges. As the terms appearing in the books are crucially important in the process of analyzing the text, we extracted the list of Arabic terms based on the English one.

The data of the MPG is currently used to serve an operational system for automatic interpretation of dreams available at: http://www.maherinfo.com/, developed by Daoud Maher Daoud.

In the following screenshot (figure 78), the user was looking for “fo’ad” (heart), and did not find any result for this word, but it retrieved results for “qalb” (heart).

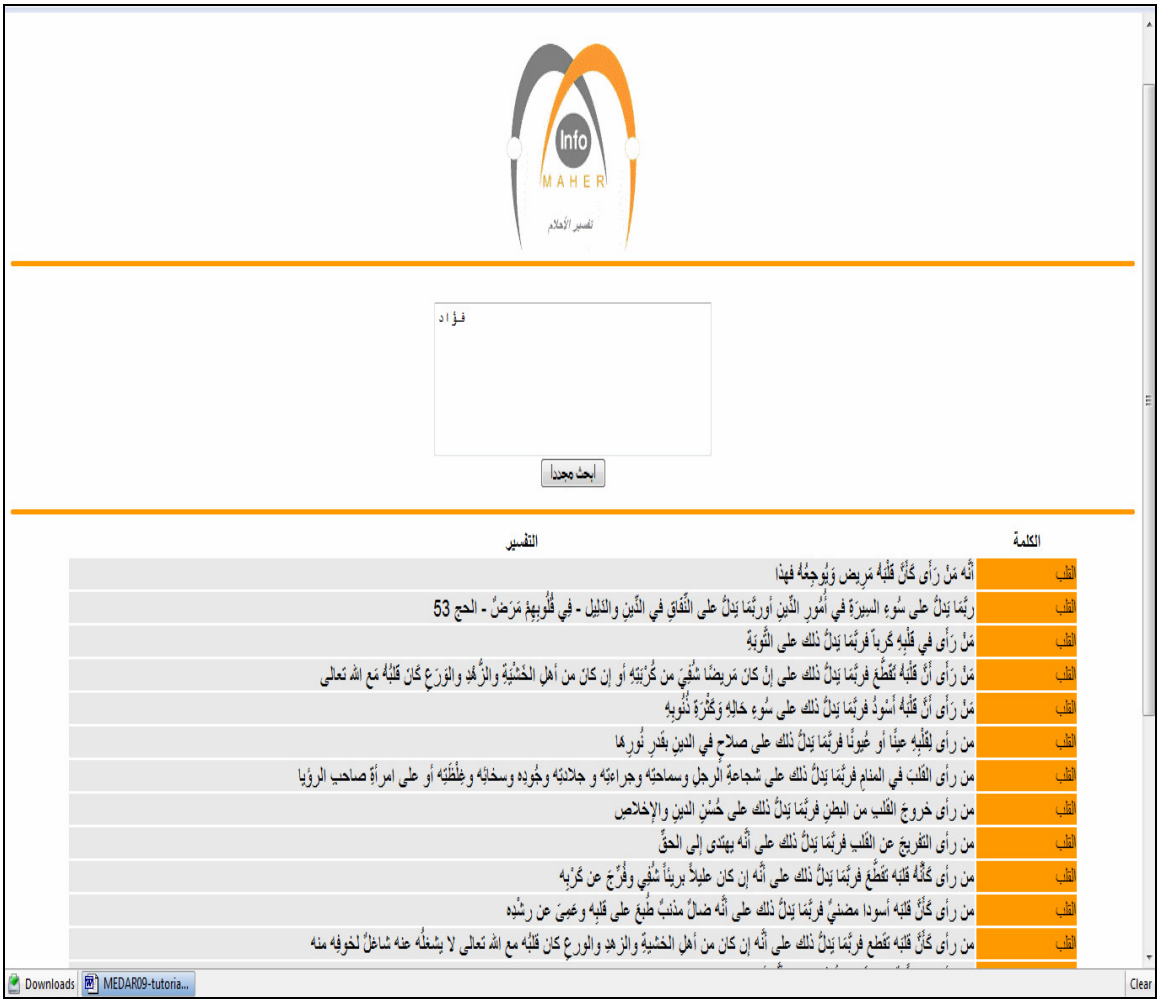


Fig. 78: Dream interpretation service

VIII.2.4 **Evaluation**

VIII.2.4.1 Evaluation criteria

The usefulness of preterminology will be evaluated by evaluating the service of dream interpretation that uses our MPG. As the MPG is expected to contain more hidden terminology that matches the terminology used by average users, we will evaluate the precision and recall of the system (IR model).

To conduct this evaluation, we fetched the access log files of maherinfo.com to see the received requests, we selected all requests sent during a particular week, which has 50 search terms (after eliminating repetitions).

VIII.2.4.2 Baseline

a. *Aljayyash.net*

Aljayyash compiles various dream interpretation books and offers a simple search service, using an external search engine. The user sends a term, and the system retrieves all the dreams that have the requested term.

b. *Alburaq.net*

Alburaq.net has a dream interpretation service, available freely online. It offers the service by making exact matching between the indexed text and the request.

VIII.2.4.3 MPG-Tafseer Evaluation

The three systems use the same books for interpretation; an interpretation is an explanation of a particular dream. MPG-Tafseer scored a higher recall than the other two systems as it gave 165 relevant interpretations to the 50 requests using the query expansion through the MPG, 11 only had 0 results.

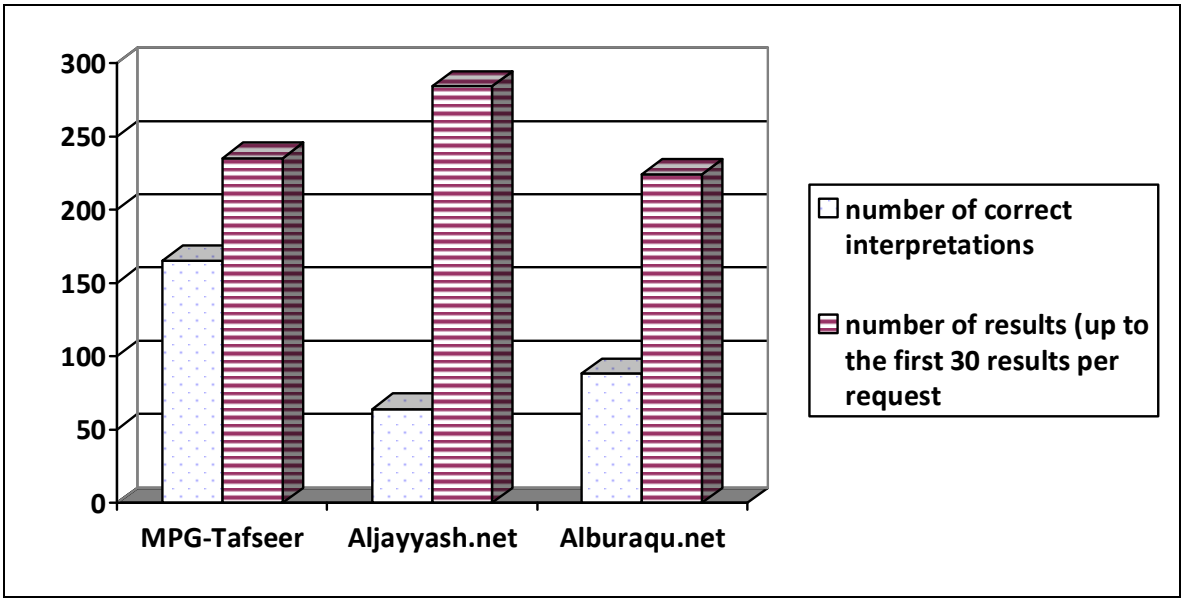


Fig. 79: Comparison between MPG-Tafseer, and Aljayyash.net and Alburaq.net

As shown in figure 79, MPG-Tafseer improved the precision significantly, as it scored 70%, while the closest (al Buraq) has 39%.

VIII.3 Assessments and Observations

Two experiments on two graphs constructed using passive approaches have been shown in this chapter.

DSR-MPG achieved better lexical and linguistic coverage than the current terminological database built by professionals and even other general purpose dictionaries. One essential aspect about MPG-DSR is the utilization of access log files, which contain digital (non-textual) content that is used by the community but not recognized by lexical or terminological repositories. That experiment shows that passive approaches can be effective in resolving latent terminology.

Tafseer-MPG built an Arabic thesaurus-like preterminology for the domain of dreams interpretation. Using the graph structure and indirect edge constructions, the graph was expanded and many synonyms were found. The produced preterminology was helpful in improving the service and achieving better recall and precision.

Chapter IX Active Contribution Experiments

Introduction

This chapter describes two active contribution experiments. The first one is done on the Arabic-based JeuxDeMots, and the second one was the contribution gateway of the DSR. This chapter shows that it is indeed possible to elicit contributions through a non linguistic activity.

IX.1 Arabic-Based MPG through JDM

IX.1.1 Objectives and Overview

This experiment shows the results of the constructed MPG through the indirect active contributions gathered from the Arabic version of JeuxDeMots.

This version has been launched in January 2010 to entertain an average Arabic audience. This experiment focuses on the potential of indirect active contributions to preterminology and the quality and coverage of the contributed material, taking into consideration the amount of time needed to collect such contributions.

IX.1.2 Evaluation Criteria

IX.1.2.1 Usability

A usable serious game (a game with a purpose) entertains online players, and achieves its purpose of building a special kind of knowledge.

This section will investigate the potentials of an MPG-based indirect active contribution by evaluating the amount of participation and contribution to the game, and the growth of the number of registered players through the 9 months the game was tested.

IX.1.2.2 Contributed Material

The second evaluation criterion deals with the contributed data itself, in terms of (1) the correctness of the contributed correspondences and lexical units, and (2) the coverage of absent terminology (to prove the hypothesis that active contribution can be effective in resolving the terminological gap).

For measuring the correctness, it is easy to test a set of lexical units, but for the coverage, it is difficult to consider JDM-MPG as a complete repository and compare it to other mature systems. However, it is possible to test if it contains preterms and correspondences that are absent in other repositories, and to evaluate its growth rate. That will give us an idea of the characteristics of the produced preterminology.

IX.1.3 Baseline

IX.1.3.1 General Purpose Arabic Lexicons

To evaluate the uniqueness of the contributed data, general purpose dictionaries like Alburaq and Google dictionaries will be used.

IX.1.3.2 Arabic WordNet

This project tries to construct an Arabic WordNet (AWN 2010), following the development process of Princeton's WordNet and Euro WordNet. It utilizes the Suggested Upper Merged Ontology (SUMO) (SUMO 2010) as an interlingua to link the Arabic WordNet to previously developed WordNets.

IX.1.4 **Results**

IX.1.4.1 Statistics

The game was seeded with around 750 Arabic terms, selected from Alburaq.net.

Figure 80 shows the growth of number of registered players starting from January. Note that online visitors can also play without registration as guests. So far we have around 55 players.

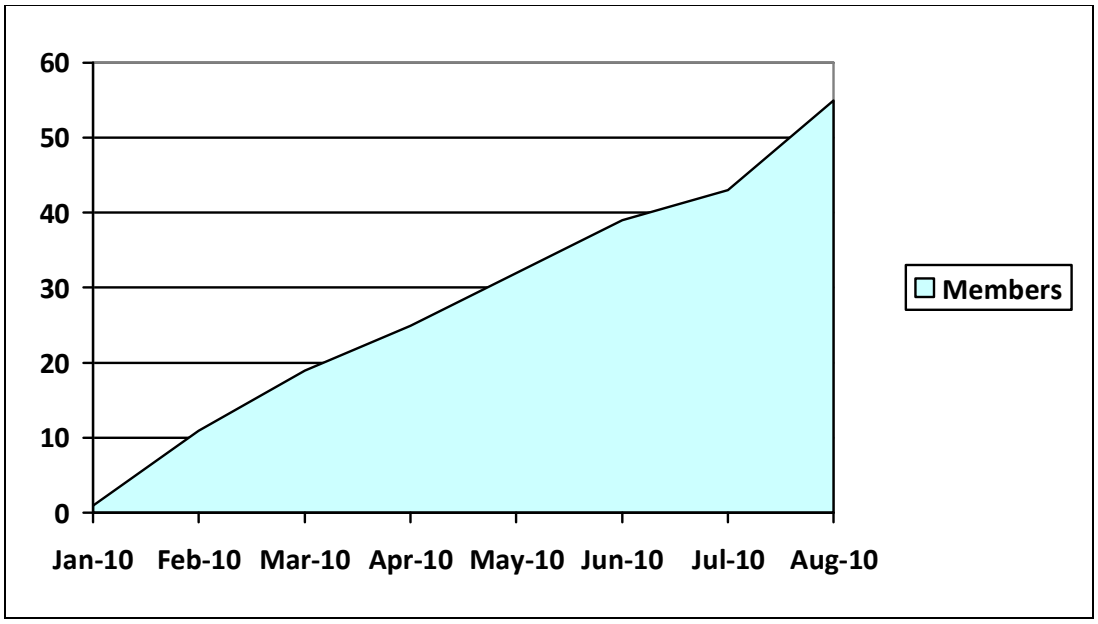


Fig. 80: Number f registered players at the Arabic JeuxDeMots each month

Figure 81 shows the growth of the number of terms contributed during the 8 months period starting from January 1st.

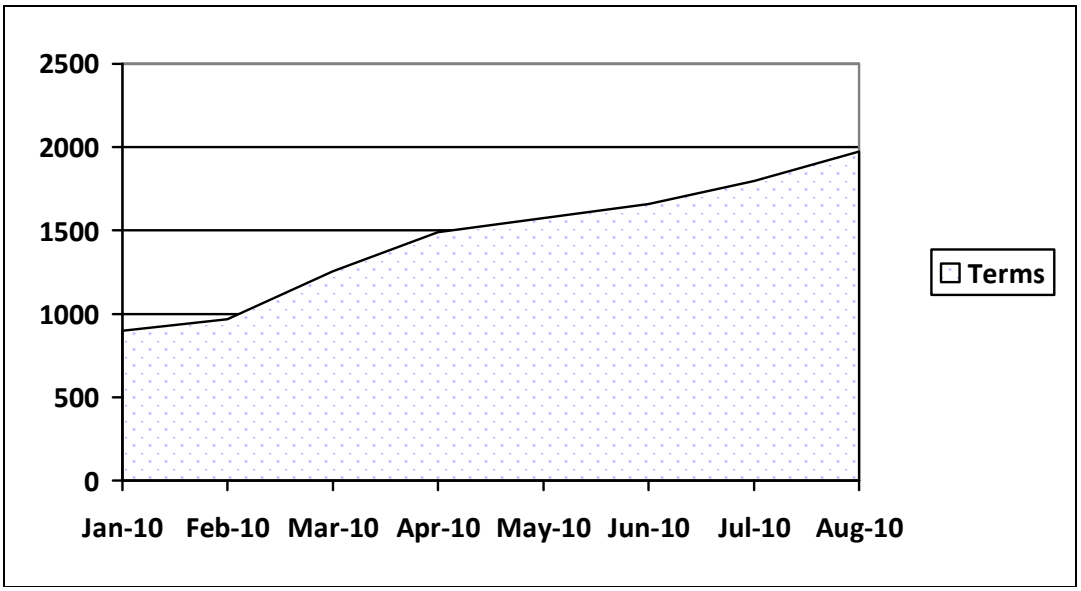


Fig. 81: Cumulative number of contributed terms each month

Figure 82 shows the growth of the number of relations added to the graph during the same 8 months period.

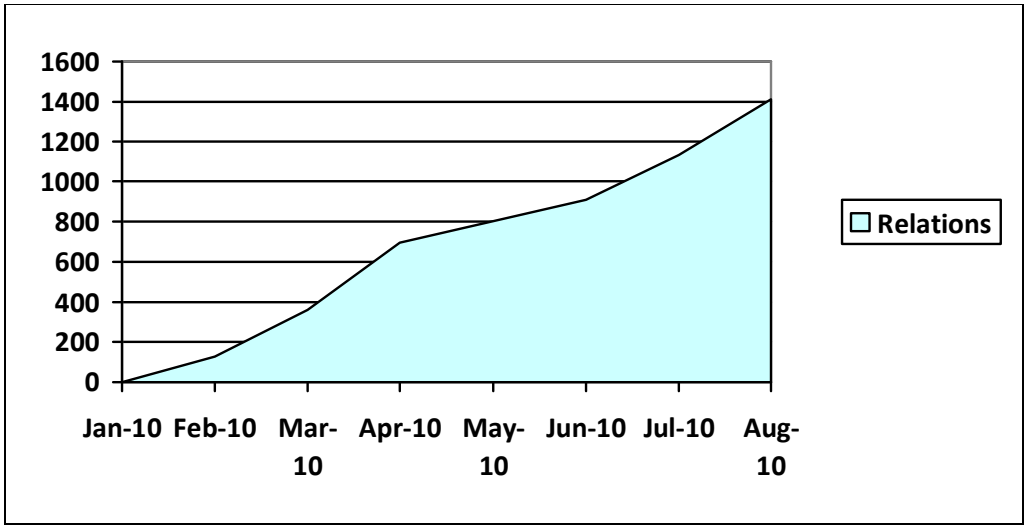


Fig. 82: Cumulative number of contributed relations

IX.1.4.2 Sample Graph

The following is a sample sub-graph from JDMAR-MPG. It started by the term “tabeeb” which means “doctor of medicine” (figure 83). The corresponding JeuxDeMots sub-graph is shown on figure 84.

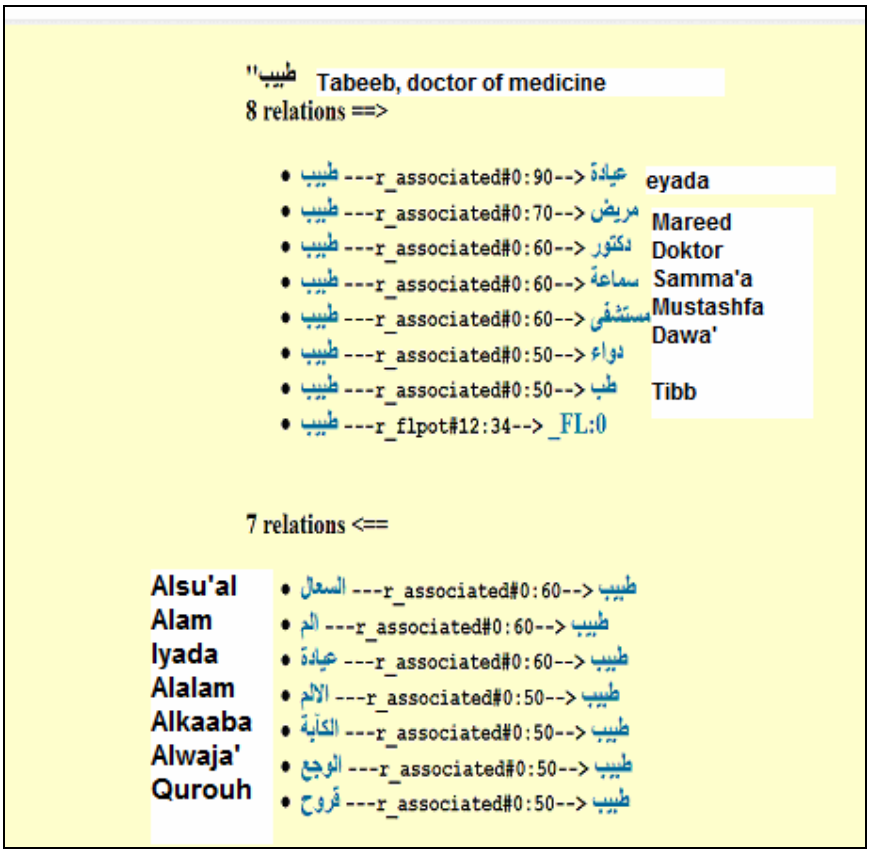


Fig. 83: Sample JDMAR sub-graph

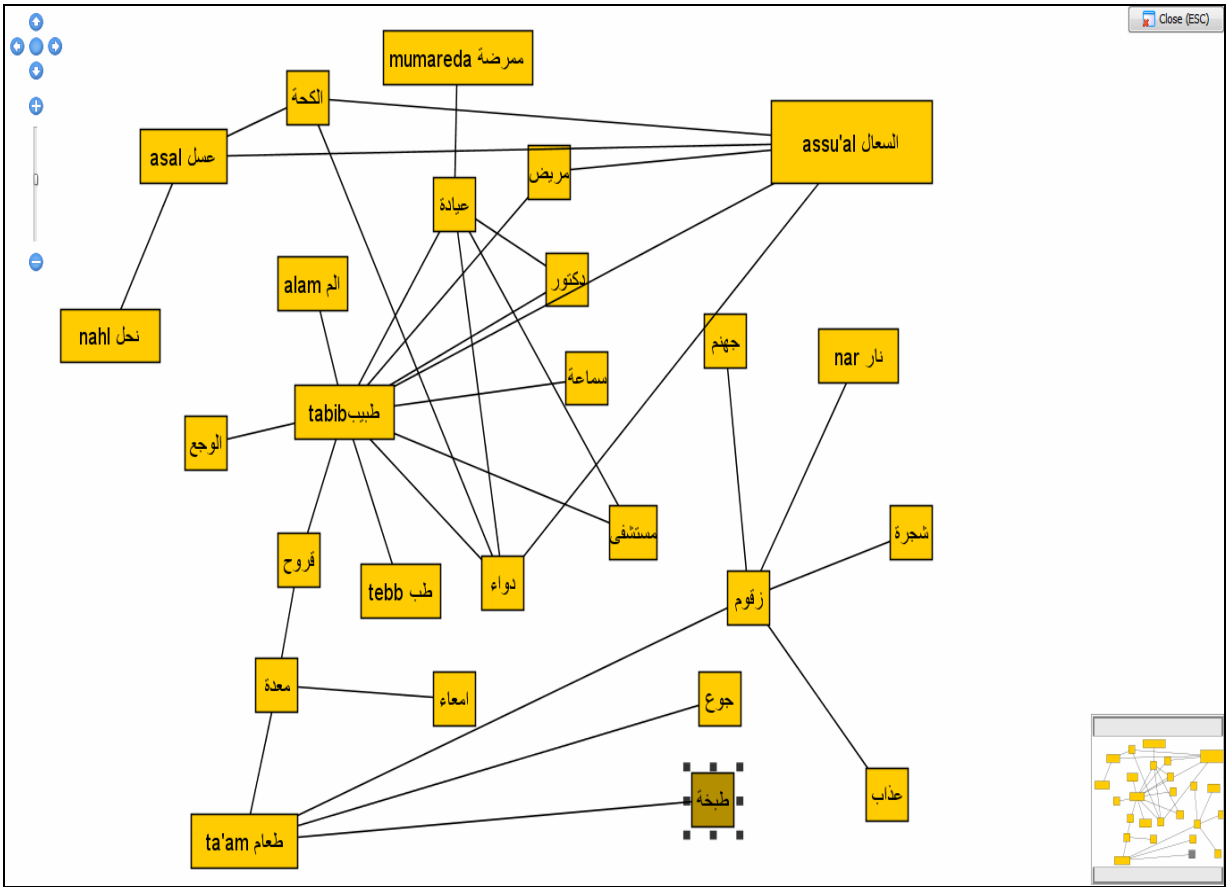


Fig. 84: Sample MPG sub-graph

IX.1.5 **Evaluation**

IX.1.5.1 Usability

JDMAR proved the potential of indirect active contributions to preterminology of a language with poor resources such as Arabic by maintaining a quite acceptable growth rate throughout the experiment compared to other instances of JeuxDeMots, like the Japanese.

It achieved its purpose of collecting preterminology through a non linguistic (non-lexical) activity.

The second and secondary purpose was to attract and entertain Arabic users. The game did have loyal players (and guests) from 18 different countries. 20 registered players answered a survey question to know if they are playing for entertainment, or only to contribute to a lexical resource, 3/20 answered that they play for entertainment, 16/20 said that they play for entertainments and contribution, and 1/20 said he plays only to contribute.

IX.1.5.2 Quality

Out of the 1400 contributed relations, 200 relations were selected, regardless of the rw weight.

Around 175 out of 200 corresponded to a meaningful functional or ontological relation, and 25 were inaccurate. Table 12 shows the categorization.

Table 12: Categorization of the contributed data

Category	Number of relations
Synonyms, writing variations, dialectal variations, plural, singular…	77
Antonyms	15
Generalization	23
Specialization	19
Other meaningful	41
Other inaccurate	25
Total	200

IX.1.5.3 Coverage

Out of the 200 relations selected previously, only 46 were found in the Arabic WordNet (as a specific final relation).

To test the availability of the new content in the produced preterminology, 200 preterms from JDMAR-MPG were selected randomly. Only 125 were found in Alburaq.net, and only 65 were found in the Arabic WordNet.

The reason is that active contribution through serious games receives dialectal and new preterms that are not available in traditional repositories, such as the terms in table 13 .

Table 13: Sample preterms

Preterm in Arabic	Transliteration	English translation
فيسبوك	facebook	Facebook
تويتر	twitter	Twitter
شريط كاسيت	shareet kasit	cassette tape
قطايف	qatayef	in the Levant area it is a sort of sweet crepe commonly served during the month of Ramadan
توجيهي	tawjeehi	the general secondary examination in Jordan and Palestine (the west bank and Gazza)
مفلوز	mfalwez	someone who has the influenza
بيّاره	bayyara	citrus (orange, lemon, Clementine,…) farm

Such a repository with its natural growth in terms of members and content can significantly enrich Arabic preterminology (terminology) in various domains.

IX.2 DSR-Contribution Gateway

IX.2.1 Overview and Objectives

Traditional collaborative approaches depend on professionals or non-professionals in the development of lexical resources. However, the activity of the contribution itself is usually a linguistic or lexical activity during which a contributor inputs his terminological knowledge purposefully for the sake of contribution. Such an activity does not satisfy the contribution factors *CFs* as it needs a high motivation.

The suggested approach in active contribution for preterminology depends on attaching the contribution activity to a highly attractive activity (in order to satisfy the *M* factor).

As an experiment, MPG-DSR has been enlarged (1) by a contribution gateway that offers attractive activities to the visitors of the DSR, and (2) by receiving contributions through these

services, particularly, the Contribute While Searching, and the Contribute While Reading activities.

This section describes the experiments of offering these services and receiving contributions from the visitors of Toyo Bunko's archive during a period of 4 months. We also evaluated the usability of the service and the contributed data, in terms of its importance as a possible solution for absent terminology.

IX.2.2 **Evaluation Criteria**

IX.2.2.1 Application Performance

a. CWS

As the Contribute While Searching offers a CLIR service using the DSR-MPG, the performance of this service in terms of precision and recall should be evaluated to measure the effectiveness of MPG in a multilingual search application.

b. CWR

The usability of this application can be measured through the number of users it serves and the amount of contribution it receives.

IX.2.2.2 Lexical Capabilities

The growth rate of the number of contributions and the uniqueness of the contributed multilingual preterminology can be used to measure the potential of a non linguistic active contribution approach.

IX.2.3 **Baseline**

IX.2.3.1 Current DSR Search

The current DSR website offers a search service where all the archived books are indexed (figure 85), to be searched and retrieved by visitors.

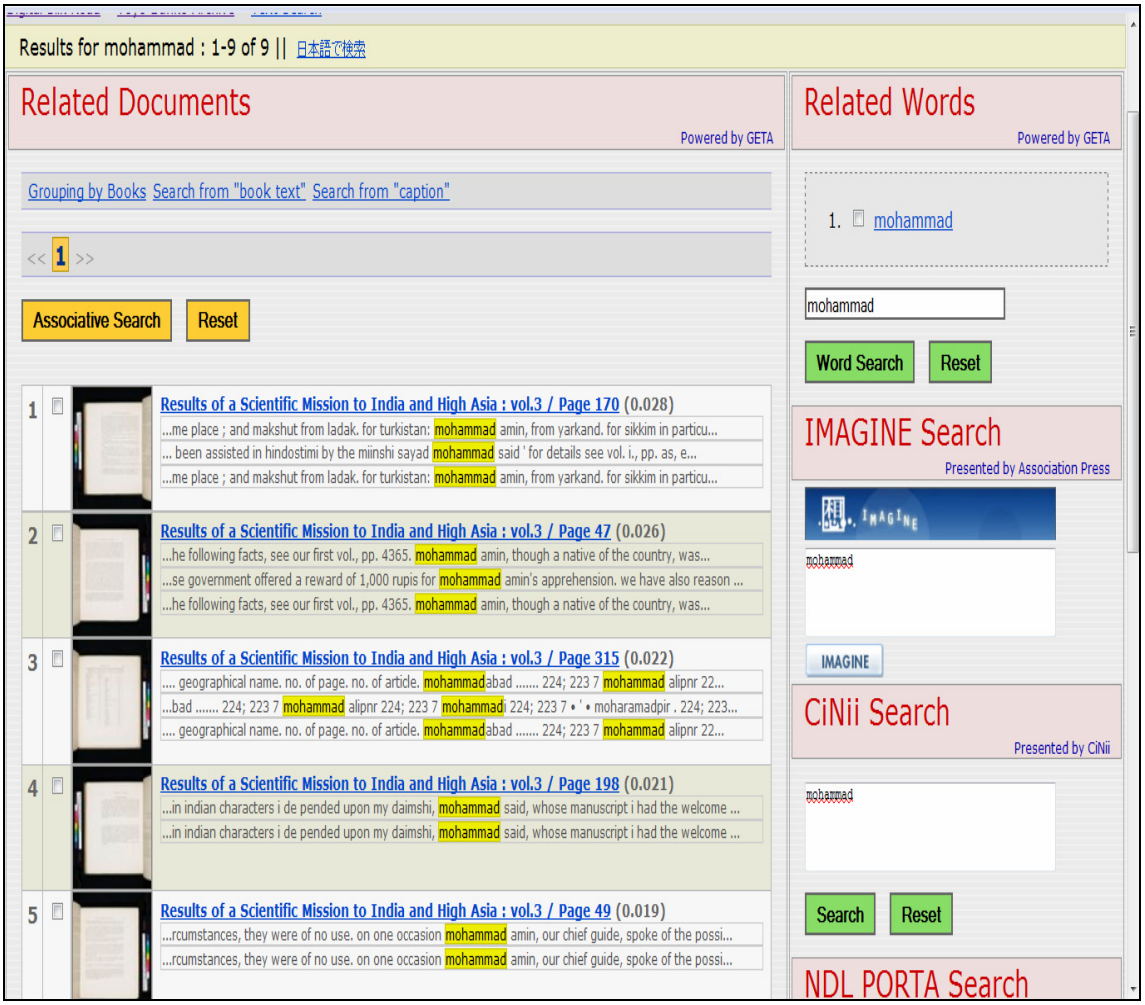


Fig. 85: DSR current search engine

IX.2.3.2 Current Reading View

The current reading view shows the original scanned image of the book, and behind it there is a text box that contains the OCR text extracted form the image. Visitors can browse the book through a set of appropriate links (See figure 86).

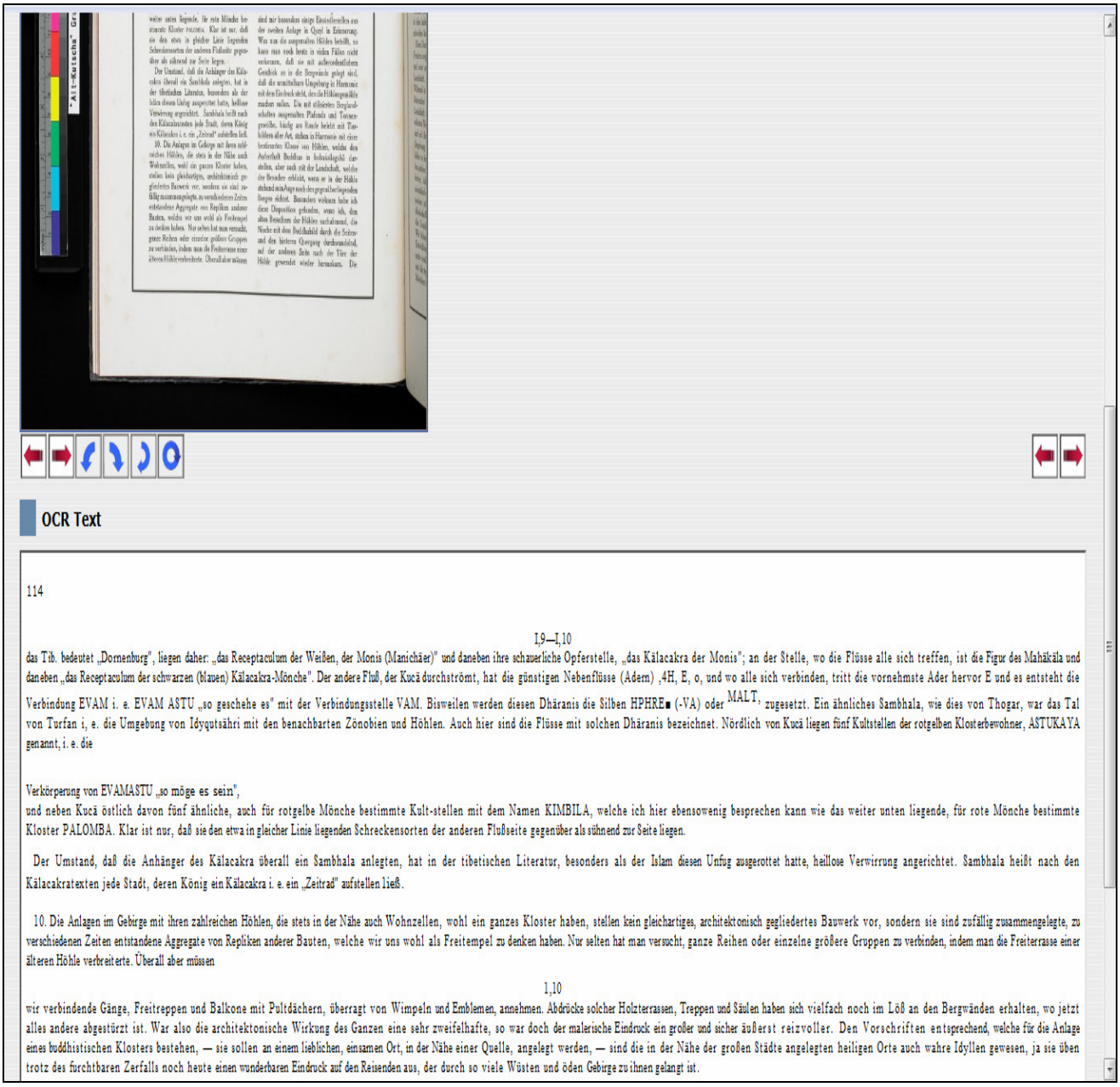


Fig. 86: A sample DSR page

IX.2.3.3 Traditional Repositories

The contributed data will be evaluated based on the content available in Alburaq.net and google.com as general purpose dictionaries, in order to investigate the nature of these lexical data and trace their importance.

IX.2.4 **Experiment Settings**

- URL: http://dsr.nii.ac.jp/pTMDB/ is the general URL that shows the multilingual search engine. To access a page through the gateway in a CWR mode, one uses the URL: http://dsr.nii.ac.jp/pTMDB/makepage2.jsp? + the URL of the original page of the book.

- Referred websites: each page has a link to the contribution gateway so that users can switch to contribution mode (See figure 87).

Fig. 87: Switching to contribution mode by following the above "beta" links

- Visitors: any Internet visitor of the website.
- MPG: the MPG that serves these two applications is the MPG constructed using passive approaches.
- Launching date: the gateway was launched around the middle of April, 2010; later it was linked through all the pages of the archived books. Since that time, all readers of the books can use it.

IX.2.5 **Results and Statistics**

Figure 88 shows the total number of visits up until the end of each month. Since April 2010, the gateway received around 580 visits, 180 of which (31%) led to a contribution (see figure 89).

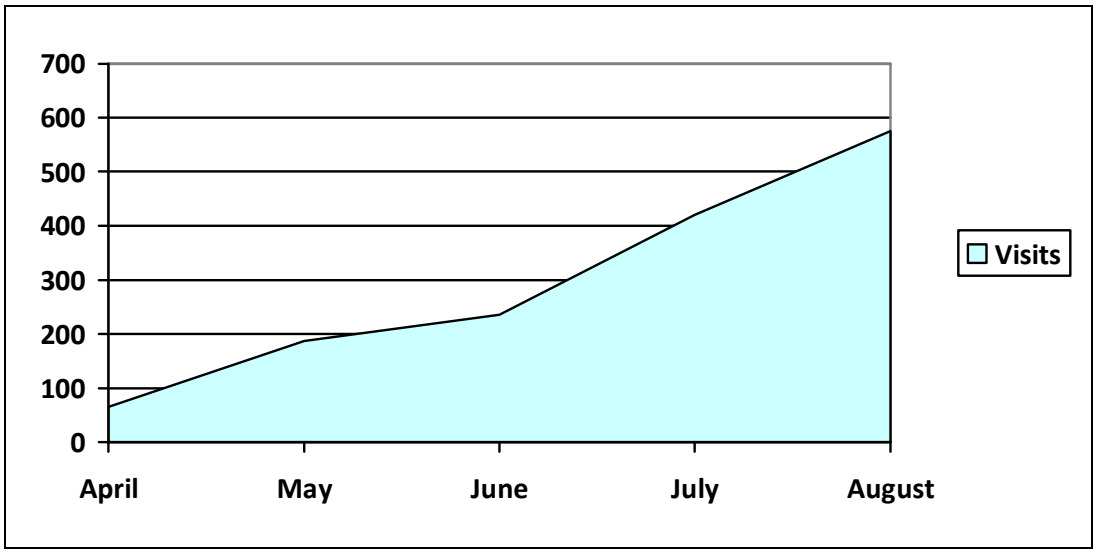


Fig. 88: Cumulative number of visits to the contribution gateway of the DSR-MPG

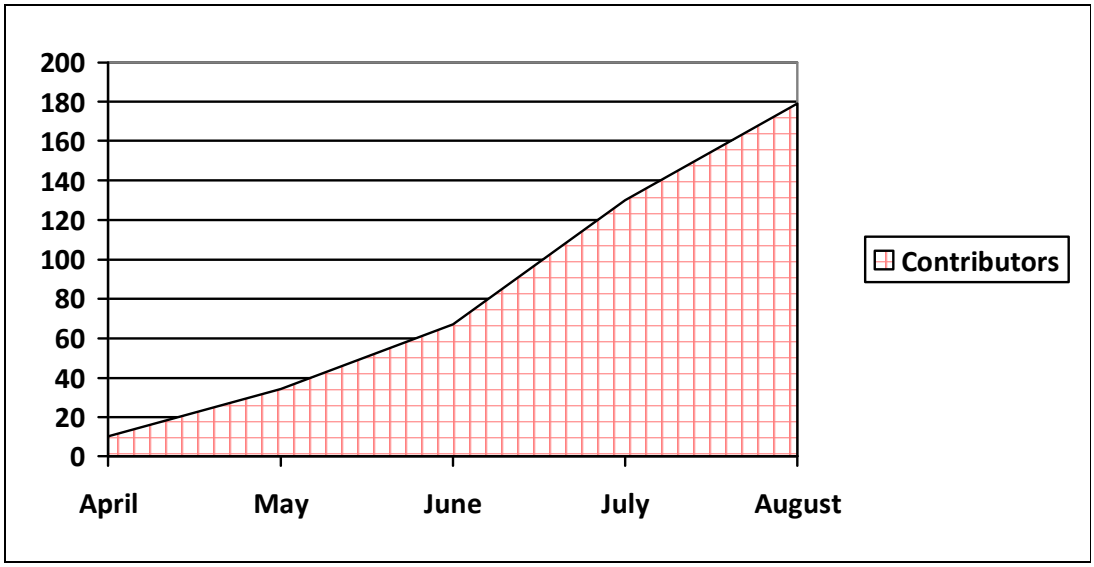


Fig. 89: Cumulative number of contributions

In a contribution session, several lexical units may be contributed. That is why 920 lexical units have been received from 180 contribution sessions. This is around 5.1 lexical contributions/contribution session.

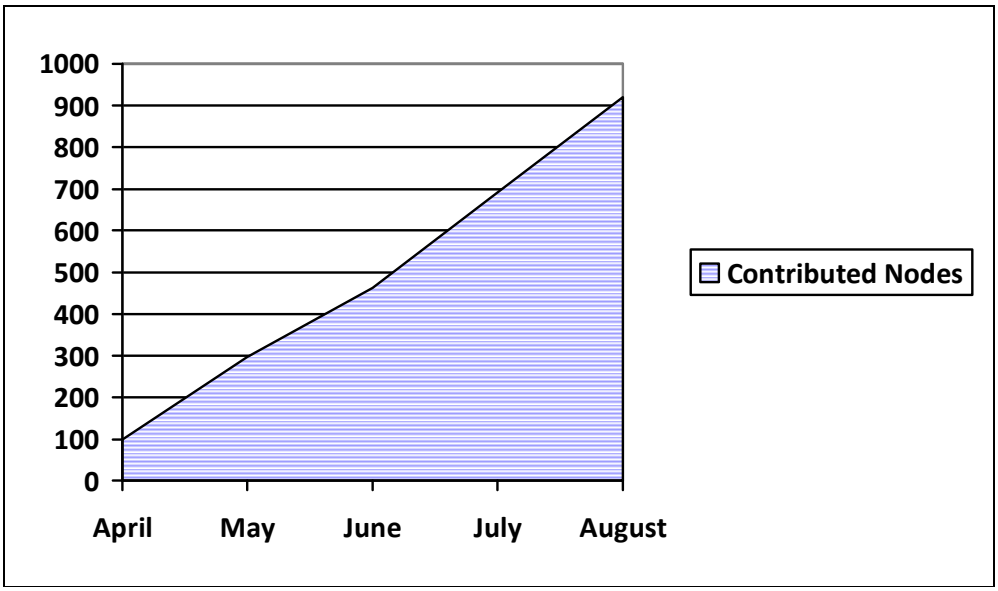


Fig. 90: Cumulative number of contributed nodes over the period

Table 14 shows the amount of contributions for each language.

Table 14: Contribution language

Language	Number of contributed nodes
Arabic	390
French	239
Japanese	139
Chinese	95
German	44
Russian	26
Italian	22
Thai	2

IX.2.6 **Evaluation**

IX.2.6.1 Application Performance

a. *CWS*

During the 4-month period, CWS received 156 requests. The performance is similar to the performance of the current system when the search terms are in English. However, CWS received 68 requests in languages other than English and Japanese. 44 of these requests were translated by MPG-DSR successfully and results were found, while the monolingual search retrieved zero results for such requests. The access rate is hence 44/68 = ~65%.

b. *CWR*

CWS offers the same functionalities as a normal DSR page, it only adds the option of selecting a term from the text to be translated and edited. This feature helps readers who need linguistic assistance, especially for translating special preterms.

This application serves also as a simple term look-up for the DSR-MPG. During the experiment, the system received around 500 visits and maintained a simple term translation service upon selection.

IX.2.6.2 Lexical Coverage

The Arabic contributions have been evaluated: 345 terms out of 390 have been correctly translated (88.4%). The incorrect terms contained inaccurate translations that needed post-editing or general-purpose translations rather than specific translations.

The majority of the terms were multi-word terms, and the 900 contributions added 540 new nodes to the MPG.

We then tried to find out if the contributed data is already available in traditional resources. As the contributed 345 resulted from around 61 contribution session, 70 Arabic terms from different sessions (around 1 term from each contribution session) were selected and searched in Alburaq and Google Dictionary. The results were as follows (see table 15):

- Google Dictionary: 26 terms were found.
- Alburaq: 19 terms were found.

Table 15: Results from searching sample contributed preterms

	Sample contributed preterms from DSR-MPG	**Google dictionary**	**Alburaq**
Number of preterms	70	26	19
Percentage	100%	37.1%	27.1%

IX.3 Assessments and Observations

Passive approaches proved to make a good use of what is available in digital format, especially for hidden terminology. However, in the case where the term is not established yet in some languages, human active involvement in the maintenance of the MPG is very necessary.

That is why MPG was experimented with 3 active contribution applications in two different systems. The Arabic JeuxDeMots achieved its purpose of entertaining average online players and collecting a significant amount of preterms and relations. Its content (preterms, and monolingual correspondences) was unique and contained spoken words that are not available in standard databases, and the amount of growth was stable and promising.

The other active contribution technique is a direct one; although the contribution activity is embedded in a non-linguistic activity which is a normal activity of the online community itself. Contribute While Reading and Contribute While Searching have shown a good potential in attracting contributions from average domain experts. The contributed data were distinctive because they did not appear using the passive approaches, and a low percentage of the contributed terms was covered by classical dictionaries.

Conclusions, Perspectives and Future Work

Nowadays there is a massive production of terms in various domains (50 terms daily, 17,500 yearly) (Alfadhel 2007). Most of these terms are produced in well-resourced languages, i.e. English, French… Finding equivalences in different languages is very important and essential for various applications. For many languages where the community is not active in term production and term translation, more terms will be latent or even absent, which causes a lexical gap.

Traditionally, terminologists are responsible for developing terminology. Their work on a term involves building the correspondence between the symbolic representation, the concept and terminological information to situate the term in the terminological sphere (information like definitions). Building such correspondences is an exhausting task which affects the cost and coverage. On the other hand, collaborative approaches try to prepare the work of terminologists by amateur volunteers which is a promising trend in acquiring knowledge. However, it may be difficult to induce volunteers to be involved in a linguistic activity, even if they are actually familiar with the domain.

Automatic approaches use textual and lexical resources to develop terminology, but are limited to the available resources and the used techniques. Also, not all terminological knowledge is available in textual format (latent terminology). Besides, even when automatic approaches are effective in finding terms, it is difficult to build the same correspondences established by terminologists.

This thesis proposed to use unconventional approaches and to develop preliminary resources for terminology, made of "preterms" and easy-to-build but loose correspondences, called *preterminology*. These approaches target latent and absent terminology. Following a growing trend in modern computational lexicography, and computational linguistics in general, we represent a preterminological resource as a potentially very large lexical graph (*Multilingual Preterminological Graph "MPG"*).

The methods for constructing a preterminology depend on the community of the domain. That community can passively contribute to preterminology by analyzing its produced resources, and it can actively contribute by being directly involved in the contribution process.

A system to develop and maintain MPGs and preterminologies has been designed and implemented. SEpT (7 in French) (System for Eliciting preTerminology) is a computer system that functions as a preterminology elicitor through building and maintaining MPGs. SEpT interact with MPG and with an external application, offers facilities to view the graph and receive active contributions. An MPG can be developed with a combination of passive and active approaches.

Passive approaches depend on domain-dedicated digital (textual or non textual) resources, while active approaches depend on human direct or indirect contribution. JeuxDeMots is an example of active contribution applications that is used to enrich preterminology. Two other active contribution applications have been suggested and experimented within this thesis: Contribute While Reading and Contribute While Searching. Both try to attract contributions through an activity that does not involve heavy linguistic and terminological tasks.

Passive approaches in developing MPGs have been evaluated; the produced graph achieved better linguistic and lexical coverage than other terminological repositories, due to the careful utilization of digital resources, especially access log files. Active approaches showed their potential in few a months of experiments where unique data have been contributed directly and indirectly.

An MPG can contain very dynamic correspondences that can handle a variety of ontological and lexical relations. Developing further methods and techniques to exploit the MPG structure is a suggested future work.

Looking for other digital (non-textual) resources seems very promising, as the experiments with access log files have shown. Utilizing similar digital resources for building a terminology could perhaps improve the way terms are recorded and adopted.

Other future research axes are the development of active contribution scenarios and applications, as human computation proved its effectiveness in various knowledge acquisition systems.

Conclusions, perspectives et travaux futurs

Aujourd'hui, il y a une production massive de termes dans divers domaines (50 termes par jour, 17.500 par an) (Alfadhel 2007). La plupart de ces termes sont produits dans des langues bien dotées en ressources, à savoir l'anglais, le français … Trouver des équivalences dans différentes langues est très important, et essentiel pour diverses applications. Pour de nombreuses langues où la communauté n'est pas active en terme de production et de traduction de termes, il y aura plus de termes qui seront latents ou même absents, ce qui entraîne un fossé lexical.

Traditionnellement, les terminologues sont chargés d'élaborer la terminologie. Leur travail sur un terme consiste à construire la correspondance entre la représentation symbolique, le concept, et l'information terminologique qui situe le terme dans la sphère terminologique (information comme des définitions). Construire de telles correspondances est une tâche épuisante qui influe sur le coût et la couverture. D'autre part, les approches collaboratives essayent de préparer le travail des terminologues par des bénévoles amateurs, ce qui est une tendance prometteuse dans l'acquisition de connaissances. Cependant, il peut être difficile d'inciter des bénévoles à participer à une activité linguistique, même s'ils sont réellement familiers avec le domaine.

Les approches automatiques emploient des ressources textuelles et lexicales pour développer la terminologie, mais sont limitées aux ressources disponibles et aux techniques utilisées. Aussi, ce ne sont pas toutes les connaissances terminologiques qui sont disponibles en format textuel (terminologie latente). D'ailleurs, même lorsque les approches automatiques sont efficaces pour trouver des termes, il est difficile de construire les mêmes correspondances que celles établies par les terminologues.

Cette thèse a proposé d'utiliser des approches non conventionnelles et de développer des ressources préliminaire pour la terminologie, constituées de "prétermes" et de correspondances faciles à construire, mais "lâches", appelées *preterminologie*. Ces approches ciblent la terminologie latente et la terminologie absente. Suite à une tendance croissante en lexicographie computationnelle moderne, et en linguistique computationnelle en général, nous représentons une ressource preterminologique comme un graphe lexical potentiellement très grand *(multilingual Preterminological Graph "MPG").*

Les méthodes de construction d'une préterminologie dépend de la communauté du domaine. Cette communauté peut contribuer passivement à de la préterminologie en analysant ses ressources produites, et elle peut contribuer activement en étant directement impliquée dans le processus de contribution.

Un système pour développer et maintenir des MPG et des préterminologies a été conçu et implémenté. Sept (7 en français) (Système pour Éliciter de la préTerminologie) est un système informatique qui fonctionne comme un éliciteur de préterminologie, grâce à la construction et à la gestion d'un MPG. SEpT interagit avec un MPG et avec une application externe, offre des

facilités pour afficher le graphe et pour recevoir des contributions actives. Un MPG peut être développé avec une combinaison d'approches passives et actives.

Les approches passives dépendent de ressources numériques (textuelles ou non textuelles) dédiées à un domaine, alors que les approches actives dépendent de la contribution directe ou indirecte de l'homme. JeuxDeMots est un exemple d'application à contribution active, qui est utilisée pour enrichir de la préterminologie. Deux autres applications de contribution active ont été proposées et expérimentées dans cette thèse: Contribuer En Lisant (CWR) et Contribuer En Cherchant (CWS). Les deux tentent d'attirer des contributions par le biais d'une activité qui n'implique pas de lourdes tâches linguistiques et terminologiques.

Des approches passive pour développer des MPG ont été évaluées ; le graphe produit a obtenu une meilleure couverture linguistique et lexicale que les autres référentiels terminologiques, en raison de l'utilisation adéquate des ressources numériques, notamment des fichiers de journaux d'accès (logs). Les approches actives ont montré leur potentiel durant quelques mois d'expériences, lors desquelles des données uniques ont été contribuées directement et indirectement.

Un MPG peut contenir des correspondances très dynamiques qui peuvent représenter une variété de relations ontologiques et lexicales. Développer de nouvelles méthodes et techniques pour exploiter la structure MPG est un travail proposé pour l'avenir.

Chercher d'autres ressources numériques (non-textuelles) semble très prometteur, comme les expériences avec des fichiers de journaux d'accès (logs) l'ont montré. Utiliser des ressources numériques similaires pour la construction d'une terminologie pourrait peut-être améliorer la manière dont les termes sont enregistrées et adoptés.

D'autres axes de recherche futurs sont l'élaboration de scénarios et d'applications à contribution active, puisque le "calcul humain" a prouvé son efficacité dans divers systèmes d'acquisition de connaissances.

Bibliography

Ahn, L. v. (2005). "Human Computation". Computer Science. Pittsburgh, Carnegie Mellon University PhD degree, 87 p.

Ahn, L. v. (2006). "Games With A Purpose". IEEE Computer Magazine, Vol. 39(6), pp. 96-98.

Ahn, L. v. and L. Dabbish (2008). "General Techniques for Designing Games with a Purpose." Communications of the ACM, pp. 58-67.

Archer, V. (2009). "Graphes linguistiques multiniveau pour l'extraction de connaissances : l'exemple des collocations." Informatics. Grenoble, Université Joseph Fourier. PhD, 247 p.

Argote, L. and P. Ingram (2000). "Knowledge transfer: A Basis for Competitive Advantage in Firms." Organizational Behavior and Human Decision Processes, Vol. 82, pp. 150-169.

Auger, A. and C. Barrière (2008). "Pattern-based approaches to semantic relation extraction: A state-of-the-art." Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication, Vol. 14(1), pp. 1-19.

Aussenac-Gilles, N. and M.-P. Jacques (2008). "Designing and evaluating patterns for relation acquisition from texts with Caméléon." Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication, Vol. 14(1), pp. 45-73.

Aussenac-Gilles, N. and D. Sörgel (2005). "Text analysis for ontology and terminology engineering." Applied Ontology, Vol. 1(1), pp. 35-46.

B.S.I. (1963). "Recommendations for the selection, formation and definition of technical terms." London, England, British Standard Institution, 21 p.

Baalabaki, M. and R. Baalabaki (2007)." المورد القريب Almawrid Alqareeb", دار العلم للملايين, 351 p.

Barrière, C. (2009). "The web as a source of informative background knowledge." MT Summit XII - Workshop: Beyond Translation Memories: New Tools for Translators MT, 8 p.

Bellynck, V., C. Boitet, et al. (2005). "ITOLDU, a Web Service to Pool Technical Lexical Terms in a Learning Environment and Contribute to Multilingual Lexical Databases." Computational Linguistics and Intelligent Text Processing, Springer Berlin / Heidelberg, pp. 324-332.

Berment, V. (2004). "Méthodes pour informatiser les langues et les groupes de langues « peu dotées »." Informatics. Grenoble/France, Université Joseph Fourier - Grenoble I, PhD, 277 p.

Bernardi, U., A. Bocsak, et al. (2005). "Are we making ourselves clear? Terminology management and machine translation at Volkswagen." 10th EAMT conference "Practical applications of machine translation", pp. 44-49.

Bey, Y. (2008). "Aides informatisées à la traduction collaborative bénévole sur le Web." Informatics. Grenoble/France, Université Joseph Fourier - Grenoble I. PhD, 270 p.

Bey, Y., C. Boitet, et al. (2006). "The TRANSBey prototype: an online collaborative Wiki-based CAT environment for volunteer translators." LREC-2006: Fifth International Conference on Language Resources and Evaluation. Third International Workshop on Language Resources for Translation Work, Research & Training (LR4Trans-III). Genoa, Italy, pp. 49-54.

Boitet, C., M. Mangeot, et al. (2002). "The PAPILLON project: cooperatively building a multilingual lexical data-base to derive open source dictionaries & lexicons." Proceedings of the 2nd workshop on NLP and XML, Vol. 17, 3 p.

Boitet, C. and N. Nédeau (1997). "Mémoire des mots en traitement automatique des langues et «étymologie synchronique»." LTT, Tunis.

Calberg-Challot, M., D. Candel, et al. (2008). "Une analyse méthodique pour l'extraction terminologique dans le domaine du nucléaire." Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication, Vol. 14(2), pp. 183–203.

Callison-Burch, C. and R. S. Flournoy (2001). "A program for automatically selecting the best output from multiple machine translation engines." MT Summit VIII: Machine Translation in the Information Age, Proceedings, Santiago de Compostela, Spain, 18-22 September 2001, pp. 63-66.

Chen, A. (2002). "Cross-Language Retrieval Experiments at CLEF 2002." in CLEF-2002 working notes, 16 p.

Chklovski, T. and Y. Gil (2005-a). "Improving the Design of Intelligent Acquisition Interfaces for Collecting World Knowledge from Web Contributors." Third International Conference on Knowledge Capture (K-CAP), Banff, Canada, October 2-5, 2005, pp. 35-42.

Chklovski, T. and Y. Gil (2005-b). "An Analysis of Knowledge Collected from Volunteer Contributors." Twentieth National Conference on Artificial Intelligence (AAAI-05), Pittsburgh, Pennsylvania, July/2005, Vol. 2, pp. 564-570.

Claveau, V. and M.-C. L'Homme (2005). "Structuring terminology using analogy-based machine learning." In Proceedings of TKE'2005 (Terminology and Knowledge Engineering), Copenhagen, Denmark, 12 p.

Copestake, A., T. Briscoe, et al. (1994). "Acquisition of lexical translation relations from MRDS." Machine Translation Vol. 9(3-4) / September, 1994, pp. 183-219.

Corréard, M.-H. and M. Mangeot (1999). "XML- A Solution For LDBs, EDs and MRDs?" COMPLEX 1999, Pécs, Hungary, 6 p.

Cruse, D. A. (1986). "Lexical Semantics." Cambridge University Press, Avon, Great Britain, 328 p.

CYC (1991). "Ideas for Applying Cyc." MCC Technical Report Number ACT-CYC-407-91, 32 p.

Daoud, M. (2007). "Vers des passerelles interactives d'accès multilingue (iMAG)". Informatique. Grenoble, Université Joseph Fourier. Master 2.

Daoud, M., C. Boitet, et al. (2009). "Constructing multilingual preterminological graphs using various online-community resources." the Eighth International Symposium on Natural Language Processing (SNLP2009), Thailand, pp. 116-121.

Daoud, M., C. Boitet, et al. (2009). "Building a Community-Dedicated Preterminological Multilingual Graphs from Implicit and Explicit User Interactions". Second International Workshop on REsource Discovery (RED 2009), co-located with VLDB 2009, Lyon, France, 8 p.

Daoud, M., K. Kageura, et al. (2010). "Passive and Active Contribution to Multilingual Lexical Resources through Online Cultural Activities." NLPKE10, Beijing, China, 4 p.

Daoud, M., K. Kageura, et al. (2009). "Using Wikipedias to Initialize Multilingual Preterminological Lexical Resources." Wikimedia Conference Japan 2009 (WCJ 2009), Tokyo, Japan.

Daoud, M., A. Kitamoto, et al. (2008). "A CLIR-Based Collaborative Construction of Multilingual Terminological Dictionary for Cultural Resources." Translating and the Computer 30, London-UK, 12 p.

Désilets, A., L. Gonzalez, et al. (2006). "Translation the Wiki Way." Proceedings of the WIKISym 2006 - The Conference Wiki of the 2006 International Symposium. Odense, Denmark. August 21-23, 2006, pp. 19-32.

Desmet, I. and S. Boutayeb (1994). "Terms and words: Propositions for terminology." Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication Vol. 1(2), pp. 303-325.

Diab, M. and N. Habash (2009). "Arabic Dialect Processing." MEDAR09. April, 2009, Cairo, Egypt. Tutorial.

Dubuc, R. (1992). "Manuel pratique de terminologie." Brossard (Québec), 3rd edition, Linguatech, 194 p.

Etzioni, O., K. Reiter, et al. (2007). "Lexical translation with application to image searching on the web." MT Summit XI, Copenhagen, Denmark, pp.175-182.

Even, S. (1979). "Graph Algorithms." Computer Science Press, 213 p.

Francopoulo, G., M. George, et al. (2006). "Lexical Markup Framework (LMF)." International Conference on Language Resources and Evaluation - LREC 2006 (2006), Gênes, Italy, 4 p.

Gaschler, J. and M. Lafourcade (1994). "A Case of Building and Manipulating a Dictionary with Very Simple Tools: the FEM Dictionary." ICLA, Penang, Malaysia, pp. 34-37.

Ghazal, L. (1977). "The New Methodology for Assigning Arabic Terminology." IERA, Rabat.

Gibbon, D. (2002). "On Lexicon Macrostructures." E-MELD workshop, Detroit.

Gouda, Y. (1991). "Dreams and their meanings in the Old Arab tradition." Vantage Press (New York), 471 p.

Gruber, T. (2009). "Ontology." Encyclopedia of Database Systems, pp. 1963-1965.

Halskov, J. and C. Barrière (2008). "Web-based extraction of semantic relation instances for terminology work." Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication, Vol. 14(1), pp. 20-44

Hartley, A. and C. Paris (1997). "Multilingual Document Production: from Support for Translating to Support for Authoring." Machine Translation, Vol. 12(1-2), pp. 109-29.

Heid, U., S. Jauß, et al. (1996). "Term Extraction With Standard Tools for Corpus Exploration -- Experience from German." Proceedings of the 4th International Congress on Terminology and Knowledge Engineering - TKE'96, pp. 26-30.

Huang, F., Y. Zhang, et al. (2005). "Mining key phrase translations from web corpora." Human Technology Conference and Conference on Empirical Methods in Natural Language Processing, HLT-EMNLP-2005, Vancouver, pp. 483-490.

Huynh, C.-P., C. Boitet, et al. (2008). "SECTra_w : an Online Collaborative System for Evaluating, Post-editing and Presenting MT Translation Corpora." LREC-08 Conference. Marrakech, Morocco, 8 p.

ISO-1087-1 (2000). "Terminology work -- Vocabulary -- Part 1: Theory and application." ISO, Switzerland.

ISO-6156 (1987). "Magnetic tape exchange format for terminological/ lexicographical records (MATER)." International Organization for Standardization, Genéva.

Jones, G., F. Fantino, et al. (2008). "Domain-Specific Query Translation for Multilingual Information Access Using Machine Translation Augmented With Dictionaries Mined From Wikipedia". Proceedings (CLIA-2008), Hyederabad, India, 8 p.

Joubert, A. and M. Lafourcade (2008). "JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes." In proc of JADT'2008, Ecole normale supérieure Lettres et sciences humaines, Lyon, France, pp. 657-666.

Kageura, K. (2002). "The Dynamics of Terminology: A descriptive theory of term formation and terminological growth", Terminology and Lexicography Research and Practice 5, 322 p.

Kageura, K. and B. Umino (1998). "Methods of automatic term recognition." Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication, Vol. 3(2), pp. 259-289.

Kawazoe, A., L. Jin, et al. (2008). "The development of a schema for the annotation of terms in the BioCaster disease detection/tracking system." Journal of Applied Ontology, 9 p.

Kim, A. J. (2000). "Community Building on the Web: Secret Strategies for Successful Online Communities." Addison Wesley, 352 p.

Kim, Y. G., S. I. Yang, et al. (2005). "Terminology construction workflow for Korean-English patent MT." MT Summit X. Phuket, Thailand, 5 p.

Kit, C. and X. Liu (2008). "Measuring mono-word termhood by rank difference via corpus comparison." Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication, pp. 204-229.

Kübler, N. (2002). "Creating a term base to customise an MT system: reusability of resources and tools from the translator's point of view." LREC-2002: Third International Conference on Language Resources and Evaluation. Workshop: Language resources for translation work and research, Las Palmas Canary Islands, 27 May 2002, pp. 44-48.

Kwong, O. Y., B. K.Tsou, et al. (2002). "Alignment and extraction of bilingual legal terminology from context profiles." COLING 2002: Second international workshop on computational terminology (COMPUTERM 2002). Taipei,Taiwan, 7 p.

Lafourcade, M. and V. Prince (2001). "Relative Synonymy and Conceptual vectors." NLPRS2001, Tokyo, Japan. pp. 127-134

Loerch, U. (2000). "An Introduction to Graph Algorithms." Auckland, New Zealand, University of Auckland.

Louise, G., P. James, et al. (1996). "The role of lexicons in natural language processing." Comm. ACM, Vol. 39(1), pp. 63-72.

Malik, M. G. A., C. Boitet, et al. (2008). "Hindi Urdu machine transliteration using finite-state transducers." Proceedings of the 22nd International Conference on Computational Linguistics COLING 2008, Vol. 1, pp. 537-544.

Mangeot, M. (2001). "Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue." Informatics, Université Joseph-Fourier - Grenoble I PhD, 296 p.

Mangeot, M. (2006). "Dictionary Building with the Jibiki Platform. Software Demonstration." Proc. of EURALEX 2006, Torino, Italy, 5 p.

Mangeot, M. (2009). "Projet Mot à mot : élaboration d'un système lexical multilingue parle biais de dictionnaires bilingues." journées scientifiques LTT 2009, Lisbonne, Portugal, 15-17 octobre 2009, 12 p.

Mariam, L. G., L. Gillam, et al. (2005). "Terminology and the Construction of Ontology." Terminology, Vol. 11, pp. 55-81.

Mihalcea, R. and T. Chklovski (2003). "Open mind word expert: Creating large annotated data collections with web users' help". In Proceedings of 4th International Workshop on Linguistically Interpreted Corpora (LINC-03) held in conjunction with EACL-2003, 8 p.

Miller, G. A. (1995). "WordNet: A Lexical Database for English." Communications of the ACM, Vol. 38(11), pp. 39-41.

Murata, T., M. Kitamura, et al. (2003). "Implementation of collaborative translation environment 'Yakushite Net'." MT Summit IX. New Orleans, USA, pp. 479-482.

Nagata, M., T. Saito, et al. (2001). "Using the web as a bilingual dictionary." Proceedings of the workshop on Data-driven methods in machine translation, Vol. 14, pp. 1-8

Nguyen, H. T., C. Boitet, et al. (2007). "PIVAX, an online contributive lexical data base for heterogeneous MT systems using a lexical pivot." SNLP-2007. Bankok, Thailand, 6 p.

Nikkhou, M. and K. Choukri (2005). "Survey on Arabic Language Resources and Tools in the Mediterranean Countries." ELDA/NEMLAR, 44 p.

Oard, D. (1999). "Global Access to Multilingual Information." Fourth International Workshop on Information Retrieval with Asian Languages. Taipei-Taiwan, 8 p.

Och, F. J. and H. Ney (2003). "A Systematic Comparison of Various Statistical Alignment Models." Computational Linguistics, Vol. 29(1), pp. 19-51

Ogden, C. K. and I. A. Richards (1923). "The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism." University of Cambridge, 362 p.

Ono, K., A. Kitamoto, et al. (2008). "Memory of the Silk Road -The Digital Silk Road Project-." Proceedings of (VSMM08), Project Papers, Limassol, Cyprus, pp. 437-444.

Ono, K., A. Kitamoto, et al. (2007). "Digital Silk Road Project: Current Status and Future Perspectives." e-Culture Workshop, 24th APAN Meeting.

Ostransky, B. (2005). "The art of medieval arab oneirology." Archiv orientální, Vol. 73(4), pp. 407-428.

Pennington, J., E. Smith, et al. (1994). "LANGUAL: A food-description language." Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication, Vol. 1(2), pp. 277-289.

Popescu-Belis, A. (2003). "An experiment in comparative evaluation: Humans vs. Computers." Proceedings of the Ninth Machine Translation Summit., New Orleans, Louisiana, USA, 8 p.

Prince, V. and S. Ferrari (2000). "Création et extension automatiques de dictionnaires terminologiques multilingues spécialisés." Ressources et évaluation en ingéniérie des langues, ed. De Boeck, Paris: pp. 211-223.

Resnik, P. and N. A. Smith (2003). "The web as a parallel corpus." Computational Linguistics, Vol. 29(3), pp. 349-380.

Rey, A. (1995). "Essays on Terminology." John Benjamins Publishing Company, 223 p.

Richardson, M. (2008). "Learning about the World through Long-Term Query Logs." ACM Transactions on the Web (TWEB), Vol. 2(4), 27 p.

Richardson, M. and P. Domingos (2003). "Building large knowledge bases by mass collaboration." International Conference On Knowledge Capture, Sanibel Island, FL, USA, pp. 129-137.

Sadat, F., M. Yoshikawa, et al. (2003). "Bilingual terminology acquisition from comparable corpora and phrasal translation to cross-language information retrieval." ACL-2003: 41st Annual meeting of the Association for Computational Linguistics, Vol. 2, pp. 141 - 144.

Sager, J. C. (1990). A Practical Course in Terminology Processing Amesterdam/Philadelphia, John Benjamins Publishing Company, 266 p.

Schmitz, K.-D. (2006). "Terminology and Terminological Databases." Encyclopedia of Languages and Linguistics, Vol. 12, 9 p.

Sérasset, G. (1994). "Interlingual lexical organisation for multilingual lexical databases in nadia." COLING 94, Vol. 1, pp. 278-282.

Sérasset, G. (2004). "A Generic Collaborative Platform for Multilingual Lexical Database Development." COLING 2004, Geneva, Switzerland, Aug. 2004, pp. 73-79. .

Sérasset, G., F. Brunet-Manquat, et al. (2006). "Multilingual legal terminology on the Jibiki platform: the LexALP project." Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, COLING ACL-2006. Sydney, Australia, Association for Computational Linguistics, pp. 937-944.

Shimohata, S., M. Kitamura, et al. (2001). "Collaborative Translation Environment on the Web." In Proceedings of the MT Summit VIII, pp. 331-334.

Singh, P., T. Lin, et al. (2002). "Open Mind Common Sense: Knowledge acquisition from the general public." Lecture Notes In Computer Science; Vol. 2519. On the Move to Meaningful Internet Systems, 2002 - DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002, pp. 1223-1237.

Speer, R. (2007). "Open Mind Commons: An Inquisitive Approach to Learning Common Sense". Workshop on Common Sense and Intelligent User Interfaces, Honolulu-Hawaii, 5 p.

Stermsek, G., M. Strembeck, et al. (2007). "A User Profile Derivation Approach based on Log-File Analysis." in Hamid R. Arabnia & Ray R. Hashemi, ed., 'IKE' , CSREA Press, pp. 258-264.

TEI (2009). "TEI P5: Guidelines for Electronic Text Encoding and Interchange." Oxford - Providence - Charlottesville - Nancy, 1420 p.

Temmerman, R. (2000). "Towards New Ways of Terminology Description: The sociocognitive approach." John Benjamins, 258 p.

Tercedor, M. and C. I. López-Rodríguez (2008). "Integrating corpus data in dynamic knowledge bases: The Puertoterm project." Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication, Vol. 14(2), pp. 159–182.

Tiedemann, J. (2003). "Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing." Studia Linguistica Upsaliensia 1, 142 p.

Trippel, T. (1999). "Terminology for Spoken Language Systems." Universität Bielefeld, Fakultät für Linguistik und Literaturwissenschaft, 25. Mai 1999, PhD, 149 p.

Tufiş, D. (2004). "Term translations in multilingual corpora: discovery and consistency check." Fourth International Conference on Language Resources and Evaluation, LREC-2004. Lisbon, Portugal, pp.1981-1984.

Uchida, H. and M. Zhu (2003). "The Universal Networking Language specification, version 3.0 UNDL Foundation." 43 p.

UN-Geo (2002). "Glossary of Terms for the Standardization of Geographical Names." UN, New York.

Valderrábanos, A. S., A. Belskis, et al. (2002). "TExtractor: a multilingual terminology extraction tool." Proceedings of the second international conference on Human Language Technology Research, San Diego, California, pp. 393-398.

Vo-Trung, H. (2004). "Reuse of free online MT engines to develop a meta-system of multilingual machine translation." EsTAL 2004 (Espana for Natural Language Processing), Vol. 3230, pp. 303-313.

Vo-Trung, H., H. K. Phan, et al. (2005). "FEV Dictionary, a product of the generic solutions to import it Vietnamese in Papillon project." LTT 2005, Bruxelles, Belgium, 8 p.

Yassin, Y. A. (2003). "Why Arabic Is the Most Difficult Language for Localization." Globalization Insider, Vol. XII (3.6), 5 p.

Netography

Alfadhel, A. (2007). "Saudi Terminology Data Bank." Retrieved 11/2007, 2007, from http://gdis.kacst.edu.sa/resources.html.

Aljayyash. (2010). "Tafseer Al Ahlaam." from http://dreams.aljayyash.net/.

Apache. (2008). "Solr." Retrieved 20-3-2008, from http://lucene.apache.org/solr/.

Arabization.org. (2010). "المنظمة العربية للتربية و الثقافة و العلوم- مكتب تنسيق التعريب." Retrieved 1/7/2010, 2010, from http://www.arabization.org.ma/.

AWN. (2010). "Arabic WordNet." Retrieved 1/7/2010, 2010, from http://www.globalwordnet.org/AWN/

Babylon. (2009). "Babylon Dictionary." Retrieved 5/5/2009, 2009, from http://www.babylon.com/define/98/English-Arabic-Dictionary.html.

CYC. (2008). "THE FACTory." Retrieved 12/1/2009, from http://game.cyc.com/.

DicML. (2003). "DicML." Retrieved 1/7/2010, 2010, from http://lucasso.republika.pl/DicML/DicML_en.html

EOLSS. (2008). "EOLSS." Retrieved 12 January 2009, from http://www.eolss.net/.

ETB. (2009). "Eurotermbank." Retrieved 1/4/2010, 2010, from

http://www.eurotermbank.com.

EVROTERM. (2010). "EVROTERM." Retrieved 1/4/2010, 2010, from http://evroterm.gov.si.

EURAC. (2007). "LexALP Information System." Retrieved 11/2007, from http://217.199.4.152:8080/general/lexalp/index.php.

EuroFIR. (2008). "LanguaL." Retrieved 1/4/2010, 2010, from http://langual.org.

EUSKALTERM. (2007). "Basque Public Term Bank." Retrieved 10/10/2008, 2008, from http://www1.euskadi.net/euskalterm/.

Facebook. (2010). "Facebook." Retrieved 1/7/2010, 2010, from http://facebook.com.

FACTory. (2010). "FACTory." Retrieved 1/9/2010, 2010, from http://game.cyc.com/.

FAO. (2008). "FAO TERMINOLOGY." Retrieved 1/9/2008, 2008, from http://www.fao.org/faoterm.

Fujitsu. (2009). "ATLAS 14." Retrieved 1/4/2010, 2010, from http://www.fujitsu.com/global/services/software/translation/atlas/improvements.html.

Google. (2008). "Google Translate." Retrieved 1 June 2008, 2008, from http://translate.google.com.

Google. (2009). "Google Image Labeler." Retrieved 11/3/2009, 2009, from http://images.google.com/imagelabeler/?src=b.

Google. (2010). "Google Dictionary." Retrieved 1/7/2010, 2010, from http://www.google.com/dictionary.

GraphML. (2010). "The GraphML File Format." Retrieved 1/7/2010, 2010, from http://graphml.graphdrawing.org/.

Himsolt, M. (2010). "GML: A portable Graph File Format." Retrieved 1/7/2010, 2010, from http://www.infosun.fim.uni-passau.de/Graphlet/GML/gml-tr.html.

IATE. (2008). "Inter-Active Terminology for Europe." Retrieved 10/10/2008, 2008, from http://iate.europa.eu.

IBM. (2010). "Honyaku no osama." Retrieved 1/9/2010, 2010, from http://www-06.ibm.com/software/jp/internet/king/.

IDRC. (2009, 10 January 2009). "The Water Demand Management Glossary (Second Edition)." from http://www.idrc.ca/WaterDemand/IDRC_Glossary_Second_Edition/index.html.

IEC. (2008). "Electropedia." Retrieved 10/10/2008, 2008, from http://dom2.iec.ch/iev/iev.nsf/welcome?openform.

IMF. (2010). "IMF Terminology." Retrieved 1/4/2010, 2010, from http://www.imf.org/external/np/term/index.asp.

Java. (2010). "Java programming language." Retrieved 1/7/2010, 2010, from http://java.com.

JDMAR. (2010). "Arabic JeuxDeMots." Retrieved 1/7/2010, 2010, from http://javalig.imag.fr/jdmar/jdm-accueil.php.

jGrapht. (2010). "jGrapht." Retrieved 1/7/2010, 2010, from http://jgrapht.net.

LCL. (2010). "TermExtractor." Retrieved 2/7/2010, 2010, from http://lcl2.di.uniroma1.it/termextractor/.

LIRMM. (2010). "JeuxDeMots." Retrieved 10/9/2010, 2010, from http://www.lirmm.fr/jeuxdemots/.

LISA (2010). "TBX." Retrieved 1/9/2010, 2010 from http://www.lisa.org/Term-Base-eXchange.32.0.html

Mangeot, M. (1999). "Accès Internet au dictionnaire FEM (français-anglais-malais)." GETA, CLIPS, IMAG, Grenoble, Dictionnaire trilingue d'usage. http://clips.imag.fr/geta/services/fem.

Mangeot, M. and G. Sérasset. (2007). "The Papillon Project" Retrieved 11/2007, 2007, from http://www.papillon-dictionary.org/Home.po.

Merriam-Webster. (2010). "Dictionary and Thesaurus - Merriam-Webster Online." Retrieved 1/7/2010, 2010, from http://www.merriam-webster.com/.

Mindpixel. (2009). "Mindpixel." Retrieved 12/1/2009, from http://mindpixel.com/.

Mozilla. (2006). "Mozilla Localization Project." Retrieved 14-may-2007, from http://www.mozilla.org/projects/l10n/.

NII. (2003). "Digital Silk Road." Retrieved 1/9/2008, 2008, from http://dsr.nii.ac.jp/index.html.en.

NII. (2008). "Digital Archive of Toyo Bunko Rare Books." Retrieved 1 June 2008, 2008, from http://dsr.nii.ac.jp/toyobunko/.

Oxford dictionaries. (2010, 1/7/2010). "How many words are there in the English language?" from http://www.oxforddictionaries.com/page/howmanywords.

PalTel. (2010). "Qamous Al buraq." from http://www.alburaq.net/dictionary/transform.cfm.

Papillon. (2010). "Jibiki API." Retrieved 1/5/2010, 2010, from http://papillon.imag.fr/papillon/Api.po.

reCAPTCHA. (2009). "reCAPTCHA." Retrieved 11/3/2009, 2009, from http://recaptcha.net/.

SimpleDict. (2010). "SimpleDict dictionary." Retrieved 1/7/2010, 2010, from http://sourceforge.net/projects/simpledict/.

StarDict. (2010). "StarDict Disctionary." Retrieved 1/7/2010, 2010, from http://www.stardict.org/.

SUMO. (2010). "Suggested Upper Merged Ontology (SUMO)." Retrieved 1/8/2010, 2010, from http://www.ontologyportal.org/.

Sun-Microsystems. (2007). "Software Globalization." 2007, from http://developers.sun.com/dev/gadc/des_dev/i18ntaxonomy/intro.html.

Systran. (2009). "Systran Web Tranlstor." Retrieved 20/12/2009, 2009, from www.systransoft.com/.

TagCrowd. (2010). "TagCrowd." Retrieved 1/6/2010, 2010, from http://tagcrowd.com/.

WHO-EMRO. (2009). "The unified health lexicon." Retrieved 10 January 2009, from http://www.emro.who.int/umd/.

Wikipedia-A. (2008). "Wikipedia." Retrieved 1 June 2008, 2008, from http://www.wikipedia.org/.

Wikipedia-B. (2010). "Ibn Sirin." Retrieved 1/2/2010, from http://en.wikipedia.org/wiki/Ibn_Sirin.

Wiktionary. (2008). "Wiktionary." Retrieved 1/9/2008, 2008, from http://en.wikipedia.org/wiki/Wiktionary.

WordReference. (2008). "Online Language Dictionaries." Retrieved 1/9/2008, 2008, from http://www.WordReference.com.

XDXF. (2010). "XDXF: XML Dictionary eXchange Format." Retrieved 1/7/2010, 2010, from http://xdxf.sourceforge.net/.

Yahoo. (2008). "Yahoo Terms!" Retrieved 20-3-2008, from http://developer.yahoo.com/search/content/V1/termExtraction.html.

yWorks. (2010). "yEd." Retrieved 1/7/2010, 2010, from http://www.yworks.com/en/products_yed_about.html.

Appendix 1: Sample Multilingual Synonyms from DSR-MPG

Table 16 shows sample multilingual synonyms from the DSR-MPG.

Table 16: Sample multilingual synonyms

English	French	Japanese	Arabic	Thai	German	Chinese	Portuguese	Russian	Italian	Swidish	Avar	Sanskrit
sanskrit	Sanskrit	サンスクリット	لغة سنسكريتية	ภาษาสันสกฤต	Sanskrit	梵语	Sânscrito	Санскрит	Lingua sanscrita	Sanskrit	*****	संस्कृत
yellow river	Huang He	黄河	النهر الأصفر	แม่น้ำฮวงโห	Gelber Fluss	黄河	Rio Amarelo	Хуанхэ	Fiume Giallo	Huanghe	*****	हुआंग हे नदी
himalaya	Himalaya	ヒマラヤ山脈	هيمالايا	เทือกเขาหิมาลัย	Himalaya	喜马拉雅山脉	Himalaia	Гималаи	Himalaya	Himalaya	*****	हिमालय
indian ocean	Océan Indien	インド洋	محيط هندي	มหาสมุทรอินเดีย	Indischer Ozean	印度洋	Oceano Índico	Индийский океан	Oceano Indiano	Indiska oceanen	*****	सिन्धु महासागर
sikkim	Sikkim	シッキム州	سيكيم	รัฐสิกขิม	Sikkim	锡金邦	Siquim	Сикким	Sikkim	Sikkim	*****	सिक्किम
botanists	Botanique	植物学	علم النبات	พฤกษศาสตร์	Botanik	植物學	Botânica	Ботаника	Botanica	Botanik	*****	वनस्पति विज्ञानं
muhammed	Mahomet	ムハンマド・	محمد بن عبد الله	มุฮัมมัด	Mohammed	穆罕默德	Maomé	Мухаммед	Maometto	Muhammed	*****	मुहम्मद
mughal period	Empire moghol	ムガル帝国	امبراطورية مغول الهند	จักรวรรดิโมกุล	Mogulreich	莫卧儿帝国	Império Mogol	Империя Великих Моголов	Gran Mogol	Mogulriket	*****	मुगल साम्राज्य
mahabharata	Mahâbhârata	マハーバーラタ	مهاباراتا	มหาภารตะ	Mahabharata	摩诃婆罗多	Mahabharata	Махабхарата	Mahābhārata	Mahabharata	*****	महाभारतं
majapahit	Majapahit	マジャパヒト王国	إمبراطورية ماجاباهيت	อาณาจักรมัชปาหิต	Majapahit	满者伯夷	*****	*****	Majapahit	Majapahitriket	*****	मजापहित साम्राज्य
brahma	Brahmâ	ブラフマー	براهما	พระพรหม	Brahma	梵天	Brahma	Брахма	Brahma	Brahma	*****	ब्रह्मा

pilgrimage to mecca	Hajj	ハッジ	حج	ฮัจญ์	Haddsch	朝覲	Hajj	Хадж	Hajj	Hajj	*****	*****
chinese emperor	Empereur de Chine	*****	*****	จักรพรรดิ	Kaiserreich China	中国皇帝	*****	*****	*****	*****	*****	*****
ancient egypt	Égypte antique	古代エジプト	مصر القديمة	อียิปต์โบราณ	Altes Ägypten	古埃及	Antigo Egipto	Древний Египет	Antico Egitto	Forntida Egypten	*****	*****
islamabad	Islamabad	イスラマバード	إسلام أباد	อิสลามาบัด	Islamabad	伊斯兰堡	Islamabad	Исламабад	Islamabad	Islamabad	*****	*****
kant	Emmanuel Kant	イマヌエル・カント	إيمانويل كانت	อิมมานูเอล คานท์	Immanuel Kant	伊曼努尔·康德	Immanuel Kant	Кант, Иммануил	Immanuel Kant	Immanuel Kant	*****	इमान्युएल काण्ट
executive power	Pouvoir exécutif	行政	سلطة تنفيذية	อำนาจบริหาร	Exekutive	行政部门	Poder executivo	Исполнительная власть	Potere esecutivo	Verkställande makt	*****	*****
arabian nights	Les Mille et Une Nuits	千夜一夜物語	ألف ليلة وليلة	อาหรับราตรี	Tausendundeine Nacht	一千零一夜	As Mil e uma Noites	Тысяча и одна ночь	Le mille e una notte	Tusen och en natt	*****	*****
british possessions	Territoire britannique d'outre-mer	イギリスの海外領土	*****	อาณานิคมโพ้นทะเลของสหราชอาณาจักร	Britische Überseegebiete	英國海外領土	Territórios britânicos ultramarinos	*****	Territori britannici d'oltremare	Brittiska besittningar och protektorat	*****	*****
majapahit	Majapahit	マジャパヒト王国	إمبراطورية ماجاباهيت	อาณาจักรมัชปาหิต	Majapahit	满者伯夷	*****	*****	Majapahit	Majapahitriket	*****	मजापहित साम्राज्य
allah	Allah	アッラーフ	الله (إسلام)	อัลลอฮ์	Allah	安拉	Alá	Аллах	Allah	Allah	*****	*****
kuran	Coran	クルアーン	القرآن	อัลกุรอาน	Koran	古兰经	Alcorão	Коран	Corano	Koranen	*****	*****
cuneiform writing	Cunéiforme	楔形文字	كتابة مسمارية	อักษรรูปลิ่ม	Keilschrift	楔形文字	Escrita cuneiforme	Клинопись	Scrittura cuneiforme	Kilskrift	*****	*****
kharosthi script	Alphabet kharoṣṭhī	カローシュティー文字	*****	อักษรขโรษฐี	Kharoshthi-Schrift	佉卢文	*****	Кхарошхи	Kharoshthi	Kharosti	*****	*****

medieval period	Moyen Âge	中世	عصور وسطى	สมัยกลาง	Mittelalter	中世纪	Idade Média	Средние века	Medioevo	Medeltiden	*****	*****
byzantine architecture	Architecture byzantine	ビザンティン建築	عمارة بيزنطية	สถาปัตยกรรมไบแซนไทน์	Byzantinische Architektur	拜占庭式建筑	Arquitectura bizantina	Архитектура Византии	Architettura bizantina	Bysantinsk arkitektur	*****	*****
tiflis	Tbilissi	トビリシ	تبليسي	ทบิลิซิ	Tiflis	第比利斯	Tbilisi	Тбилиси	Tbilisi	Tbilisi	Тбилиси	*****
moscow	Moscou	モスクワ	موسكو	มอสโก	Moskau	莫斯科	Moscovo	Москва	Mosca (Russia)	Moskva	Москва	*****
caspian seas	Mer Caspienne	カスピ海	بحر قزوين	ทะเลแคสเปียน	Kaspisches Meer	裡海	Mar Cáspio	Каспийское море	Mar Caspio	Kaspiska havet	Каспий	*****
geography	Géographie	地理学	جغرافيا	ภูมิศาสตร์	Geographie	地理学	Geografia	География	Geografia	Geografi	География	*****
bakou	Bakou	バクー	باكو	บากู	Baku	巴库	Bacu	Баку	Baku	Baku	Баку	*****

Appendix 2: Lexical Translation Using Wikipedia

The following method translates a term "term" using Wikipedia; it uses two private methods (tool1 and findCateg). The code lists all the languages that MPG is dealing with through their Wikipedia URLs. The code was automatically generated by extracting all the URLs of a Wikipedia Article.

```
public String translatewikipedia(int term, String slang) {
 String str = "";
 try {
 URL url=new URL("http://en.wikipedia.org/wiki/"+n[term].getPreterm().replaceAll(" ", "_"));
 URLConnection yc = url.openConnection();
 BufferedReader in = new BufferedReader(new InputStreamReader(yc.getInputStream()));
            String inputLine;
            while ((inputLine = in.readLine()) != null) {
                str = str + inputLine;
            }
            in.close();
        } catch (Exception e) {
            System.out.println(e.getMessage());
        }
        int currcount = pcount;
        if ((!str.contains("<title>Error</title>")) && str.length() > 100) {
           String categ = findCateg(str);
           String en = tool1("http://en.wikipedia.org/wiki/", "utf-8", str, term);
           String de = tool1("http://de.wikipedia.org/wiki/", "utf-8", str, term);
           String fr = tool1("http://fr.wikipedia.org/wiki/", "utf-8", str, term);
           String pl = tool1("http://pl.wikipedia.org/wiki/", "utf-8", str, term);
           String ja = tool1("http://ja.wikipedia.org/wiki/", "utf-8", str, term);
           String it = tool1("http://it.wikipedia.org/wiki/", "utf-8", str, term);
           String nl = tool1("http://nl.wikipedia.org/wiki/", "utf-8", str, term);
           String pt = tool1("http://pt.wikipedia.org/wiki/", "utf-8", str, term);
           String es = tool1("http://es.wikipedia.org/wiki/", "utf-8", str, term);
           String ru = tool1("http://ru.wikipedia.org/wiki/", "utf-8", str, term);
           String sv = tool1("http://sv.wikipedia.org/wiki/", "utf-8", str, term);
           String zh = tool1("http://zh.wikipedia.org/wiki/", "utf-8", str, term);
           String no = tool1("http://no.wikipedia.org/wiki/", "utf-8", str, term);
           String fi = tool1("http://fi.wikipedia.org/wiki/", "utf-8", str, term);
           String ca = tool1("http://ca.wikipedia.org/wiki/", "utf-8", str, term);
           String vo = tool1("http://vo.wikipedia.org/wiki/", "utf-8", str, term);
           String ro = tool1("http://ro.wikipedia.org/wiki/", "utf-8", str, term);
           String tr = tool1("http://tr.wikipedia.org/wiki/", "utf-8", str, term);
           String uk = tool1("http://uk.wikipedia.org/wiki/", "utf-8", str, term);
           String eo = tool1("http://eo.wikipedia.org/wiki/", "utf-8", str, term);
           String cs = tool1("http://cs.wikipedia.org/wiki/", "utf-8", str, term);
           String sk = tool1("http://sk.wikipedia.org/wiki/", "utf-8", str, term);
           String hu = tool1("http://hu.wikipedia.org/wiki/", "utf-8", str, term);
           String da = tool1("http://da.wikipedia.org/wiki/", "utf-8", str, term);
           String id = tool1("http://id.wikipedia.org/wiki/", "utf-8", str, term);
           String he = tool1("http://he.wikipedia.org/wiki/", "utf-8", str, term);
           String lt = tool1("http://lt.wikipedia.org/wiki/", "utf-8", str, term);
           String ko = tool1("http://ko.wikipedia.org/wiki/", "utf-8", str, term);
           String sr = tool1("http://sr.wikipedia.org/wiki/", "utf-8", str, term);
           String sl = tool1("http://sl.wikipedia.org/wiki/", "utf-8", str, term);
           String ar = tool1("http://ar.wikipedia.org/wiki/", "utf-8", str, term);
           String bg = tool1("http://bg.wikipedia.org/wiki/", "utf-8", str, term);
           String et = tool1("http://et.wikipedia.org/wiki/", "utf-8", str, term);
           String hr = tool1("http://hr.wikipedia.org/wiki/", "utf-8", str, term);
           String new1 = tool1("http://new.wikipedia.org/wiki/", "utf-8", str, term);
           String vi = tool1("http://vi.wikipedia.org/wiki/", "utf-8", str, term);
           String te = tool1("http://te.wikipedia.org/wiki/", "utf-8", str, term);
           String nn = tool1("http://nn.wikipedia.org/wiki/", "utf-8", str, term);
           String fa = tool1("http://fa.wikipedia.org/wiki/", "utf-8", str, term);
           String th = tool1("http://th.wikipedia.org/wiki/", "utf-8", str, term);
           String gl = tool1("http://gl.wikipedia.org/wiki/", "utf-8", str, term);
           String el = tool1("http://el.wikipedia.org/wiki/", "utf-8", str, term);
           String ceb = tool1("http://ceb.wikipedia.org/wiki/", "utf-8", str, term);
           String simple=tool1("http://simple.wikipedia.org/wiki/", "utf-8", str, term);
           String ms = tool1("http://ms.wikipedia.org/wiki/", "utf-8", str, term);
           String eu = tool1("http://eu.wikipedia.org/wiki/", "utf-8", str, term);
           String ht = tool1("http://ht.wikipedia.org/wiki/", "utf-8", str, term);
           String bs = tool1("http://bs.wikipedia.org/wiki/", "utf-8", str, term);
           String bpy = tool1("http://bpy.wikipedia.org/wiki/", "utf-8", str, term);
           String lb = tool1("http://lb.wikipedia.org/wiki/", "utf-8", str, term);
           String ka = tool1("http://ka.wikipedia.org/wiki/", "utf-8", str, term);
           String is = tool1("http://is.wikipedia.org/wiki/", "utf-8", str, term);
           String sq = tool1("http://sq.wikipedia.org/wiki/", "utf-8", str, term);
           String la = tool1("http://la.wikipedia.org/wiki/", "utf-8", str, term);
           String br = tool1("http://br.wikipedia.org/wiki/", "utf-8", str, term);
```

```
String hi = tool1("http://hi.wikipedia.org/wiki/", "utf-8", str, term);
String az = tool1("http://az.wikipedia.org/wiki/", "utf-8", str, term);
String bn = tool1("http://bn.wikipedia.org/wiki/", "utf-8", str, term);
String mk = tool1("http://mk.wikipedia.org/wiki/", "utf-8", str, term);
String mr = tool1("http://mr.wikipedia.org/wiki/", "utf-8", str, term);
String sh = tool1("http://sh.wikipedia.org/wiki/", "utf-8", str, term);
String tl = tool1("http://tl.wikipedia.org/wiki/", "utf-8", str, term);
String cy = tool1("http://cy.wikipedia.org/wiki/", "utf-8", str, term);
String io = tool1("http://io.wikipedia.org/wiki/", "utf-8", str, term);
String pms = tool1("http://pms.wikipedia.org/wiki/", "utf-8", str, term);
String lv = tool1("http://lv.wikipedia.org/wiki/", "utf-8", str, term);
String ta = tool1("http://ta.wikipedia.org/wiki/", "utf-8", str, term);
String su = tool1("http://su.wikipedia.org/wiki/", "utf-8", str, term);
String oc = tool1("http://oc.wikipedia.org/wiki/", "utf-8", str, term);
String jv = tool1("http://jv.wikipedia.org/wiki/", "utf-8", str, term);
String nap = tool1("http://nap.wikipedia.org/wiki/", "utf-8", str, term);
String nds = tool1("http://nds.wikipedia.org/wiki/", "utf-8", str, term);
String scn = tool1("http://scn.wikipedia.org/wiki/", "utf-8", str, term);
String be = tool1("http://be.wikipedia.org/wiki/", "utf-8", str, term);
String ast = tool1("http://ast.wikipedia.org/wiki/", "utf-8", str, term);
String ku = tool1("http://ku.wikipedia.org/wiki/", "utf-8", str, term);
String wa = tool1("http://wa.wikipedia.org/wiki/", "utf-8", str, term);
String af = tool1("http://af.wikipedia.org/wiki/", "utf-8", str, term);
String be_x_old=tool1("http://be-x-old.wikipedia.org/wiki/", "utf-8",str, term);
String an = tool1("http://an.wikipedia.org/wiki/", "utf-8", str, term);
String tg = tool1("http://tg.wikipedia.org/wiki/", "utf-8", str, term);
String ksh = tool1("http://ksh.wikipedia.org/wiki/", "utf-8", str, term);
String fy = tool1("http://fy.wikipedia.org/wiki/", "utf-8", str, term);
String cv = tool1("http://cv.wikipedia.org/wiki/", "utf-8", str, term);
String vec = tool1("http://vec.wikipedia.org/wiki/", "utf-8", str, term);
String zh_yue = tool1("http://zh-yue.wikipedia.org/wiki/", "utf-8", str, term);
String roa_tara = tool1("http://roa-tara.wikipedia.org/wiki/", "utf-8", str, term);
String ur = tool1("http://ur.wikipedia.org/wiki/", "utf-8", str, term);
String qu = tool1("http://qu.wikipedia.org/wiki/", "utf-8", str, term);
String sw = tool1("http://sw.wikipedia.org/wiki/", "utf-8", str, term);
String uz = tool1("http://uz.wikipedia.org/wiki/", "utf-8", str, term);
String ga = tool1("http://ga.wikipedia.org/wiki/", "utf-8", str, term);
String bat_smg = tool1("http://bat-smg.wikipedia.org/wiki/", "utf-8", str, term);
String mi = tool1("http://mi.wikipedia.org/wiki/", "utf-8", str, term);
String gd = tool1("http://gd.wikipedia.org/wiki/", "utf-8", str, term);
String ml = tool1("http://ml.wikipedia.org/wiki/", "utf-8", str, term);
String yo = tool1("http://yo.wikipedia.org/wiki/", "utf-8", str, term);
String co = tool1("http://co.wikipedia.org/wiki/", "utf-8", str, term);
String kn = tool1("http://kn.wikipedia.org/wiki/", "utf-8", str, term);
String pam = tool1("http://pam.wikipedia.org/wiki/", "utf-8", str, term);
String yi = tool1("http://yi.wikipedia.org/wiki/", "utf-8", str, term);
String hsb = tool1("http://hsb.wikipedia.org/wiki/", "utf-8", str, term);
String nah = tool1("http://nah.wikipedia.org/wiki/", "utf-8", str, term);
String ia = tool1("http://ia.wikipedia.org/wiki/", "utf-8", str, term);
String li = tool1("http://li.wikipedia.org/wiki/", "utf-8", str, term);
String sa = tool1("http://sa.wikipedia.org/wiki/", "utf-8", str, term);
String als = tool1("http://als.wikipedia.org/wiki/", "utf-8", str, term);
String hy = tool1("http://hy.wikipedia.org/wiki/", "utf-8", str, term);
String tt = tool1("http://tt.wikipedia.org/wiki/", "utf-8", str, term);
String roa_rup = tool1("http://roa-rup.wikipedia.org/wiki/", "utf-8", str, term);
String zh_min_nan = tool1("zh-min-nan.wikipedia.org/wiki/", "utf-8", str, term);
String pag = tool1("http://pag.wikipedia.org/wiki/", "utf-8", str, term);
String map_bms = tool1("http://map-bms.wikipedia.org/wiki/", "utf-8", str, term);
String am = tool1("http://am.wikipedia.org/wiki/", "utf-8", str, term);
String wuu = tool1("http://wuu.wikipedia.org/wiki/", "utf-8", str, term);
String fo = tool1("http://fo.wikipedia.org/wiki/", "utf-8", str, term);
String nrm = tool1("http://nrm.wikipedia.org/wiki/", "utf-8", str, term);
String vls = tool1("http://vls.wikipedia.org/wiki/", "utf-8", str, term);
String lmo = tool1("http://lmo.wikipedia.org/wiki/", "utf-8", str, term);
String nds_nl = tool1("http://nds-nl.wikipedia.org/wiki/", "utf-8", str, term);
String rm = tool1("http://rm.wikipedia.org/wiki/", "utf-8", str, term);
String se = tool1("http://se.wikipedia.org/wiki/", "utf-8", str, term);
String ne = tool1("http://ne.wikipedia.org/wiki/", "utf-8", str, term);
String war = tool1("http://war.wikipedia.org/wiki/", "utf-8", str, term);
String fur = tool1("http://fur.wikipedia.org/wiki/", "utf-8", str, term);
String lij = tool1("http://lij.wikipedia.org/wiki/", "utf-8", str, term);
String nov = tool1("http://nov.wikipedia.org/wiki/", "utf-8", str, term);
String sco = tool1("http://sco.wikipedia.org/wiki/", "utf-8", str, term);
String dv = tool1("http://dv.wikipedia.org/wiki/", "utf-8", str, term);
String bh = tool1("http://bh.wikipedia.org/wiki/", "utf-8", str, term);
String diq = tool1("http://diq.wikipedia.org/wiki/", "utf-8", str, term);
String pi = tool1("http://pi.wikipedia.org/wiki/", "utf-8", str, term);
String kk = tool1("http://kk.wikipedia.org/wiki/", "utf-8", str, term);
String ilo = tool1("http://ilo.wikipedia.org/wiki/", "utf-8", str, term);
String os = tool1("http://os.wikipedia.org/wiki/", "utf-8", str, term);
String zh_classical=tool1("zh-classical.wikipedia.org/wiki/", "utf-8",str, term);
String mt = tool1("http://mt.wikipedia.org/wiki/", "utf-8", str, term);
String frp = tool1("http://frp.wikipedia.org/wiki/", "utf-8", str, term);
String lad = tool1("http://lad.wikipedia.org/wiki/", "utf-8", str, term);
String fiu_vro = tool1("http://fiu-vro.wikipedia.org/wiki/", "utf-8", str, term);
String pdc = tool1("http://pdc.wikipedia.org/wiki/", "utf-8", str, term);
String csb = tool1("http://csb.wikipedia.org/wiki/", "utf-8", str, term);
```

```
String bar = tool1("http://bar.wikipedia.org/wiki/", "utf-8", str, term);
String kw = tool1("http://kw.wikipedia.org/wiki/", "utf-8", str, term);
String to = tool1("http://to.wikipedia.org/wiki/", "utf-8", str, term);
String mn = tool1("http://mn.wikipedia.org/wiki/", "utf-8", str, term);
String haw = tool1("http://haw.wikipedia.org/wiki/", "utf-8", str, term);
String km = tool1("http://km.wikipedia.org/wiki/", "utf-8", str, term);
String gv = tool1("http://gv.wikipedia.org/wiki/", "utf-8", str, term);
String ps = tool1("http://ps.wikipedia.org/wiki/", "utf-8", str, term);
String ang = tool1("http://ang.wikipedia.org/wiki/", "utf-8", str, term);
String tk = tool1("http://tk.wikipedia.org/wiki/", "utf-8", str, term);
String ln = tool1("http://ln.wikipedia.org/wiki/", "utf-8", str, term);
String ie = tool1("http://ie.wikipedia.org/wiki/", "utf-8", str, term);
String tpi = tool1("http://tpi.wikipedia.org/wiki/", "utf-8", str, term);
String gu = tool1("http://gu.wikipedia.org/wiki/", "utf-8", str, term);
String ty = tool1("http://ty.wikipedia.org/wiki/", "utf-8", str, term);
String crh = tool1("http://crh.wikipedia.org/wiki/", "utf-8", str, term);
String wo = tool1("http://wo.wikipedia.org/wiki/", "utf-8", str, term);
String jbo = tool1("http://jbo.wikipedia.org/wiki/", "utf-8", str, term);
String ay = tool1("http://ay.wikipedia.org/wiki/", "utf-8", str, term);
String sc = tool1("http://sc.wikipedia.org/wiki/", "utf-8", str, term);
String si = tool1("http://si.wikipedia.org/wiki/", "utf-8", str, term);
String zea = tool1("http://zea.wikipedia.org/wiki/", "utf-8", str, term);
String eml = tool1("http://eml.wikipedia.org/wiki/", "utf-8", str, term);
String ky = tool1("http://ky.wikipedia.org/wiki/", "utf-8", str, term);
String ig = tool1("http://ig.wikipedia.org/wiki/", "utf-8", str, term);
String or = tool1("http://or.wikipedia.org/wiki/", "utf-8", str, term);
String mg = tool1("http://mg.wikipedia.org/wiki/", "utf-8", str, term);
String cbk_zam = tool1("http://cbk-zam.wikipedia.org/wiki/", "utf-8", str, term);
String kg = tool1("http://kg.wikipedia.org/wiki/", "utf-8", str, term);
String glk = tool1("http://glk.wikipedia.org/wiki/", "utf-8", str, term);
String arc = tool1("http://arc.wikipedia.org/wiki/", "utf-8", str, term);
String rmy = tool1("http://rmy.wikipedia.org/wiki/", "utf-8", str, term);
String gn = tool1("http://gn.wikipedia.org/wiki/", "utf-8", str, term);
String mo = tool1("http://mo.wikipedia.org/wiki/", "utf-8", str, term);
String so = tool1("http://so.wikipedia.org/wiki/", "utf-8", str, term);
String kab = tool1("http://kab.wikipedia.org/wiki/", "utf-8", str, term);
String ks = tool1("http://ks.wikipedia.org/wiki/", "utf-8", str, term);
String stq = tool1("http://stq.wikipedia.org/wiki/", "utf-8", str, term);
String ce = tool1("http://ce.wikipedia.org/wiki/", "utf-8", str, term);
String bcl = tool1("http://bcl.wikipedia.org/wiki/", "utf-8", str, term);
String udm = tool1("http://udm.wikipedia.org/wiki/", "utf-8", str, term);
String mzn = tool1("http://mzn.wikipedia.org/wiki/", "utf-8", str, term);
String pap = tool1("http://pap.wikipedia.org/wiki/", "utf-8", str, term);
String cu = tool1("http://cu.wikipedia.org/wiki/", "utf-8", str, term);
String sah = tool1("http://sah.wikipedia.org/wiki/", "utf-8", str, term);
String tet = tool1("http://tet.wikipedia.org/wiki/", "utf-8", str, term);
String sd = tool1("http://sd.wikipedia.org/wiki/", "utf-8", str, term);
String lo = tool1("http://lo.wikipedia.org/wiki/", "utf-8", str, term);
String ba = tool1("http://ba.wikipedia.org/wiki/", "utf-8", str, term);
String pa = tool1("http://pa.wikipedia.org/wiki/", "utf-8", str, term);
String iu = tool1("http://iu.wikipedia.org/wiki/", "utf-8", str, term);
String na = tool1("http://na.wikipedia.org/wiki/", "utf-8", str, term);
String got = tool1("http://got.wikipedia.org/wiki/", "utf-8", str, term);
String bo = tool1("http://bo.wikipedia.org/wiki/", "utf-8", str, term);
String dsb = tool1("http://dsb.wikipedia.org/wiki/", "utf-8", str, term);
String chr = tool1("http://chr.wikipedia.org/wiki/", "utf-8", str, term);
String cdo = tool1("http://cdo.wikipedia.org/wiki/", "utf-8", str, term);
String hak = tool1("http://hak.wikipedia.org/wiki/", "utf-8", str, term);
String om = tool1("http://om.wikipedia.org/wiki/", "utf-8", str, term);
String my = tool1("http://my.wikipedia.org/wiki/", "utf-8", str, term);
String sm = tool1("http://sm.wikipedia.org/wiki/", "utf-8", str, term);
String ee = tool1("http://ee.wikipedia.org/wiki/", "utf-8", str, term);
String ug = tool1("http://ug.wikipedia.org/wiki/", "utf-8", str, term);
String as = tool1("http://as.wikipedia.org/wiki/", "utf-8", str, term);
String ti = tool1("http://ti.wikipedia.org/wiki/", "utf-8", str, term);
String av = tool1("http://av.wikipedia.org/wiki/", "utf-8", str, term);
String bm = tool1("http://bm.wikipedia.org/wiki/", "utf-8", str, term);
String zu = tool1("http://zu.wikipedia.org/wiki/", "utf-8", str, term);
String nv = tool1("http://nv.wikipedia.org/wiki/", "utf-8", str, term);
String cr = tool1("http://cr.wikipedia.org/wiki/", "utf-8", str, term);
String pih = tool1("http://pih.wikipedia.org/wiki/", "utf-8", str, term);
String ss = tool1("http://ss.wikipedia.org/wiki/", "utf-8", str, term);
String ve = tool1("http://ve.wikipedia.org/wiki/", "utf-8", str, term);
String bi = tool1("http://bi.wikipedia.org/wiki/", "utf-8", str, term);
String rw = tool1("http://rw.wikipedia.org/wiki/", "utf-8", str, term);
String ch = tool1("http://ch.wikipedia.org/wiki/", "utf-8", str, term);
String xh = tool1("http://xh.wikipedia.org/wiki/", "utf-8", str, term);
String kl = tool1("http://kl.wikipedia.org/wiki/", "utf-8", str, term);
String ik = tool1("http://ik.wikipedia.org/wiki/", "utf-8", str, term);
String bug = tool1("http://bug.wikipedia.org/wiki/", "utf-8", str, term);
String dz = tool1("http://dz.wikipedia.org/wiki/", "utf-8", str, term);
String ts = tool1("http://ts.wikipedia.org/wiki/", "utf-8", str, term);
String tn = tool1("http://tn.wikipedia.org/wiki/", "utf-8", str, term);
String kv = tool1("http://kv.wikipedia.org/wiki/", "utf-8", str, term);
String tum = tool1("http://tum.wikipedia.org/wiki/", "utf-8", str, term);
String xal = tool1("http://xal.wikipedia.org/wiki/", "utf-8", str, term);
String st = tool1("http://st.wikipedia.org/wiki/", "utf-8", str, term);
```

```
            String tw = tool1("http://tw.wikipedia.org/wiki/", "utf-8", str, term);
            String bxr = tool1("http://bxr.wikipedia.org/wiki/", "utf-8", str, term);
            String ak = tool1("http://ak.wikipedia.org/wiki/", "utf-8", str, term);
            String ab = tool1("http://ab.wikipedia.org/wiki/", "utf-8", str, term);
            String ny = tool1("http://ny.wikipedia.org/wiki/", "utf-8", str, term);
            String fj = tool1("http://fj.wikipedia.org/wiki/", "utf-8", str, term);
            String lbe = tool1("http://lbe.wikipedia.org/wiki/", "utf-8", str, term);
            String ki = tool1("http://ki.wikipedia.org/wiki/", "utf-8", str, term);
            String za = tool1("http://za.wikipedia.org/wiki/", "utf-8", str, term);
            String ff = tool1("http://ff.wikipedia.org/wiki/", "utf-8", str, term);
            String lg = tool1("http://lg.wikipedia.org/wiki/", "utf-8", str, term);
            String sn = tool1("http://sn.wikipedia.org/wiki/", "utf-8", str, term);
            String ha = tool1("http://ha.wikipedia.org/wiki/", "utf-8", str, term);
            String sg = tool1("http://sg.wikipedia.org/wiki/", "utf-8", str, term);
            String ii = tool1("http://ii.wikipedia.org/wiki/", "utf-8", str, term);
            String cho = tool1("http://cho.wikipedia.org/wiki/", "utf-8", str, term);
            String rn = tool1("http://rn.wikipedia.org/wiki/", "utf-8", str, term);
            String mh = tool1("http://mh.wikipedia.org/wiki/", "utf-8", str, term);
            String chy = tool1("http://chy.wikipedia.org/wiki/", "utf-8", str, term);
            String aa = tool1("http://aa.wikipedia.org/wiki/", "utf-8", str, term);
            String ng = tool1("http://ng.wikipedia.org/wiki/", "utf-8", str, term);
            String kj = tool1("http://kj.wikipedia.org/wiki/", "utf-8", str, term);
            String ho = tool1("http://ho.wikipedia.org/wiki/", "utf-8", str, term);
            String mus = tool1("http://mus.wikipedia.org/wiki/", "utf-8", str, term);
            String kr = tool1("http://kr.wikipedia.org/wiki/", "utf-8", str, term);
            String hz = tool1("http://hz.wikipedia.org/wiki/", "utf-8", str, term);
            String tokipona = tool1("http://tokipona.wikipedia.org/wiki/", "utf-8", str, term);
            if (pcount > currcount) {
                 return "";
             }
        }
        return "";
    }
private String tool1(String langUrl, String enc, String str, int term) {
        try {
            int start = str.indexOf(langUrl);
            if (start > 1) {
                int end = str.indexOf("\"", start);
                String target = str.substring(start, end);
                target = target.substring(target.lastIndexOf("/") + 1);
                System.out.println(pcount++);
                String langCode = "";
                start = str.indexOf(langUrl) + "http://".length();
                end = str.indexOf(".wiki", start);
                langCode = str.substring(start, end);
                langCode = langCode.trim();
                int currid = addnode(java.net.URLDecoder.decode(target,
                    enc).replaceAll("_", " "), langCode);
                addtedge(term, currid);
                return java.net.URLDecoder.decode(target, enc).replaceAll("_", " ");
            }
            return "*****";
        } catch (Exception e) {
            System.out.println(e.getMessage());
            return "*****";
        }
    }

    private static String findCateg(String str) {
        try {

            int start = str.indexOf("<div id='catlinks' class='catlinks'>");
            if (start > 1) {
                int end = str.indexOf("</div>", start);
                String target = str.substring(start, end).replaceAll("\\<.*?\\>", " +
");
                return target;
            } else {
                return "00";
            }

        } catch (Exception e) {
            System.out.println(e.getMessage());
            return "*****";
        }

    }
```

Appendix 3: Access Log Files Analysis

The following Java method initializes an MPG by analyzing an access log file (filename), it uses a private method called AddArcs.

```
public void accessloginit(String filename) {
        File file = new File(filename);
        FileInputStream fis = null;
        BufferedInputStream bis = null;
        DataInputStream dis = null;
        String t = "";
        int count = 0;
        int nodeCount = 0;
        String currSsn = "";
        String ssn = "";
        String tempS = "";
        int[] temp = new int[2000];
        int sCount = 0;
        try {
            fis = new FileInputStream(file);
            bis = new BufferedInputStream(fis);
            dis = new DataInputStream(bis);
            BufferedReader inp = new BufferedReader(new InputStreamReader(new
              FileInputStream(file), "UTF8"));
            String ss = "x";
            while ((ss = inp.readLine()) != null) {
                if (ss.contains("input=")) {
                    t = findTerm(ss).trim();
                    ssn = currSsn;
                    currSsn = ss.substring(0, ss.indexOf("$$$$$")).trim();
                    int theid = addnode(t, "en");
                    nodeCount++;
                    if ((!currSsn.equals(ssn)) && (count != 0)) {
                        AddArcs(temp, sCount, ssn);
                        temp = new int[2000];
                        temp[0] = theid;
                        sCount = 1;
                    } else if (t != "") {
                        temp[sCount] = theid;
                        sCount++;
                    }
                    count++;
                }
            }
        } catch (Exception e) {
            e.printStackTrace();
        }
    }
private void AddArcs(int[] tmp, int count, String IP) {
        try {
            System.out.println(IP + " : " + count);
            for (int i = 0; i < count; i++) {
                for (int j = i; j < count; j++) {
                    addedge(tmp[i], tmp[j]);
                }
            }
        } catch (Exception e) {
            System.out.print(e.getMessage());

            e.getStackTrace();
        }
    }
```

Résumé

Notre motivation est de combler le fossé terminologique qui grandit avec la production massive de nouveaux concepts (50 quotidiens) dans divers domaines, pour lesquels les termes sont souvent inventés d'abord dans une certaine langue bien dotée, telle que l'anglais ou le français. Trouver des termes équivalents dans différentes langues est nécessaire pour de nombreuses applications, telles que la RI translingue et la TA. Cette tâche est très difficile, particulièrement pour certaines langues très utilisées telles que l'arabe, parce que (1) seule une petite proportion de nouveaux termes est correctement enregistrée par des terminologues, et pour peu de langues ; (2) des communautés spécifiques créent continuellement des termes équivalents sans les normaliser ni même les enregistrer (terminologie latente) ; (3) dans de nombreux cas, aucuns termes équivalents ne sont créés, formellement ou informellement (absence de terminologie).

Cette thèse propose de remplacer le but impossible de construire d'une manière continue une terminologie à jour, complète et de haute qualité pour un grand nombre de langues par celui de construire une preterminologie, en utilisant des méthodes non conventionnelles et des contributions passives ou actives par des communautés d'internautes : extraction de termes parallèles potentiels non seulement à partir de textes parallèles ou comparables, mais également à partir des logs (traces) des visites à des sites Web tels que DSR (Route de la Soie Digitale), et à partir de données produites par des jeux sérieux. Une préterminologie est un nouveau genre de ressource lexicale qui peut être facilement construit et a une bonne couverture.

Suivant en ceci une tendance croissante en lexicographie computationnelle et en TALN en général, nous représentons une préterminologie multilingue par une structure de graphe (Preterminological Multilingual Graph, MPG), où les nœuds portent des prétermes et les arcs des relations préterminologiques simples (synonymie monolingue, traduction, généralisation, spécialisation, etc.) qui sont des approximations des relations (terminologiques ou ontologiques) usuelles.

Un Système complet pour Éliciter une Préterminologie (SEPT) a été développé pour construire et maintenir des MPG. Des approches passives ont été expérimentées en développant un MPG pour le site Web culturel de DSR, et un autre pour le domaine de l'onirologie arabe : les ressources produites ont atteint une bonne couverture informationnelle et linguistique. L'approche indirecte par contribution active est testée depuis 8-9 mois sur l'instance arabe du jeu sérieux JeuxDeMots.

Abstract

Our motivation is to bridge the terminological gap that grows with the massive production of new concepts (50 daily) in various domains, for which terms are often first coined in some well-resourced language, such as English or French. Finding equivalent terms in different languages is necessary for many applications, such as CLIR and MT. This task is very difficult, especially for some widely used languages such as Arabic, because (1) only a small proportion of new terms is properly recorded by terminologists, and for few languages; (2) specific communities continuously create equivalent terms without normalizing and even recording them (latent terminology); (3) in many cases, no equivalent terms are created, formally or informally (absence of terminology).

This thesis proposes to replace the impossible goal of building in a continuous way an up-to-date, complete and high-quality terminology for a large number of languages by that of building a preterminology, using unconventional methods and passive or active contributions by communities of internauts: extracting potential parallel terms not only from parallel or comparable texts, but also from logs of visits to Web sites such as DSR (Digital Silk Road), and from data produced by serious games. A preterminology is a new kind of lexical resource that can be easily constructed and has good coverage.

Following a growing trend in computational lexicography and NLP in general, we represent a multilingual preterminology by a graph structure (Multilingual Preterminological Graph, MPG), where nodes bear preterms and arcs simple preterminological relations (monolingual synonymy, translation, generalization, specialization, etc.) that approximate usual terminological (or ontological) relations.

A complete System for Eliciting Preterminology (SEpT) has been developed to build and maintain MPGs. Passive approaches have been experimented by developing an MPG for the DSR cultural Web site, and another for the domain of Arabic oneirology: the produced resources achieved good informational and linguistic coverage. The indirect active contribution approach is being tested since 8-9 months using the Arabic instance of the JeuxDeMots serious game.