Department of
Information Engineering
and Computer Science DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive, 5 - 38123 POVO, Trento - Italy
http://disi.unitn.it

SEMANTIC NAME MATCHING

Enrico Bignotti

July 2013

Technical Report # DISI-13-029

Thanks: Prof. Fausto Giunchiglia, Juan Pane, Phd, Prof. Raffaella Bernardi

UNIVERSITÀ DEGLI STUDI DI TRENTO

Facoltà di Lettere e Filosofia

Corso di Laurea in Filosofia e Linguaggi della Modernità

Semantic Name Matching

Relatore
Dott. ssa Raffaella Bernardi

Laureando
Enrico Bignotti

Correlatore
Prof. Fausto Giunchiglia
Dott. Juan Pane

Anno Accademico 2011/2012

Acknowledgements

For this work, I would like to thank my advisor prof.ssa Bernardi, for directing me to a fascinating scope of research, that lead me to work with my two co-advisors, prof. Fausto Giunchiglia, whose philosophy of work I greatly admire and I look forward to continue my academic experience with him, and Juan Pane, PhD, who helped me greatly in expanding my research scope, teaching me how to work and write; plus, he showed great patience and understanding. In addition, I would like to thank Enzo Maltese, Alethia Hume, and Aliaksandr Autayeu for their feedback and kindness in pointing out how to develop my thesis.

Outside the scope of this thesis, I want to thank my girlfriend Laura, who helped me discover new worlds, and my loyal friends, especially those from my high school years, the one I enjoy travelling and chatting the most with, and all those people kind enough to show me sympathy throughout all these years.

Finally, I want to thank my family, especially my father Antonio and my mother Carla, who provide and love me, always believing in me, and, last but not least, I want to thank Tanino, for I hope he would be proud of me, and what I have become.

Abstract

I nomi sono dovunque nell'universo quotidiano; qualsiasi cosa di cui noi riteniamo importante parlare ha un nome (un luogo, un film, una persona, etc. . .); perciò, essi sono studiati in molte aree del sapere. In filosofia, i nomi hanno creato un acceso dibattito nel secolo scorso, soprattutto riguardo ai temi della referenza (come fanno i nomi ad indicare qualcosa?) e il significato (i nomi hanno un valore semantico?); nonostante posizioni autorevoli come quelle di Mill, Searle e Kripke, nessuna teoria è ancora accettata universalmente. In sociologia, i nomi sono studiati in quanto classe, poiché esistono nomi che vengono usati in aggiunta al nome originale, ovvero le *varianti* (comunemente detti pseudonimi), che dipendono dal contesto e dai fini di chi li utilizza. Infine, in geografia vengono studiate le *variazioni* multilingue (es., traduzioni) e monolingue (es., errori di battitura) che influiscono sui nomi di entità geopolitiche. Sia le variazioni che le varianti sono (specialmente) problematiche anche per il campo del *name matching*, un'area dell'informatica che si occupa di scoprire istanze che differiscono ortograficamente ma che si riferiscono alla stessa entità; quest'area è il principale ambito del nostro lavoro. La nostra applicazione del name matching è all'interno di un network P2P di utenti, basato sulle entità, che consta di tre livelli: locale (gli utenti), comunità (gruppi di utenti), globale (tutte le entità). Le entità a livello locale sono una visione parziale delle entità reali, ovvero come sono viste dagli utenti, mentre a livello globale sono conservate le entità reali in quanto tali, cio é al di là dei punti di vista personali. Le variazioni e le varianti sono problematiche in quanto possono cambiare radicalmente la struttura del nome in base a fattori linguistici (ovvero, variazioni) e sociali (ovvero, varianti) — difficili da formalizzare con un approccio automatico. Per risolvere questo problema proponiamo una tassonomia delle variazioni e varianti, che serva a predire questi fattori nei nomi delle entità, e cambiare l'architettura di quest'ultimi per rappresentare, oltre al nome originale, sia le varianti sia le variazioni, al fine di rendere pi veritiera la rappresentazione dell'entità stessa. Il nostro approccio è nuovo, perch é importa nozioni multidisciplinari, mai usate prima nell'informatica, da varie aree (filosofia, sociologia, geografia). Inoltre, sfruttiamo i risultati e le scoperte di aree vicine al name matching (es., NED, NER, entity linking), ma adattandoli ed espandendoli per il nostro approccio.

Abstract

Names are studied in different fields, and, among the issues they present, name variations (e.g., translations, misspellings, etc. . .) and name variants (e.g., pseudonyms) pose a challenge to name matching, i.e., discovering instances that differ typographically but represent the same entity. Our scenario for name matching is a P2P, entity-based network of users divided in local level (the users), community level (groups of users), and global level (all the entities). Entities at local level are a partial view of the real word entity, represented at the global level. In this framework, name variations and name variants change the orthography of names because of linguistic and social factors, and their presence depends on the scenario level considered. Thus, they are hard to tackle by an automatic approach such as name matching. Our proposed solutions is to use a taxonomy we created to understand and predict the variations and variants of different entity names, and divide the entity name in different entries to accommodate the original name plus variations and variants. Our approach is novel because we take advantage of a multidisciplinary method, drawing from various fields (i.e., philosophy, sociology and geography) importing terms and views not found in computer science. We also draw from areas close to name matching, building from their findings and expanding them.

Contents

List of Tables

Chapter 1

Introduction

Names are ubiquitous and pervasive in our everyday life, and this is proven by how many different fields study names and their issue. For instance, philosophy witnessed a heated debate throughout the XXth century, focusing on the problems of reference i.e., whether a name has a semantic value, and the problem of meaning, i.e., how a name can indicate or single out an entity. Although many views tried to tackle these problems, no universal agreement has been found yet. In addition, sociology studies show that names are more like a 'class' of types of names, i.e., *name variants*, varying according to different features (e.g., social contexts). Because of this, it is hard to define them (e.g., social context tend to be blurred, if not overlapping), leading to confusion when attempting to refer to a bearer. Finally, geography shows that name not only variate because of social factors, but also because of *linguistic* factors, i.e., mono- and multilingual name variations. By monolingual variation, we mean a variation that happens in one language only, e.g., misspellings, whereas by multilingual variation we mean any translation or transliteration affecting the name when being 'used' by a language other than its own, e.g., translations. Both affect the name's orthography, and further hinder the process of reference.

All these issue, name variants and name variations especially, also affect the scope of computer science; moreover, they affect a particular task of this area, i.e., *name matching*. Name matching consists in "discovering instances that differ typographically (different surface appearance) but represent the same concept/entity" [54]. In this scope, an entity is "a 'thing' which can be distinctly identified [. . .] a specific person, company, or event is an example of an entity" [6]; plus, an entity has *attributes* representing its properties. Consequently, "entities with the same set of attribute [are] *entity types*" [6]. In our case, name matching is used to tackle the issue of name variation ad name variants in a *distributed* scenario, i.e., a P2P network of users, whose structure consists of three levels: local (the level of users, who have their own personal repository of entities), community (the level where a group of users, i.e., a community, can share their entities), and global (the 'universal' level, i.e., a *global* repository of all the entities). Both local and global level represent a different view of entities (and thus different names); in fact, local level represent a *partial* view of an entity, as an entity may be present in multiple repositories with different attribute values, while the global level represent the 'universal' view of an entity, i.e., the real world entity *per se.* This distributed scenario is also affected by name variants and name variations, that can severely modify the 'surface' of names and that vary depending on linguistic and social contexts, which are hard to formalize and tackle for name matching. Plus, monolingual variations and name variants happen at local level, since the user would generally have entities from one language, while multilingual variations are pervasive at global level. In addition, there is the need for entities name to represent not only their original names, but also name variations and name variants, i.e., their architecture should be able to accommodate and show how many variants and variations affect the entity name.

Our proposed solutions are as follows. Firstly, we provide a taxonomy of name variations, which draws from both personal research and state-of-the-art literature, and is divided into four sections, categorizing both multilingual variations (i.e., full and part-of translations) and monolingual variations (i.e., misspellings and format changes), also including name variants (as they may be subjects to name variations). Then, by checking with schema.org, a Web-based taxonomy of types, we use our taxonomy to capture patterns in the way different types are affected by name variations, thus creating guidelines for choosing the correct strategy to tackle them. Secondly, we describe our way of modifying the entity name structure by dividing it in different entries, where the various types of names will be stored. Then, we propose how to implement these solutions in our scenario by illustrating their usage in a step by step fashion.

Our approach is novel, because we take advantage of a multidisciplinary method, drawing from various fields such as philosophy, sociology and geography, importing terms and views not found in the field of computer science. Plus, we take the standard definition of name matching in a wider context, i.e., not just matching strings but also considering factors outside syntax and spellings, working in a distributed scenario. In addition, we also draw from closely related areas, e.g., named entity recognition, named entity disambiguation, and entity linking, building from some of their findings and expanding them for the purpose of our approach.

Finally, the structure of the thesis will be as follows. In Chapter 2 we will provide an overview on the philosophical debate surrounding the issue of reference and meaning of names, whereas in Chapter 3 we will illustrate the findings of sociology and geography on name variations and variants, respectively. Then, we will state our problem in Chapter 5, followed by an overview on fields close to our approach on name matching, and how we are different from them, i.e., Chapter 6. Finally, in Chapter 7 we describe in detail our proposed solutions for our scenario.

Chapter 2

What Are Names?
A Philosophical Overview

In this chapter, after a brief definition of what is a name, we will proceed in sketching the philosophical debate surrounding this subject in the last century. We will distinguish three approaches to names: Millianism, descriptivism, and causal theories. Yet, before that, we need to distinguish between nouns, proper nouns, proper names, and names.

According to [25], nouns refer to a word class having the following properties:

- "*It contains among its members those words that denote persons or concrete objects*".
- "*Its member head phrases -noun phrases- which characteristically function as subject or object in clause structure and refer to the participants in the situation described in the clause, to the actor, patient, recipient, and so on*".
- "*It is the class to which categories of number, gender and case have their 'primary' application in languages which have these grammatical properties*".

In other words, nouns are "the category containing words denoting all kinds of physical objects, such as persons, animals and inanimate objects [...] there are also innumerable abstracts nouns" [26].

This class contains three subclasses: common nouns, proper nouns and pronouns. Disregarding pronouns (as they're not the subject of this work), the main difference between proper and common nouns is *reference*, i.e., the act of singling out a real word entity. In fact, while common nouns do not refer to a single or specific entity (they may do so, but indirectly, e.g. 'The thief broke into a house'), proper nouns do. Plus, since proper nouns belong to a class, i.e., nouns, used to indicate single words (e.g., 'cat', 'love'), they are always considered as single words (e.g., 'Fausto', 'Juan'). Furthermore, proper nouns do not have the full range of determiners, i.e., they can't be used with all determiners (e.g., 'A Juan' loses the uniqueness of reference of 'Juan'), and lack article contrast, e.g., 'Paris' vs. *The Paris.

On the other hand, proper names have also unique referents, but they are distinguished from proper nouns because, while proper noun are single word units, proper names are noun phrases (NPs). In fact, proper nouns characteristically function as the head of NPs serving as proper names [26]; for instance, the proper name 'Jessica Alba' consists of two proper nouns: 'Jessica' and 'Alba'. Nonetheless, proper names may contain other parts of speech, e.g., while 'The University of Cambridge' and 'Cambridge University Press' are proper names, although containing the proper noun 'Cambridge', their heads, 'University' and 'Press', are common nouns. It is then clear that both proper names, e.g. 'Fausto Giunchiglia', and proper nouns, e.g., 'Fausto', deal with unique referents, but, while proper nouns are single syntactical units, proper names are syntagms, i.e., NPs — but are they different from names? If we look at the literature, especially in the area of computer science, it seems that name is a shorter and broader equivalent of both proper name and proper nouns, so we will follow this custom. Therefore, throughout this work we will use 'name' to mean both proper names and proper nouns, specifying when and if needed.

This simple, linguistic based distinction would be sufficient for the whole work, but even in its simplicity, it hides deep and complicated issues. In fact, among the various fields which deal with names (e.g., sociology), philosophy investigated the matter as well, especially in the latest century. According to philosophy, the two main issues that names arise are whether names have a meaning, and how does reference work. In other words, the first problem addresses the possibility of names having a semantic value (in addition to their syntactic one) whereas the second one tries to capture the misterious way by which proper names are somehow 'attached to' things in the world.

Indeed, these two problems surfaced philosophy of language in the last century, starting with J.S. Mill in the middle of the XIX century [40]. Subsequently, as soon as the works Frege [16] and Russel [50] shed some light on matter, a new model imposed itself, with Searle becoming its most passionate defender [51]. Nonetheless, thanks to the inquiries of Kripke in the '70s [31], the whole matter went into questioning again. We will now turn to a brief overview of three main stances on the matter at hand, and a final overview on the present framework.

2.1 Millianism

This is the first stance on the two questions concerning reference and meaning, and derives its name from the philosopher John Stuart Mill, which proposed it in *A System of Logic* [40], published in 1843.

Mill addresses the two issues in Book I, agreeing with Hobbes that a name is "a word taken at pleasure to serve for a mark [...] which being pronounced to others, may be to them a sign of what thought the speaker had before in his mind" [40]. Then, he divides names into *general* and *singular* names (which are essentially proper names), i.e., names that can be "truly affirmed [...] of each of an indefinite number of things" [40], or that can be capable of "being truly affirmed [...] of the same thing" [40], respectively. After these first two classes, he goes on dividing names further by other binary classes (which we won't go into detail here), until he describes the most relevant class for proper names: *connotative* and *non-connotative* (or to use a more modern term, *denotative*). In fact, all names have a connotation and a denotation, i.e., they both connote or imply some attribute(s) and denote or single out individuals that fall under that description. In other words, if we follow the two philosophical issues concerning names, connotation deals with the meaning of the name (i.e., both proper nouns and proper names), whereas denotation indicates its reference.

On the other hand, "proper names [and proper nouns] are not connotative; they denote the individuals who are called by them; but they do not indicate or imply any attribute as belonging to those individuals" [40]. For instance, the town lying at the end of the river Dart, aptly named Dartmouth, would still be called the same even if the river were to change its course, for proper names are "attached to the object themselves, and are not dependent upon the continuance of any attribute of the object" [40]. Thus, Mill claims that proper names have no other meaning but their referent.

Even though Millianism is a simple and intuitive theory, backed up by common sense, there are some problems this theory has a hard time coping with. Firstly, Mill's account of the actual reference seems incomplete. It has been noted that while proper names may be used for actually present individuals (e.g., one could point them after calling their name), proper names are also used to refer to individuals not being present or not even existing at the moment of the utterance. If "John Smit" is just a label for a certain individual, how does it succeed in referring to him when he is not present? Furthermore, there are the so-called 'Frege's puzzles' [53] which are unresolvable by the Millian approach: the first one is the problem of informative identity statements, whereas the second one is the problem of existence statements with proper names as their subjects.

Informative identity statements show that, if we agree with Mill that names have no meaning other than the object they denote, then sentences like "Cicero is Tully" or "Hesperus is Phosphorous" are trivial identity statements (i.e., they simply affirm that the two entities are equal). But they are not, for the speaker does not simply wish to affirm that they are equal. Rather, the speaker hints at some new information about them—such layer is lost with Millianism.

Existence statements with proper names as their subjects show that, since Mill assumes that proper names solely stand for the object they are names of, this leads to the fact that 'Juan Pane exists' makes reference to Juan Pane and then redundantly asserts its existence, while 'Deadpool never existed' seems paradoxically to refer to Deadpool and then assert of him that he never existed. Another criticism by Kripke works along the line of the second puzzle, suggesting that the question 'Does N exist?' cannot be meaningfully when 'N' is replaced with a proper name. In fact, "if I know to whom the [name] has been applied, the answer is automatically 'Yes'. If I dont know to whom it is applied, what am I asking? [...] [O]n Mill's account of proper names, I cannot even understand the question" [31].

Indeed, all these flaws that plague Millianism were dealt with effectively almost fifty years later, starting with Frege, then Russel, and finally Searle, as we will see in the next section.

2.2 Descriptivism

In this section we will describe the second major theory concerning proper names on the issues of reference and meaning, i.e., descriptivism (also known as description theory).

This theory nowadays is split in many different sub-theories (e.g., Katz's DNT [29]) which underline different aspects of its core assumptions or try to avoid its flaws. Notwithstanding the great variety of 'flavours' this theory has, there is a broad enough definition that its followers may agree on. Basically, names refer in virtue of being associated with a definite description or set of definite descriptions that are uniquely true of the individual to which the name refers; in addition, the name has a meaning which consists of the description associated with it.

As we said, this core thought has been declined in order to accomodate for criticism or deeper intuition, but regardless of this, one can see that three philosopher stand out in defining the theory's framework: Frege, Russel and Searle. Therefore, Subsection 2.2.1 will describe Frege's contribution to descriptivism, while Subsection 2.2.2 and 2.2.3 will illustrate Russel and Searle additions, respectively. Finally, Subsection 2.2.4 will introduce the debate between Searle and Kripke, which led to the rise of a third view on the matter of name, i.e., causal theories.

2.2.1 Frege's 'On Sense and Nominatum'

As we said, the first reaction to Millianism, and thus the first sketch of descriptivism, was done by Frege in his treatment of the problem of informative identity statements in *On Sense and Nominatum* (*Über Sinn und Bedeutung*).

Before introducing Frege's arguments, we must explain what does he mean by sense (*Sinn*) and nominatum (*Bedeutung*) (sometimes translated as denotation). In fact, according to Frege's theory, the denotation of a name is the object it picks out; the sense of the name is the mode of presentation of the object, i.e., "a difference in the way in which the designated objects are given" [16].

This distinction is drawn in order to avoid the same problems that Millianism encounters when dealing with the first Frege's puzzle, i.e., the problem of informative identity statements. Following this distinction, informative identity statements may be informative because e.g., 'Hesperus' and 'Phosphorous' express different descriptive senses, i.e., 'The last star visible in the morning' and 'The first star visible in the evening', respectively. This intuition may be easily applied to proper name overall, affirming that their sense is a definite description, as the core of descriptivism claim.

As for the second puzzle, Frege argues that saying 'Aristotle never existed' is simply to say that there was not some object satisfying the descriptive sense expressed by 'Aristotle', e.g., 'The greatest student of Plato'.

Indeed, we can already see the foundation of descriptivism, even though Frege itself notices some issue that may arise. For instance, different people express or have in mind different descriptions for the same name, e.g., I could understand by 'Antonio' 'my father', while one of his co-workers could understand 'my boss'. Furthermore, my father could get fired, and thus the reference would change, since he would become 'my ex-boss' to his former co-workers.

Apart from these issues, the idea of a descriptive content being the meaning (i.e., the sense) of the name led to a more complete and coherent definition of this theory, i.e., Russel's account.

2.2.2 Russel's account

In order to defend descriptivism, Russell elaborates on Frege's intuitions although taking another route.

First of all, he makes an important distinction between what he calls 'ordinary' proper names and 'logically' proper names. Logically proper names are indexicals such as 'this' and 'that', which directly refer (in a Millian sense, as we saw) to objects of immediate acquaintance. For Russell, ordinary proper names are abbreviated definite descriptions [50], i.e., sentence like 'The 'x'' which could be used to describe Santa Claus as 'The fat, old gentleman with the red cape. . . '.

Thus, a name is just an abbreviation for a definite description. Furthermore, following his own theory for descriptions, definite descriptions (and hence names) have no reference at all and their meanings (in the Fregean sense) are just the truth conditions of their logical equivalent.

This relevance given to definite description stemmed the usually called *famous deed descriptivism*, which claims that we can use salient definition for a name, which narrows the assignment of reference. For instance, 'Aristotle' could be described via 'The teacher of Alexander the Great'. Unfortunately, most of the times these descriptions involve other names, which would need to be broken down accordingly; therefore, this transforms the process in a complex, if not endless, process. Furthermore, it seems unclear what method should be used to decide which description is more salient in respect to the other viable ones. In fact, many entities never actually accomplished anything or acted notably.

All these (plus Frege's) problems found a valid address in the works of Searle, the most passionate defender of this theory to date.

2.2.3 Searle, or the Last Defender of Descriptivism

While first sketching the framework of his theory in the early 60's with *Proper names*, Searle developed it fully in *Speech Acts* at the end of the '70's.

To better situate his thoughts in the general framework of descriptivism, we would like to make a lengthy quote: "Anyone who uses a proper name must be prepared to substitute an identifying description [...] of the object referred to by a proper name. If he were unable to do this, we should say that he did not know whom or what he was talking about, and it is this consideration which inclines us to say that proper names must have a sense, and that the identifying description constitutes that sense" [51].

This clearly proves how close Searle is to the other description theory followers, for he claims that proper names have some kind of meaning or sense, and this meaning is descriptive; more precisely, it is uniquely descriptive in nature. Nonetheless, he still needs to confront himself with the flaws and problems this theory faces since Frege. In order to avoid such problems, Searle proposes a 'cluster theory' of meaning for proper names.

Searle starts by wondering, if we were to gather all the available descriptions of Aristotle, what are the conditions making us say "This is Aristotle?" "[T]he conditions, the descriptive power of the statement, is that a sufficient but so far unspecified number of these statements (or descriptions) are true of the object" [51]. Once we obtain a satisfying (albeit unspecified) set of descriptions, these will function as the truth conditions of the question "Is this Aristotle?". Thus, according to Searle, 'Aristotle' refers to 'Aristotle' not because there is some single identifying description expressing the sense of the name 'Aristotle'. Rather, it is because the entity Aristotle satisfies most or a (relative and context-dependant) sufficient number of the identifying descriptions amounting as the unique referent of the name. Consequently, this move is clearly able to explain why different speakers associate different identifying descriptions with the same name.

Yet, in view of the open question of how many and which definite descriptions are to be considered satisfying for an effective reference, Searle considers his account to be much "looser" than the traditional ones. Overall, many argue that this theory does resolve the more direct problems of the traditional Fregean view, while retaining both the account of the reference relationship and the ability to address its puzzles, e.g., informative identity statements and existential sentences.

2.2.4 Kripke vs. Searle, or the Rise of a Third View

Not long after *Speech Acts* was published, descriptivism faced its most fierce adversary: Saul Kripke. In a trio of lectures later published as *Naming and Necessity*, he listed a series of theses both to represent the theses supported by descriptivism, and to develop three main arguments against it, usually known as: the problem of rigidity (sometimes referred to as 'modal' argument), the problem of unwanted necessity (sometimes referred to as 'epistemic' argument), and the problem of ignorance and error (sometimes referred to as 'semantic' argument).

The 7 theses are listed in lecture I and are the following [31]:

1. To every name or designating expression 'X', there corresponds a cluster of properties, namely the family of those properties ϕ such that A believes 'ϕ X'.

2. One of the properties, or some conjointly, are believed by A to pick out some individual uniquely.

3. If most, or a weighted most, of the ϕ 's are satisfied by one unique object y, then y is the referent of 'X'.

4. If the vote yields no unique object, 'X' does not refer.

5. The statement, 'If X exists, then X has most of the ϕ 's' is known a priori by the speaker.

6. The statement, 'If X exists, then X has most of the ϕ 's' expresses a necessary truth (in the idiolect of the speaker).

7. For any successful theory, the account must not be circular. The properties which are used in the vote must not themselves involve the notion of reference in such a way that it is ultimately impossible to eliminate.

As Kripke notes, "7. is not a thesis but a condition on the satisfaction of the other theses. In other words, Theses 1.-6. cannot be satisfied in a way which leads to a circle, in a way which does not lead to any independent determination of reference" [31] Furthermore, these theses suggest that a descriptivism is "weaker", i.e., its claim are so vague and non committing that they are hardly refutable. 1. clearly refers to Searle's cluster theory, although not stating that the set of properties ϕ is the meaning of X, while 2. stipulates the epistemic position of the speaker. Then, 3. takes the properties in 1. and 2. and turns them into a mechanism of reference, showing the process of reference assignment theorized by descriptivism, and .4 states what happens when no object satisfies the properties. Finally, as Kripke reckons, 5. follows from 1.-3. "and 5. and 6. really just say that a sufficiently reflective speaker grasps this theory of proper names. Knowing this, he therefore sees that 5. and 6. are true" [31].

After dissecting the descriptiont theory, he begins to deploy its three arguments. Firstly, he explains the problem of rigidity. Following Searle example, he considers the name "Aristotle" and the descriptions "The greatest student of Plato", "The founder of logic" and "The teacher of Alexander". Aristotle obviously satisfies all of the descriptions (and many of the others we commonly associate with him), but it is not a *necessary* (hence 'modal' argument) truth that if Aristotle existed then Aristotle was any one, or all, of these descriptions, contrary to thesis 6., for he might have existed and not have become known to posterity at all or he might have died in infancy.

This intuition led Kripke to define names as 'rigid designators'. A rigorous (among many used in his works) definition would be: "a singular term

T *refers to object O at possible world W* iff O is the object that is (semantically) relevant for determining the truth value at W of sentences containing T. A singular term T is a rigid designator iff T refers to the same object with respect to all possible worlds, i.e., it refers to the same individual in every possible world in which that individual exists" [53]. On the other hand, descriptivism doesn't take the multi-world possibility, i.e., doesn't consider counterfactuals, which leads to inconsistencies, e.g., Aristotle died as an infant and people refer to him as "The greatest philosopher of antiquity", which would rather indicate Plato. In other words, since Aristotle could have done other actions, he is not identical to his definite descriptions.

The problem of unwanted necessity (or 'epistemic argument') is rather simple, for it states that, if thesis 5. is to hold, the properties of a name (e.g., Hesperus is visible in the evening) should be known *a priori* by the speaker. In fact, if the meaning of 'Hesperus' is 'the evening star', then 'Hesperus is the evening star' appears trivial and *a priori.* Yet this is not true, as one had to be physically on Earth or *a posteriori* know somehow this to state it. This is also true for historical figures that are distant in the past from the standpoint of the speaker, as the speaker cannot verify whether a property of the referent is true or not.

Finally, Kripke's last problem against descriptive theories (or 'semantic' argument) consists in pointing out that people may associate inaccurate descriptions with proper names. Its famous example considers Kurt Gödel and its proof on the incompleteness of arithmetic, which is probably its most recurring definite description, for many would know him as "The one who proved the incompleteness of arithmetic". Suppose he hadn't proved it, and he stole it from his friend Schmidt, who mysteriously disappears. Following 3., if most, or a weighted most, of the ϕ 's (properties associated with Gödel) are satisfied by one unique object y (Schmidt), then y (Schmidt) is the referent of 'X' (Gödel). This means that " when we talk about 'Gödel', are in fact always referring to Schmidt. But it seems to me that we are not" [31]. Of course, this would force descriptivism to hold a counter-intuitive proposition.

These arguments, together with others from followers of Kripke, led many philosopher (except Searle, who basically claimed that Kripke fell under the 'straw-man' fallacy) to abandon descriptivism in favour of half-way versions of it (e.g., Burge, which claims that proper names work like complex demonstratives, i.e., capable of singling out a salient entity [53]) or causal theories.

2.3 Causal Theories of Proper Names

Whether or not one finds any of the opposing side's arguments sensible or convincing, Kripke tried to avoid the burden of proof by saying that he just wanted to get "a clearer picture", rather than establishing a new theory. Nonetheless, its theory is generally known as *causal theory* and is based, rather than on the descriptive content of a name, on two aspects of reference: reference fixing and reference borrowing. Reference fixing is defined as a "dubbing" [31], generally through perception, even though it could happen via description. Reference-fixing is by perception when a speaker actually performs a naming ceremony (or baptism), e.g., "I call/name/ baptise/etc. you X", on a perceived object. For instance, the name 'Neptune' was fixed by description, stipulated by the astronomer Leverrier to refer to whatever was the planetary cause of observed perturbations in the orbit of Uranus [53]. Once this event is enacted by any of the methods just described, how can someone who is not acquainted with the newly named entity refer to it? Krikpe argues that there is a causal chain that links from the first users of the name to all the possible users. Speakers thus effectively 'borrow' (hence the reference borrowing term) their reference from speakers earlier in the chain. It must be noted that any borrower does not need to identify lenders; all that is required is that borrowers are appropriately linked to their lenders through communication [53]. However, as Kripke points out, there must be an actual intention by the borrower to successfully refer to the entity the lender was also referring to. One may note that, because of his focus on reference, Kripke in on the same line of argument of Mill —names have only reference, and no meaning whatsoever.

Yet this new theory come with a price, i.e., new problems. The most notable one is the lack of effective answers to the phenomenon of reference change. This problem was noted and put forth by Gareth Evans [12], who cites the case of 'Madagascar', unknowingly referred to the African island as 'Madagascar' when the natives actually used the term to refer to a part of the mainland. Evans claims that Polo clearly intended to use the term as the natives do, but somehow changed the meaning of the term 'Madagascar' to refer to the island as it is known today. Furthermore, Evans provides another example to strengthen his criticism: imagine that two newborn babies who, after being baptised, are inadvertently switched. Since nobody finds out about the error, "the man universally known as 'Jack' is so-called because a woman dubbed some other baby with that name" [12].

Michael Devitt, another supporter (albeit with some *caveat*), argues that repeated groundings in an object can account for reference change [10]. In other words, once a sufficient number of groundings in a long period of time happen, the reference will actually change. This proposal seem to suffer the same indeterminate problem that affects Searle cluster theory, i.e., the threshold for the reference fixing is highly context dependant.

2.4 The Present of Proper Names

Despite this century-long debate, there are still many other views surrounding the matter of names. For instance, Kaplan's use of the indexical, which have a clear linguistic meaning other than mere reference (I, for instance means something like 'the speaker of the current utterance'), the two-dimensional semantics, etc...[1] have tried different angles from previous theories, but still no final evidence has been put forth. Furthermore, different flavours of descriptivism, after the Kripke-Searle debate, are trying to overcome the well-known limits of the theory (e.g., the NDT, Nominal Description Theory, by Katz [29]), but they are generally refuted on the basis of being *ad hoc*, rather than trying effectively to create a well founded method to answer the two issues of proper names in philosophy.

Nevertheless, we will see that many points of discussion from the philosophical point of view will be reprised, i.e., chapter 5, section 6.1 and chapter 7, but overall, thanks to philosophy we now have a clearer understanding of the theoretic issues with name. Therefore, we hope that this chapter helped in showing how relevant names are both in academic works and our daily lives. While we did not aim to solve the philosophical quarrel, we want to turn to the pragmatic side of the matter of names and take advantage of the research in computer science, in hope of shedding light on the matter.

[1] See the Name entry at the Stanford Encyclopedia of Philosophy (http://plato.stanford.edu/entries/names/) for an overview on the other theories, in addition to those presented here.

Chapter 3

Which Name in What Context. Names in Sociology and Linguistics

Now that we discussed the issue of names in philosophy one could still feel a somewhat lack of clear definition on a more pragmatical aspect, and that the issues of reference and meaning cannot possibly exhaust the whole range of problems when dealing with names. In fact, philosophy focuses on the inner mechanism behind reference and meaning, but cares little for defining the results of these process. Indeed, even if philosophy calls anything that properly refers to an entity a name, it can be evinced from other areas (e.g., sociology) that names can serve different purposes than simply referring, while other fields (e.g., geography and linguistics) show that names may change according to the linguistic context where they are used.

If we consider sociology, we can see that names can be used to hide the true referent via pointing to a fictitious entity (i.e., alias), or they can highlight a feature of an entity, raising the salience (i.e., relevance of the feature) to the same level of the most used name; for instance, in a familiar context, a person may be referred by a nickname rather than his or her original name. Indeed, if we follow the syntactical definition of name, drawn from the distinction between proper nouns and proper names, we gave in chapter 2, it also states that proper names (and proper nouns) are distinguished from common nouns because, unlike proper name, common nouns do not have an unique referent.

While this feature of proper names allows to the conclude that every name has a referent, it does not follow that the bearer has one single way to be referred. In fact, this assumption would nod consider that many name bearers (be they humans, living or non living things), may be known under multiple names, in addition to their original one. For instance, there is the case of multiple names referring to the same geographic feature but are neither spelling variants of another nor are they related, i.e., bearer that are *polyonymous* (e.g., ‘Holland’ and ‘the Netherlands’ both refer to the same European country). In fact, because of this property, names are said to be *polysemous*, i.e., they can be shared by different entities.[1]

If we recall Frege’s contributions from subsection 2.2.1, this phenomenon was already analyzed when he considered informative identity statements such as ‘Hesperus is Phosphorous’, i.e., two names denoting the same entity [16]. Yet, it must be noted that, regardless of the validity of the argument, this analysis, and those stemmed from it, applies only to cases of the same entity being described differently according to a certain attribute of its, e.g., being a star visible either in the morning or in the evening, as in the case of Venus. In fact, philosophy does not attempt to capture pragmatical aspects of names like, e.g., pseudonyms. In other words, Frege (and philosophy at large) limits itself to definite descriptions, relegating the entity to a rather passive role, without considering sentient choices like choosing an username or a nickname. Moreover, albeit one of the its puzzles, informative identity statements represent a particular and restricted case of issues related to multiple names.

Therefore, we need to take into account other types of names, by going beyond the problems of meaning and reference, thus considering how and why when we say ‘name’, we may refer (be it knowingly or unknowingly) to a particular ‘type’ of name, i.e., a name which is not the original one of the entity, with specific features and context of usage. In fact, philosophy does not have an answer for this issue, whereas other fields of study (e.g., sociology) do [1]. However, since this issue seems to relate to a pragmatical level, i.e., it concerns social and practical aspect of the issue of names, and it appears to be relevant for our research, we analyzed a list of different types of name, to help discriminate names in different contexts. We will illustrate our findings in section 3.1.

On the other hand, there is the other issue, from a different fields of study, that needs to be investigated — name variation in multilingual contexts. In fact, while sociology deals with name *variants*, i.e., other names in addition to the subject’s original one, it does not study the process of name *variations*, i.e., which changes may a name undergo, due to mono- or multilingual contexts. By monolingual context, we mean a variation that happens in one language only, e.g., misspellings, whereas by multilingual contexts we

[1] For instance, [41] notes that Wikipedia contains over 100 people with the name ‘John Williams’.

broadly mean any translation or transliteration affecting the name when being 'used' by a language other than its own, as we said at the beginning of this chapter. Indeed, misspellings do not pose a difficult theoretical problem *per se*, but rather a pragmatical one, which will be thoroughly discussed in chapter 5. On the other hand, multilingual name variations constitute an important issue for names, especially in geography (e.g., as [39], [30], [48] and [28] show); yet, being a matter of languages and interlingual contexts, one could argue that is mainly a linguistic problem.

Nevertheless, the issue has received lot of attention from the geographic community, especially those scholar who study atlases, as there is the need to find a way to standardize the names of geographical features (i.e., cities, mountains, rivers, etc. . .) across multiple languages. In order to do so, they developed a set of concepts that address this phenomenon, and that will be illustrated in section 3.2.

3.1 The Class of Names

In this section we will illustrate the class of name variants, showing the process behind their usage and creation, and also providing a clear distinction among them. Thus, we need to distinguish names *per se*, i.e., the names from which name variants stem, and the actual name variants, which are still 'names' nonetheless. Therefore, we adopt the following terminology: 'name(s)' are defined as *the preferred or official proper noun or proper name used to refer to an entity*, while switching to other classes of names, i.e., name variants, in the other cases.

Instead of uniqueness of reference, usage is our the criterion for discriminating between official and non official names. Consequently, we call the class of non official names *pseudonym*. Pseudonyms, which literally means false name in ancient Greek, can be broadly defined as an "alternative name an entity chooses for a particular purpose, which differs from its original or true name" [1]. This is an overlapping concept with alias, a.k.a., allonym, pen name (or stage name), and nicknames. We are inclined to believe the concept of pseudonym can be used to indicate a class of alternative names because it's the most general one. In fact, pseudonyms differ from the other cases, since these other alternative names are used according to the referent's agenda and depending on the context. For instance, aliases are different from pseudonyms, for in legal cases pseudonyms are allowed as "a way to shield the privacy of rape victims" [35], whereas aliases are illegal "on their generally deceiving feature" [35]. Furthermore, pseudonyms are different from nicknames when it comes to their process of assignment. In fact, while nicknames are generally given by others (e.g., a pet name, an inside joke in a social circle, etc. . .), pseudonyms are self-selected [1]. Pen names and stages name can be considered special cases of pseudonyms in the entertainment area, whereas allonyms indicate a name of another person assumed by someone in authorship of a work of art;[2] this means that allonyms have a somewhat narrower scope than pseudonyms.

To sum up, we elected pseudonym as name of the non official name class, for it has a larger scope than the other types of name; we will now turn into a detailed description of these subclasses.

Alias: while being known under the acronym a.k.a. (also known as), this is the closest synonym for pseudonym, and it is sometimes used as its equivalent. Our motivation for distinguishing it from the latter is that alias tend to denote a willing detachment for the original name, as the individual seeks to "deny any historical connection to the previous name and its corresponding identity" [1], whereas the pseudonym is assumed "with little or no effort to deny the individual's original name, even if the original name is rarely referred to" [1]. Furthermore, allonyms can be considered a special case of alias.
Overall, while in the other cases the purpose of privacy is one of the many possibilities, it seems that this is the sole motivation to resort to aliases.
Some examples of aliases are:

- In legal cases, *John Doe* is used for a party whose true identity is unknown or must be withheld in a legal action [35]
- *The Unabomber*, alias of Theodore Kaczynski used during its trial [35]

Pen name: it may be defined as a different name used in artistic areas in order to conceal the original name or to better market the author's works with an appealing name, with notable equivalent as *stage name* for actors, singers, and entertainers at large.
Unlike an alias, a pen name has a higher level of usage (i.e., it is more salient than the original name), due to its public nature, and can result in a tougher challenge for name matching.
Famous examples of this type of pseudonym are:

- *Mark Twain*, pen name of Samuel Clemens
- *Freddie Mercury*, stage name of Farookh Bulsara
- *George Orwell*, pen name of Eric Blair

[2] http://www.collinsdictionary.com/

Nickname: it may be defined as familiar or humorous name given as a replacement for or addition to the proper name. In other words, a nickname establishes a metonomic relation with the person's name (i.e., takes a feature to refer to the whole being of the name owner), raising and underlining an attribute to reference level.

The first difference from the previous two name classes is that, while they are chosen by the bearer for his or her own reasons, a nickname can be assigned both ways, with a higher chance of being assigned by a third party. In fact, a pseudonym is typically self-selected and "emphasizes aspects of identity that the recipient of the name wishes to make known publicly" [1], whereas a nickname is given by others "to accents members of the community want to emphasize even if the recipients of the nicknames prefer to be reminded of [...] other aspects" [1]. Therefore, it is more prone to stem from a notable feature of the bearer (e.g., 'Fatty' for an overweight person), a shortening of the original name (Australian names are a famous case)[3], personality traits, and so on. Some famous example of nicknames are:

- *Tricky Dick*, nickname given to Richard Nixon during the 1950 U.S. Senate race in California
- *The King of Rock 'n' Roll*, common nickname of Elvis Presley
- *Old Possum*, Ezra Pound's nickname for the British poet Thomas Stearns Eliot

Overall, nicknames tend to be much more cultural and language specific than the other two cases, thus harder to formalize; yet, there have been attempts like [11] that, using various resources, show that name variants may be integrated into a database. Furthermore, [11] notes that nicknames may be associated with hypocorism, i.e., a lesser form of the given name used in more intimate situations, e.g., as a term of endearment or a pet name.

Thanks to these classes, we hopefully shed some light on the inner mechanisms by which a name from the pseudonym class comes to life, and how it can affect the problem of names overall. Drawing from sociology, we found a way to take into account the social factors who make names a difficult subject, beyond the (albeit relevant as well) problems of reference and meaning.

[3] A deeper analysis of the phenomenon in S. F. Kiesling, *Comparative Studies in Australian and New Zealand English: Grammar and Beyond*, edited by P. Peters, P. Collins, and A. Smith, *World Englishes*, Vol. 30, 449–452, (2010)

3.2 Endonym vs. Exonym

Now that we discussed our findings with regard to name variants, let us now turn to the findings on multilingual name variations, i.e., exonyms and endonyms from the field of geography.

An exonym is a name “used in a specific language for a geographical feature situated outside the area where that language has official status and differs in form from the name used in the official language or languages of the area where the geographical feature is situated” [9]. On the other hand, an endonym is “the name of an object in one of the languages occurring in the area where the feature is situated” [9]. For instance, ‘London’ is the endonym of the British capital, whereas its Italian exonym is ‘Londra’. Moreover, the officially romanized endonym ‘Moskva’ is not an exonym, nor is the Pinyin form ‘Beijing’. In fact, while ‘Peking’ is an exonym because it’s an actually different word, the two previous cases are just the result of switching between two different alphabets, thus leaving the name intact. In other words, exonyms are only concerned with languages, disregarding transliterations.

Furthermore, in the literature there is another concept related to these two: the *exograph*, i.e. importing a name from one language (i.e., source language) to another (i.e., target language). For instance, the Estonian ‘Viin’ for the Austrian capital ‘Wien’. We will see that this concept will be fundamental in our approach to tackle the issue of name variation in multilingual contexts

In addition, although mainly used for geographical objects, these terms are commonly used for names of other types of objects:

a) People: e.g., Napoleon is the French endonym of the famous politician, whereas its Italian exonym is Napoleone.

b) Creative works: e.g., Pride and Prejudice is the English endonym of Jane Austen’s most famous book.

c) Events: e.g., First Italian War of Independence is the English exonym of the Italian event Prima Guerra d’Indipendenza.

d) Organizations: e.g., can be considered the endonym of the international organization.

Moreover, because of the efforts of standardization of exonyms used by all the countries, we could look up all the different usage and trend of given exonym by two languages and then derive pragmatic rules (or hints, at the very least) of translation. A work in this direction is [48], where Raukko lists

100 European city names in 8 different languages, and their translated counterparts (be they exonyms, exophones or exographs). Among the findings of this research (including how heavily a language relies on exonyms rather than endonyms, as in Romance ones, or vice versa, as in Scandinavian ones), we wish to point out that Raukko finds a pattern ("The French-Spanish-Italian connection", as he names it) among French, Spanish and Italian exonyms: Italian adds final vowels (e.g., 'Toulon'[fr] vs. 'Tolone'[it]), whereas French deletes them (e.g., 'Barcellona'[sp] vs. 'Barcellone'[fr]; this counts as deletion because the 'e' is mute in French), and tends to add an 's' to its exonyms, basing this hypothesis on the fact many French endonyms end with an 's' (e.g., 'Paris', 'Nantes', 'Poitiers', etc...). In addition, Raukko stresses the importance of English exonyms, for they are the basis for any standardised transliteration and nomenclature (in fact, it is language of the Germanic family that uses exonym the most and more extensively).

Now, if we take a step back from geography, some considerations are in order. Firstly, we saw that this phenomenon of multilingual variations does not limit itself to geographical features, but to many more elements of our daily life. Because of this, it seems that the common reason behind the need of translation is whether an object is 'used' in multiple linguistic communities; in other words, the more an object is important to many languages, the more a language may need to accommodate its name to its phonetic and syntactic rules, thus 'using' the object name effortlessly.

This may explain why we have exonyms, and why they generally apply to capital cities[4], major cities or border cities with important minorities (e.g., Alto Adige or Val d'Aosta in Italy) –because they are famous (i.e., important center of commerce, tourism, political power, etc...). As Michna notes, "if a geographical object has a great importance, it will be better known in its wider environment. Such objects have a great chance of having several names (exonyms) in different language" [39]. Furthermore, claims Michna, it appears that the criteria atlas rely on may be considered similar to the usage criterion: priority to mountain ranges over valleys and lowlands, height (i.e., the highest mountains get exonyms), importance of location (i.e., passes located on major transportation routes or mountain ridges which lie on state border). In other words, "the quantitative criteria applicable to mountain ranges include the area they cover and their elevation. The qualitative criteria are equally important and usually features such as uplands or mountains are named" [39].

Of course, this intuition, although it may shed some light on the matter,

[4] "There are four names on the 100 list [of cities] that each of the eight TL's uses an exonym for, all being national capitals" [48].

carries some issues. Firstly, just like pseudonyms, it is highly context dependant, as the status of usage is based on various factors. In fact, someone (or something) may become widespread, and thus ‘used’, and then become unknown relatively quickly (even though the creation of exonyms underlines a durable status of usage). Secondly, it must be noted that this context dependency sticks to the *object*, rather than the name *per se*. For instance, if we know that the apostle ‘Giovanni’ is ‘Juan’ in Spanish, then why should not all the people named ‘Giovanni’ presents themselves as ‘Juan’ in Spain, and vice versa? In other words, should, e.g., ‘Juan Pane’ and ‘Giovanni Pane’, be considered as names referring to the same person? Plus, not every name ‘used’ in different linguistic contexts is translated, e.g., current famous political figures like US president Barack Obama, whereas older (if not ancient) political figures like ‘Caesar’ seem to generally have their name translated. Thus, we could say that the period of the name bearer may be an additional factor, in addition to ‘usage’. Another evidence supporting this intuition is the fact that classical music pieces (e.g., the ‘Ninth symphony’ by Beethoven) are translated, whereas contemporary songs titles are not.

Regardless of the many issues the intuition behind ‘usage’ leads to, it has one merit: it underlines how relevant and pervasive the issue of multilingual name variations is outside scholarly communities. Because of this, although the other issues from philosophy and sociology are indeed important (and will be explored in this work), and because of the growing interest that this type of name variations is receiving, we believe that multilingual variations are of paramount importance. Thus, they require a thoroughly investigation, while keeping in mind the other issues, e.g. name variants and reference.

Chapter 4

A Taxonomy of Name Variations

In this chapter, we will first introduce our taxonomy of name variations. As we said in section 3.2, multilingual name variations will be an important (albeit not the one and only) issue we will focus in this work. Therefore, we need to provide a detailed description not only of multilingual name variations, but also monolingual ones, i.e., name variations inside a single language. In fact, there is the need to account for them, given their frequency and relevance, even though they have received a considerable amount of attention in the past years; in fact, they have been studied in depth (e.g., [42], [8], [52], and [2]).

Furthermore, the main difference between multilingual name variations and monolingual ones is that, while multilingual name variations actually deal with translation, the monolingual one deal with language independent factors like misspellings and format (i.e., how elements of the name are positioned). Although clearly distinguished, both variations may affect the same name; for instance, 'Jon' is a misspelling (i.e., a monolingual variation) of the translation (i.e, a multilingual variation) of the Spanish name 'Juan'.

In addition, our taxonomy considers also the name variants list in Section 3.1, since they may be subject to name variations, but, given that pseudonyms are explored thoroughly in the appropriate section, we won't list them again here.

Moving to the structure of our taxonomy, we distinguish four main sections:

a) Actual translations, i.e., translations of a whole name, are called *full translations* and are illustrated in Section 4.1. It consists of a single instance, i.e., exographs, a term related to exonyms and endonyms.[1]

b) in addition to whole translations, there are cases where translation does not affect names directly, but rather the other elements of the proper name (e.g., honorifics for person names, geographical common noun for places, etc. . .), i.e., *part-of translations*. Section 4.2 investigates three different cases of part-of translation.

c) A list of all the possible cases of misspellings we could find in the literature is illustrated in Section 4.3.

d) Three possible cases of format variations (i.e., how the elements of the proper name are positioned) are illustrated in Section 4.4.

Let us now take a look at the structure of the taxonomy in more depth.

4.1 Full translation

The only element of this class of name variation is the *exograph*, and it can be defined as "importing a name from a SL (source language) [i.e., the name's original language], and adapting it to a TL (target language) [i.e., the name's final language] phonetic structure" [48]; of course, it is highly language dependent in the degree of variation the name undergoes from the SL to the TL.

To provide a better perspective of the high degree of variance in translation, consider the case of exographs between close languages (Spanish and Italian in **Case 1** and Latin and Italian in **Case 2**), or between distant languages (English and Swedish in **Case 3** and English and Italian in **Case 4**)

Case 1 'Aphrodites'[gr] vs. *Afrodita*[es], *Afrodite*[it]

Case 2 'Titianus'[lat] vs. *Tiziano*[it]

Case 3 'Batman'[en] vs. *Läderlappen*[sv]

Case 4 'Goofy'[en] vs. *Pippo*[it]

First of all, the contribution of the full translation section is that it borrows a strict terminology from the geography field to explain as clearly as possible the process of a full import from one language to another. In addition, an exograph includes the case of transliteration, since it obviously consider the case of switching between different alphabets.

[1] See Section 3.2 for a definition of these terms.

4.2 Part-of translation

Part-of translation helps pointing out that, in many cases, a name can be composed with common or proper nouns, coherently with the definition of a proper name in Chapter 2. Yet, this class of variations underlines another possibility, i.e., that while the proper noun is not translated, the common noun(s) is translated instead. For instance, 'Lake of Garda'/'Garda Lake' [en], where the proper noun 'Garda' passes the translation unto its other syntactical parts, i.e., 'Lake' or 'Lake of'. This may be due to the fact that, not having a translation for the name 'Garda' (but for the rest of the proper name), it tries to help understanding what type of entity it, by translating 'Lake' or 'Lake of'.

Strictly speaking, the common nouns translated may be considered *trigger words*, i.e., "local patterns [...] to recognise names" [46]. In our case, they serve the purpose of indicating, if translated, that we are dealing with a case of part-of translation, and whether the proper noun is translated.

- **Place**: a geographical common noun or the generic term that expresses the character of the object, e.g., hill, valley, sea, and so on.
  The following names are example of Place related part-of translation:
  - 'Lago di Garda'[it] vs. *Lake of* Garda/Garda*Lake*[en], Garda *see*[de].
  - 'Monte Bondone' vs. *Mount* Bondone[en].
- **Person**: honorific, i.e., a word or expression with connotations conveying esteem or respect when used in addressing or referring to a person. We distinguish between 'proper' honorifics and 'common' honorifics or job titles.
  Examples of 'proper' honorifics are:
  - *King*, *Queen*[en], *Papa*[it], *Heilige*[de].

  Whereas examples of 'common' honorifics are
  - *Mr*, *Prof.*[en], *Sig.na*, *Dott.*[it].

  The main difference between the two types of honorifics is that, while 'proper' ones are trigger words for translated names, since their bearer is likely to be known in many different languages, 'common' honorifics do not imply any translation; rather, their are translated into their equivalent, if available (e.g., 'Mr.'[en] vs. 'Sig.'[it]).
- **Organization**: acronyms and company endings.
  Some examples are:

- *inc.*, *ltd.*[en], *s.p.a*, *s.r.l.*[it].

While names of enterprises tend to remain untranslated, except for the case of transliteration, such qualifiers vary according to countries' law and economic structures.

Moreover, unlike the case of exographs, the process of translation applied to the part, i.e., the trigger word, is rather automatic, and easily accounted for by gazetteers [44], i.e.,"dictionaries of placenames [that] contain descriptive information about named places" [21]. In fact, considering the three cases more closely, none of them results relevantly dependant on culture or language factors, because of an overall standardization in all the cases of the part-of translation. In other words, there is an equivalence of titles of enterprise needed for commercial purposes, an equivalence of honorifics in order to facilitate communication in social event, and a low level of discrepancy between languages on geographical terms (i.e., there are few lexical gaps in the geography domain).

Arguably, there can be cases of proper names consisting of both trigger words and names, where they are both translated: in that case, we claim that we are dealing with a full translation *and* part-of translation (although the whole name is translated), since both exograph and trigger words indicate the translation of the proper noun. On the other hand, in the case of part-of translation, only trigger words are translated, and the name is left unaffected.

4.3 Misspelling

Overall, this list tries to cover as many types of misspellings as possible, and it is largely based on classifications found in the literature on the subject (especially [7], [4], and [3]).

- **Punctuation**: e.g., 'Owens Corning' vs. 'Owens-Corning'; 'IBM' vs. 'I.B.M.'
- **Capitalization**: e.g., 'citibank' vs. 'Citibank'; 'SMITH' vs. 'Smith'.
- **Spacing**: e.g., 'J.C. Penny' vs. 'J.C. Penny'.
- **Omissions**: e.g., 'Collin' vs. 'Colin'.
- **Additions**: e.g., 'McDonald' vs. 'MacDonald'.

- **Substitution**: e.g., 'Meier' and 'Meyer'. This also include a type of substitution that [7] defines as 'wrongly typed neighbouring keys' (e.g., 'n' and 'm', 'e' and 'r', etc...); the likelihood of letter substitutions obviously depends upon the keyboard layout.
  Nonetheless, it was treated as a single case together with substitution during the validation step.

- **Phonetic variation**: the phonemes are modified and the structure of a name is changed substantially: e.g., 'Sinclair' vs. 'St. Clair'; 'José' vs. Jose.

- **Switching a pair of neighbouring letters**: e.g., 'Fausto' vs. 'Fasuto'.

As many studies note, misspelling are relevant issues in database maintenance [7], legal cases, and counter terrorism [4]. In addition, among the four types of name variations, misspellings are the easiest to approach via automatic methods, as demonstrated by the large numbers of algorithms devised to tackle them.[2]

4.4 Format

This name variation deals with the positioning (i.e., format) of the name elements, and it generally applies to people names (with the only exception of some type of places, e.g. lakes as in 'Ontario Lake' vs. 'Lake Ontario'). Because of their dependency on people customs, the three following elements are types of variation of varying frequency. In other words, since they are more dependant on the entity (i.e., people) than other variations, their occurrence is variable and limited to certain entities.

Case 1 Switching between given and family name; e.g., 'Fausto Giunchiglia'/ 'Giunchiglia Fausto'

Case 2 Compound names might be given in full (potentially with different separators), one component only [7]; e.g., 'Hans-Peter'; 'SmithMiller'

Case 3 Initials (mainly for middle and given names); e.g., 'Juan Pane' vs. 'J.P'.; e.g., 'Beau Justin Agnello' vs. 'Beau J. Agnello'

[2] For and extensive overview on the state-of-the-art algorithms, see [42]

Arguably, in **Case 1**, the second option is actually the custom in Asian culture, which may lead to confusion when trying to detect the first name or the surname of an Asian person; moreover, **Case 2** is quite rare.

Indeed, **Case 3** (especially its second example) is probably the one with the highest chance of being encountered, as far as format variations are concerned. In fact, it is due to two widespread customs in different areas of the world: *Spanish format* and *middle names*. Spanish Format may be defined as the custom of Spanish speaking countries, with full names (generally used in official occasions) consisting of a given name (simple or composite) followed by two family names (surnames). The first surname is traditionally the father's first surname, and the second the mother's first surname.[3] In addition, there can be one or more middle name between the first name and the surnames. For instance, 'Juan Ignacio Pane Fernández' is the official name, while the general way of addressing is 'Juan' (given name) 'Pane' (father's first surname).

On the other hand, middle names may be defined as names consisting of a first name and a surname, with a name between them (hence middle name). In some countries there is usually only one middle name, and in the United States and Canada it is often abbreviated to the middle initial (e.g. 'James Ronald Bass' becomes 'James R. Bass', which is usually standard for signatures) or omitted entirely in everyday use (e.g. just 'James Bass'). In the United Kingdom he would usually be referred to either as 'James Bass', 'J. R. Bass' or 'James Ronald Bass', or he may choose 'Ronald Bass', and informally there may be familiar shortenings.[4] It may be the case that there are more than one middle name, depending on the custom of the language community of the name bearer.

Finally, there is also the case of authors of scholarly works names, whose format depend on the bibliographic style used to encode the bibliography; in fact, there are several bibliographic style that can change the format of the name.[5]

Style	**Mary-Claire van Leunen**	**Oren Patashnik**	**Charles Louis de la Vallee Poussin**
ieeetr, phjcp, abbrv	M.-C. van Leunen	O. Patashnik	C. L. de la Vallee Poussin
unsrt, IEEE, plain	Mary-Claire van Leunen	Oren Patashnik	Charles Louis de la Vallee Poussin
ama	Leunen Mary-Claire	Patashnik Oren	Vallee Poussin Charles Louis
cj, nar, acm	van Leunen, M.-C.	Patashnik, O.	de la Vallee Poussin, C. L.

Table 4.1: Format Changes Due to Different Bibliography Styles

[3] See the related page on Wikipedia.

[4] See the related page on Wikipedia.

[5] http://amath.colorado.edu/documentation/LaTeX/reference/faq/bibstyles.html

Consider Table 4.1, that shows examples of three different names: 'Mary-Claire van Leunen', 'Oren Patashnik', and 'Charles Louis de la Vallee Poussin', and how the different styles affect their format. Admittedly, some of these changes could also fall under some misspelling cases presented in Section 4.3, e.g., **Punctuation**. Yet, in this case the variation is somewhat standardized, as it follows a certain style, with fixed rules that guide the format, rather than occasional mistakes.

Chapter 5

Problem Statement

In this chapter we will illustrate the issue we aim to tackle, i.e., the task of matching mono- and multilingual name variations and name variants in a distributed scenario. In fact, basing on the issues described in Chapter 2 and 3, i.e., reference and meaning in philosophy, name variants in sociology and name variations in linguistics, we can see how pervasive they are among different fields — even more since the Web and the consequent overwhelming amount of data (and thus names), which is an important topic in computer science.

The area, among many, which deals with the issue of name in the scope of computer science is *name matching*, and it will be the main area of focus in this work. Broadly speaking, name matching is "discovering instances that differ typographically (different surface appearance) but represent the same concept/entity" [54].

Before moving to our framework of name matching, we need to clarify a part of this task definition, i.e., what is an entity? In computer science, an entity is defined as "a 'thing' which can be distinctly identified [. . .] a specific person, company, or event is an example of an entity" [6]; similarly, Hume defines entities as "'things' that exist in the real world" [27]. Furthermore, "the information about an entity [. . .] is expressed by a set of attribute-value pairs" [6]; in other words, *attributes*, and their relative values, represent the properties of entities. Consequently, "entities with the same set of attribute [are] *entity types*" [6] (also called *entity sets*); for instance, all entities belonging to the entity type 'Book' will have the same attributes, e.g. 'Ubik' will have attributes like 'Author', 'Genre', 'Year', and so on.

Now we will define the scenario itself. Its architecture is based on a P2P network of users, whose structure consists of three level (or layers):

Local: the level of users, who have their own personal repository of entities.

Community: the level where a certain number of users, i.e., a community, can share their entities.

Global: the 'universal' level, i.e., a *global* repository of all the entities.

Because of this structure, entities are not simply stored in centralized system, rather they are *distributed*. In fact, since every user has a repositories of entities, the same entity may be present in multiple repositories, with different possible attributes values. In other words, every entity at local level is a *partial* view of the real word one, that can be found at global level, which holds the 'universal' view of entities.

Since every entity has a name, name matching is needed at all levels, because, regardless of the level where the matching is performed, the name of an entity is an important query; thus, the problem scenario is:

1. A user queries an entity name.
2. The system has to match two strings, and must recognize that they both refer to the same entity.

Considering that the only other mandatory information available about the entity, apart from its name, is the entity type it belongs to, how do name variations and name variants represent an issue for this task?

Name Variants They are *actually different* names, underlining another view of the entity, so they may be very distant from the original entity name.

Name Variations The issues vary depending on the language context(s)

- *Monolingual Name Variations* Variations such as misspellings and format changes,[1] while adding or searching for entities names are frequent and pervasive, slightly changing the name orthography.
- *Multilingual Name Variations* They change the 'surface' of a name to the point of being unrecognizable — even worse when it comes to different alphabets (i.e., the case of transliteration [46]).[2]

Furthermore, because of the structure of our scenario, the importance of both name variations and name variants changes with the level. In fact, on local level, since a user is likely to generally have entities in his or her own

[1] See Section 4.3 and 4.4 for an overview on the cases of monolingual variations

[2] We must point out that transliteration is a peripheral problem for our scope, since the only alphabet used in our scenario is the Latin one.

language(s), monolingual variations and name variants are more common instead of multilingual variations. On the other hand, tables turn when considering global level, where multilingual name variations are predominant.

Plus, name variations and name variants represent an issue not only when querying, but also when managing the system. In fact, when matching, both the computational power required to perform the task (i.e., the CPU), and the disk (i.e., the amount of memory where entities are stored) must be considered. In fact, while both depend on the machine available to users on local level, at the global level, since multilingual variations and variants require, e.g., name dictionaries,[3] they need to be stored; thus, searching through them for candidate strings is computationally expensive, and may slow down other processes.

Furthermore, there is the need to design the architecture of the entities names, in order to account for the different types of names, be they variants or variations, that are to be stored; plus, in the case of name variations, there is the need to indicate the language of each variation. The main issue is whether it is worth distinguishing also in practice variants and variations, since they may happen at the same time, as other architectures, e.g., the ontology YAGO2,[4] do not distinguish between variants and variations; in fact, apart from 'hasPreferredName', all the possible alternatives fall under the 'isCalled' attribute, without any further distinction.[5]

[3] See Section 4.2 and 6.3 for definition and usage of named dictionaries in the literature, respectively.

[4] See [22] for an exhaustive overview of YAGO2 ontology.

[5] For instance, see 'Napoleon' at https://d5gate.ag5.mpi-sb.mpg.de/webyagospotlx/

Chapter 6

Related Works

Before illustrating our proposals for tackling the issue of multilingual name variations, we will study the approaches in the field of name matching dealing with names and give an overview of the related works and fields. Firstly, we will provide a general definition of name matching and see how it differs from our view of the task in section 6.1, then we will compare it with three other field that deal with name in computer science: named entity recognition (NER), named entity disambiguation (NED), and entity linking — dedicating a section to every one of them, stating the difference between them and our approach to name matching, plus showing how we took advantage of these field's findings.

6.1 Name Matching

As presented in chapter 5, name matching may be broadly defined as "discovering instances that differ typographically (different surface appearance) but represent the same concept/entity" [54]. This need for matching different names, disregarding complex theoretical issues like the philosophical issue of meaning, comes from the early 60s in the area of *record linkage* [15]. Basically, the idea behind this task is to tackle the problem as a "classification problem, where the basic goal is to classify entity pairs as matching or non-matching. Fellegi and Sunter propose using largely unsupervised methods for this task, based on a feature-based representation of pairs which is manually designed and to some extent problem-specific" [2].

As [8] notes, other communities, namely the database and artificial intelligence communities, took advantage of the findings in such area. While the single methods developed vary [2], overall they supported the development of more autonomous systems rather than methods for human experts, via algorithms.

In order to provide a better picture of how these algorithms work, we will follow the approach by P. Christen [7], who divides name matching algorithms in two categories: *phonetic encoding* and *pattern matching.* Algorithms belonging to the first category “attempt to convert a name string into a code according to how a name is pronounced (i.e. the way a name is spoken)” [7], whereas those belonging to the second one try to see how close the two string of characters are by calculating a “normalised similarity measure between 1.0 (strings are the same) and 0.0 (strings are totally different) [...] For some of the techniques, different approaches to calculate such a similarity exist” [7].

Algorithms based on phonetic encoding are, e.g., *Soundex*, the original version of both *Phonex* and *Phonix*, which normalize the name by keeping the first letter in a string and converting the rest into numbers according to a conversion table (which substitutes letters with numbers), obtaining a code that represents the name, e.g., ‘Peter’ is p360. Its main flaw, as [7] notes, is that “it keeps the first letter, thus any error or variation at the beginning of a name will result in a different Soundex code”. Furthermore, most of the phonetic encoding algorithms were designed basing on English phonetics, thus needing to be tuned to whichever language the names to be matched belong to, as [45] and [7] note; thus, they are not scalable for multilingual contexts. In fact, considering our framework, we stated in Chapter 5 that users at local level are more likely to have a limited set of languages (and thus translated names), with more entities name translated in known languages, which may not be English. As for global level, there are of course too many languages to consider implementing every language phonetics viable.

On the other hand, pattern matching algorithms do not incur into such issues, as their method is language independent. The first one devised is the *Levenshtein* (or *Edit distance*), “defined for strings of arbitrary length and counts differences between strings in terms of the number of character insertions and deletions needed to convert one into the other, the minimum edit distance is then the similarity” [52]. Another widely used algorithm is Jaro, which “accounts for insertions, deletions and transpositions. The algorithm calculates the number c of common characters (agreeing characters that are within half the length of the longer string) and the number of transpositions t” [7]. The similarity measure is calculated as:

$$sim_{jaro}(s1, s2) = \frac{1}{3}(\frac{c}{|s1|} + \frac{c}{|s2|} + \frac{c-t}{c}) \tag{6.1}$$

In addition, there is the *Winkler* algorithm, that builds on Jaro and improves it “by applying ideas based on empirical studies which found that fewer errors typically occur at the beginning of names. The Winkler algorithm therefore increases the Jaro similarity measure for agreeing initial characters" [7]. The similarity measure is calculated as:

$$sim_{wink}(s1, s2) = sim_{jaro}(s1, s2) + \frac{s}{8}(1.0 - sim_{jaro}(s1, s2)) \tag{6.2}$$

Yet, while these measures may effectively tackle misspellings, performance would drop if they are faced with format variations (e.g., swapped words, as in 'Fausto Giunchiglia' vs. 'Giunchiglia Fausto'). In fact, the edit cost to convert one or the another would be very high, thus making the system deem the two names not matching. To account for this, algorithms like *Longest Common Sub-string* (LCS), that "finds and removes the longest common sub-string in the two strings compared, up to a minimum lengths" [7], were developed. This overview further proves our claim that monolingual name variations have received a lot of attention in the past years and there is a wide state-of-the-art literature on the subject.

Now that we saw the algorithms, it is clear that these methods, and name matching at large, deal with strings, i.e., sequence of symbols and/or digits. Consequently, Branting describes name matching as "the task of recognizing when two different strings denote [i.e., match] the same [...] entity" [4]; similarly [7] defines name matching as "the process of determining whether two name strings are instances of the same name". Furthermore, Branting considers name matching an instance of the general problem of approximate string matching, i.e., "determining whether the edit distance between the pattern [a given name] and target [strings in which the patterns are sought] is less than a given mismatch threshold. The edit distance between two strings consists of the number of insertions, deletions, or substitutions required to make the pattern equal to the target" [3]. Indeed, here we see the issue of multilingual name variations illustrated in chapter 5, i.e., that relying on automatic methods for translation is not viable, as, if we follow Branting's definition, we would need to tune the mismatch threshold for every possible language in our system.

Turning back to Branting's definition, he further clears it by dividing the process of the name matching task into two steps:

1. It starts with one or more pattern strings (usually proper names or proper nouns) and a collection of target strings (where the patterns are to be sought).

2. It searches targets that match a pattern so closely, that both are likely to denote the same entity.

Overall, the name matching task, as [43] illustrates, does not limits itself to computer sciences, but it is also used especially in computational biology and many other areas, such as law enforcement, and then expanded and played an important role in text understanding, e.g., the co-reference resolution problem [4].

Now, if we take a step back from computer science, this what philosophy would call finding if two entities have the same referent. Furthermore, if we look at name matching from a philosophical point of view, we could say that, since it works with strings, the semantic value (if any) of the name is not taken into account. Therefore, name matching operates only on reference, considering names simply as labels identifying entities; of course, there may be the case of multiple labels for the same entity (i.e., pseudonyms), which, as we said in chapter 5 poses quite a challenge. Thus, we can already see that name matching considers names alongside Millianism and causal theories, i.e., simple labels to refer to entities, and, although being affected by the diversity of possible pseudonyms, name matching only operates on the surface of names.

Our main difference with regard to scope of the task of name matching is that we focus on very distant matches, as both name variations and variants can change a name considerably. Therefore, we cannot completely rely on algorithms available in the literature, but need to account for the language and context dependant factors, by designing name matching algorithms that also use name dictionaries of name variations as candidates for the matching. Plus, the process of matching is exclusive, either the string is matched or it is not — in our framework, we take it a step further. In fact, while on local level the name matching is expected to return true or false, i.e., the entity names exists or not, in a user repository, on a global level our approach to name matching requires, rather than a positive/negative result, to obtain a list of possible candidates, and (ideally) a perfect candidate; thus, the task looks more like *name searching*. Furthermore, name matching simply considers strings, regardless of any other information, whereas our approach considers by default the entity type of the entity whose name its being matched, and, as we will show in Chapter 7, relies on external resources for creating a list of name candidates, i.e., name dictionaries, while standard name matching does not.

6.2 Named Entity Recognition (NER)

In this section, we will introduce a field close to our approach to name matching, i.e., *named entity recognition* (sometimes referred to as NER). Named entity recognition is defined as “the identification of proper names in text and their classification as different types of named entity, e.g., persons, organizations, locations" [14], and it is a task of Natural Language Processing (NLP).

By its definition, we can already see two main differences between NER and our approach to name matching. Firstly, NER deals exclusively with texts, as its task is to find names and categorize them, whereas our task operates with other type of frameworks. In other words, while NER needs to discover names in text, generally by picking up text clues [46], in our case name matching takes for granted that the two strings to be matched are already names. Secondly, NER indicates both the process of finding names in a text and classifying them in the appropriate category, whereas our idea of name matching works backwards with respect to NER. In fact, as stated in Chapter 5, the categorization happens *before* the process of matching, since our system requires to assign an entity type, i.e., a category, to every entity present in the system.

The main contributions that we draw from NER is a type of words that function as clue when trying to find names— trigger words, i.e., "local patterns [...] to recognise names" [46]. In other words, they serve the purpose 'triggering' the system, so that it understands that the next word(s) before or after this type of words is a name, and belong to a certain type of entity. For instance, consider the following names: 'FIAT s.p.a.', 'Monte Bondone', and 'Sig.Enrico Bignotti'. These three names consist of both proper nouns (i.e., 'FIAT', 'Bondone', 'Enrico' and 'Bignotti') and common nouns (i.e., 's.p.a.', 'Monte', and 'Sig.'); as [46], [24], and [49] show that these (and many more) types of common nouns both help understand in the text that the nearby words are name and what type of words they are (e.g., 'Sig.' indicates that the following word is the name of a person).

Yet, even though trigger words may be present in a name (as one string to match could be, e.g., 'Mr. Enrico Bignotti'), in the case of name matching we obviously don't need trigger words to understand that the string is a name, since that is taken for granted for this task. Rather, in this work we devised trigger words not as simple indicators of nearby names (and their category), but as triggers of *translations*, whether on themselves or on the proper noun/name, i.e., as parts of the part-of translations, described in section 4.2.

This use of trigger words is an intuition we built upon independently from [44], that does not explicitly consider trigger words, but underlines the possible modification of parts of the name when translated from French to German (and vice versa), without any further formalization. Therefore, based on the literature available, our usage of trigger words seems to be relevant and novel.

6.3 Named Entity Disambiguation (NED)

Named entity disambiguation (commonly referred as NED) is "the task of mapping mentions of entities in a text with the object they are referencing" [18]. In other words, the purpose of NED is to find which name *variant* is referring to an entity, e.g., understanding that both 'the king of Rock 'n' Roll' and 'Elvis Aaron Presley' refer to the singer Elvis Presley, being a nickname and the full middle name, respectively.

NED recognizes if a name is a variant of an original name, basing its decision on the context given, i.e., a window of words in a text which hint to what entity is being referred [13], and lists of candidates (usually called *disambiguation dictionaries* [5], [13], [56]), containing a set of suitable names either to match (as in name matching) or to be a name variants (as in NED). Thus, the difference between our view on name matching and NED is that, while they both rely on dictionaries of names, NED also needs the " local contextual evidence" [13], i.e., context, in which the name appear. Furthermore, NED generally tackles name variants, without considering name variations.

To sum up, the standard NED system approach follows these two steps [5]

1. "Detects whether a proper name refers to a named entity included in the dictionary (*detection*)."

2. "Disambiguates between multiple named entities that can be denoted by the same proper name (*disambiguation*)."

Relying so heavily on dictionaries, NED system need large and constantly updated sources. Thus, there is a growing interest in the area of NED towards the encyclopedic knowledge obtainable by Wikipedia[1], as the work of [5] shows. Basically, the idea is to use Wikipedia to create a dictionary of named entities to disambiguate them, thus establishing a baseline method for further works, e.g, [13]. Following the relevance [5] assigns to Wikipedia, we decided then to use it as a dataset for the validation of our schema of name variations, obviously disregarding the need of encyclopedic knowledge for disambiguating. Since we couldn't obtain the original of [5], we opted for the disambiguation dictionaries developed using Wikipedia by [56], which also created a name dictionary of name translations to be used for their system, HeiNER.[2] While "the Translation Dictionary contains more than 1,5 million name entities encountered in the English Wikipedia and its translations

[1] http://en.wikipedia.org/wiki/Wikipedia

[2] http://heiner.cl.uni-heidelberg.de/index.shtml

into the 253 languages available in Wikipedia"[3], there are disambiguation dictionaries "in 16 languages for each Named Entity".[4]

Since name variants are extremely important for NED, some works in this area try also to give a more general account on names, as we do in this work; one notable example is [55]. First of all, it is one of the few works in the literature which clearly states the syntactic structure problems of name, underlining how proper names (PP) ambiguous structure, since "PP may be attached to the preceding NP[5] and form part of a single large name, as in NP[Midwest Center PP[for NP[Computer Research]]]" [55] or "it may be independent of the preceding NP, as in NP[Carnegie Hall] PP[for NP[Irwin Berlin]], where *for* separates two distinct names, Carnegie Hall and Irwin Berlin" [55]. This structural changes render the name ambiguous, since "the exact boundaries of a proper name" [55] (i.e., where the proper name start and end, e.g., 'Hebrew University in Jerusalem, Israel').

Then, [55] goes on and claims that name also show "semantic ambiguity", i.e., name variants, also considering *namesakes*, i.e., entities named "after famous people" [55] (e.g., the teddy bear, named after Theodore Roosevelt), which we did not consider, as it is a particular kind of pseudonym, very close to allonyms from a theoretical point of view (i.e., in both cases, the name is not completely new, but imported from another entity), and which would be already disambiguated in our scenario, as every entity belong to a specific type.

6.4 Entity Linking

The last field to study is *entity linking*, that "describes the task of matching references to named entities found in natural language texts to a unique identier, denoting a specic entity" [32]. Unlike the other two fields of NER and NED, entity linking differs from name matching not because it considers different problems of names, e.g., finding their name variants as NED does, but because name matching can be a task paired with entity linking, since the former is the step before referring the query name to an element in the [33]. In other words, name matching is a subtask of entity linking.

From this field, we did not import any practical finding, rather a theoretical one that strengthen our intuition illustrated in Section 3.2, i.e., translation of names because of usage in multilingual communities. In fact, general entity linking works like [57], [38], [47], are overall more focused on matching name

[3] http://heiner.cl.uni-heidelberg.de/doc.shtml

[4] http://heiner.cl.uni-heidelberg.de/doc.shtml

[5] See the definition of noun phrase (NP) in Chapter 2.

in the scope of monolingual variations (e.g., format change, misspellings, etc...), while our main purpose is to tackle the problem of multilingual name variations. In fact, their account on monolingual variations is not different from ours, as our schema of variations introduced in chapter 4 mainly relies on heuristics, whereas theirs is automatic, especially in [57] and [47], since it is based on Wikipedia. For instance, Wikipedia redirect pages are used to account for alternative names.

Furthermore, while we suggested the possibility of an 'usage' criterion behind the translation of names in section 3.2, also both [38] and [47] (expanding from an intuition in [13] for NED) developed a similar notion, called 'prominence', in order to "to estimate measures of popularity" [47], by counting " the number of in-links, the number of out-links, and the page length (in bytes)" [47]. In addition, unlike [13] (which limits itself to Wikipedia), [38] also "submitted the query string to Google and used the rank of Wikipedia pages in the Google results as an attribute for their corresponding entity". Thus, the main difference between 'being famous' and 'prominence', is that the latter is limited to names in a single language, whereas the former accounts for names in multilingual contexts; nonetheless, both underline how the context (in Wikipedia, page links) affects the name of the entity.

Chapter 7

Proposed Solutions

In this chapter, we will introduce our solutions to the issue presented in Chapter 5, i.e., the task of matching name variation and name variants in our scenario, recalling that by name variants we mean other names referring to an entity in addition to its original name, i.e., pseudonyms, while by name variations we mean changes to a name's orthography in multilingual, e.g., translations, or monolingual, e.g., misspellings, contexts.

Firstly, we provide a taxonomy of name variation in chapter 4, which draws from both personal research and state-of-the-art literature, and is divided into four sections. Section 4.1 accounts for full translation (i.e., importing one name from one language to another), relying on a geographical term related to exonyms and endonyms — the *exograph*. Section 4.2 also deal with translation, but it addresses the case of translation of 'parts' of the name, i.e., translation of common nouns instead of proper nouns, e.g. 'Lake of Garda'[en] and 'Lago di Garda'[it].

While Section 4.1 and 4.2 actually deal with translation, Section 4.3 and 4.4 deal with language independent factors like misspellings and format (i.e., how elements of the name are positioned). We would like to point out that, as for Section 4.1, our decision to import the concept of exograph together with its related concepts, which received much attention in the literature (as [39], [30], [48] and [28] show), is novel in the field of computer science. In addition, Section 4.2 imports trigger words from NER, deploying them for new purposes, as illustrated in chapter 6. On the other hand, methods for tackling misspellings and format were taken directly from the literature, as showed in Section 6.1, since they are the two 'automatic' sections of the schema.

Our taxonomy helps categorizing name variations but admittedly does not indicate what approach to follow for tackling them. Thus, what should the system do in order to match them correctly? We propose to tackle multilingual name variations with the aid of named dictionaries, and to tackle monolingual name variations by using some of the algorithms discussed in section 6.1. Moreover, we propose to import named dictionaries from various lexical resources (e.g., HeiNER), or to build upon already existing resources. In our vision, named dictionaries should be divided by full translations of names or part-of translation, as trigger words can help by narrowing down the possible matching strings and speeding up the task of name matching. On the other hand, choosing which algorithms is more suited to tackle monolingual name variations is a matter of what type of names the system deals with (e.g., a system who has few cases of person names may not need to care so much about format variation like swapping first name and surname).

Secondly, since we also need to deal with name variants and multilingual variations, we aim to modify the entity name structure by dividing it in different entries, where the various types of names will be stored.

We will illustrate the details of these contributions throughout this chapter, in addition to our proposals of implementation in our framework.

7.1 Validation from Schema.org

Firstly we need to validate our taxonomy, i.e., confirm that our classification of name variations is correct and works. To do so, we consider schema.org, i.e., "a collection of schemas that webmasters can use to markup [the content, i.e., entities of different types] their pages in ways recognized by major search providers"[1] Thus, it aims for a general, yet accurate enough, coverage of all possible entity types that can be encountered on the Web; so, it seems a good place to validate our taxonomy. The structure of schema.org is as follow: "a set of types, arranged in a multiple inheritance hierarchy where each type may be a sub class of multiple types". In other words, types are like entity types described in Chapter 5, but here there are five 'main' types from which the rest of the types stem: CreativeWork[2], Person, Place, Event, and Organization.

Starting from this model, we check all 860 types (the full list can be found in the appendix B), in order to find which ones could undergo name variations. By checking, we mean that for each type we consider via heuristics if (and how) their name could change because of name variations prioritizing the multilingual ones, and discarding those whose name variations consisted

[1] http://schema.org/

[2] This formatting on type names without spaces is imported from schema.org, and will be used in this work to indicate an entity type

only of monolingual name variations, as this would have forced us to choose every type.

Thus, we obtained the following list of 22 candidate types:

1. CreativeWork
    - Book
    - Movie
    - MusicAlbum
    - MusicRecording
    - Painting
    - TvSeries
2. Event
    - Festival
    - MusicEvent
    - SocialEvent (e.g., awards, wars, etc. . .)
    - SportEvent
3. Organization
    - Corporation
    - Organization
    - EducationalOrganization
    - SportsTeam
4. Person
    - Religious Figures and Monarchy
5. Place
    - City
    - Country
    - LandmarkOrHistoricalBuilding (e.g., Eiffel Tower; this entity type is also called LOHB)
    - Landform

We wish to point out two facts about this list. Firstly, as for 'Organization', schema.org distinguishes between 'NGO', i.e., Non Governmental Organization, and 'GovernmentOrganization', while we merge them in 'Organization' and add 'Corporation'. In fact, both NGO and GovernmentOrganization are affected by full translation and their difference in terms of attribute is low, whereas Corporation is the only case of part-of translation, allowing us to capture more entities. Secondly, Person type lack any subtype in schema.org whereas we add 'Religious Figures and Monarchy'. This subtype names are the only one of the entity type Person that falls constantly under both multilingual name variations, e.g., they have proper honorics like 'King', whereas other names are less consistent and hard to formalize (e.g., ancient artist like 'Titianus'[lat] vs. 'Tiziano'[it], while more recent artist like 'Picasso' are not translated).

Now that we have the types that show name variations, there is the need to find in detail which variations every type is more prone to fall under. Thus, we choose 15 random instances (i.e., entities) of every type, then check, for every name variation, whether it applies to the instance; the full list of tables of entity types can be found in Appendix B. For every table, the rows represent the 15 entities chosen, while the columns represent the name variations the entities may be affected by.

To sum up the trends of both name variations and name variants, we have Table 7.1, where the rows, instead of representing the 15 random entities, represent all the selected types from the schema.org list, while the columns represent the name variations that they may show.

Moreover, we would like to add some clarification for every column of the tables, i.e., both Table 7.1 and those in Appendix B.

Full Translation: this column represent all the translations, that are based on Wikipedia[3] multilingual links.

Part-of translation: in this column, because of space and formatting reasons, we just put the trigger word(s) of the name underlined.

Alternative Names: this column was added to also take into account pseudonyms for the sake of completeness. Moreover, although they are name variants and not name variations, they may be translated.

Yet, it could be that 'Alternative Names' may be the official name of the entities (i.e., the entity original name). In fact, since we opted for salience as the main criterion for deeming the name 'Alternative' or not, as explained in Section 3.1, it may be that the official name is less

[3] http://en.wikipedia.org/wiki/Wikipedia

Types	Class of Name Variations											
			Misspellings									
	Full Translation	Part-of translation	Punctuation	Capitalization	Spacing	Omissions	Additions	Substitution	Phonetic variations	Switching letters	Alternative Name	Format
Book	√		√	√	√	√	√	√		√	√	
Movie	√		√	√	√	√	√	√	√	√	√	
MusicAlbum			√	√	√	√	√	√		√	√	
MusicRecording	√		√	√	√	√	√	√		√	√	
Painting	√	√	√	√	√	√	√	√		√	√	
TVSeries	√		√	√	√	√	√	√		√	√	
MusicEvent			√	√	√	√	√	√	√	√	√	
SocialEvent	√		√	√	√	√	√	√	√	√	√	
SportEvent	√		√	√	√	√	√	√		√	√	
Festival	√		√	√	√	√	√	√		√	√	
Organization	√		√	√	√	√	√	√		√	√	
Corporation		√	√	√	√	√	√	√	√	√	√	
SportsTeam	√		√	√	√	√	√	√		√	√	
EducationalOrganization	√	√	√	√	√	√	√	√	√	√	√	
Person	√	√	√	√	√	√	√	√	√	√	√	√
City	√		√	√	√	√	√	√		√	√	
Country	√		√	√	√	√	√	√	√	√	√	
LandmarkOrHistoricalBuilding	√	√	√	√	√	√	√	√	√	√	√	
Landform	√	√	√	√	√	√	√	√		√	√	√

Table 7.1: Name Variations and Types

used that another one, although this is not a fixed occurrence.
To prove our claim, we decided to submit both names as queries to Google and Wikipedia, then labelling as 'Alternative' the one with the fewest occurrences. For instance, although 'Alice's Adventures in Wonderland' is the original name of the book by Carroll, 'Alice in Wonderland' is by far the most used (69.900.000 vs. 912.000 results on Google); since 'Alice's Adventures in Wonderland' is less used, it becomes a value of 'Alternative Name'. On the other hand, 'The Big Apple' cannot substitute in salience 'New York' ($1{,}910^9$ vs. 310^9 results, respectively).

Misspellings: because of space and formatting reasons, we just put one of the possible misspelled parts of the name.

Format: because of space and formatting reasons, we only listed one example of format variation, as the whole list can be easily obtained automatically, which may be if, e.g., one follows a certain formatting style for authors' name.[4]

Furthermore, the language considered for all the instances in the tables, thus for both Table 7.1 and those in Appendix B, are the EU official ones, i.e., English, German, French, and Spanish, plus Italian and Danish; nevertheless, the main languages considered are English and Italian. If an entity comes from either of the languages, we use the other one for translation (if available). On the other hand, if the entity comes from any other language than English and Italian, English becomes the language for translation; otherwise, we switch to Italian.

In conclusion, thanks to schema.org we have now a validation that our taxonomy works in capturing name variations, and we can also restrict the number of entity types to which the variations apply. Plus, by testing for each of the 22 selected entity types, we can obtain a pattern in the way they are affected by name variations. These patterns can be used also as guidelines to which strategy use to tackle them, e.g., relying on named dictionaries or choosing specific algorithms that address format changes in the case of ,e.g., Person entities names.

[4] See Section 4.4 for an overview of the possible bibliography style, and how they change the name.

7.2 Entity Name Architecture

In Chapter 5 we also stated that we need to consider how to 'store' them in the entity, i.e., finding a way to modify the entity architecture in order to represents its name variations and name variants. In fact, it goes against what we claim in Chapter 5 and logic to create separate entities that are different because the name of one is a pseudonym or a translation of the name of the other; for instance, 'Freddie Mercury' is the stage name of 'Farookh Bulsara', but, having overall the same attributes, should one create two separate entities? We believe not.

7.2.1 'Name' Design

Given our Chapter 3 and 4, we could go for a three-parted rendition of the entity name, with three entries: 'Preferred' (the most salient name of an entity), 'Alternative Name' (name *variants*), and 'Other'(name *variations*).

Although theoretically sound, this division does not appear to also work on the pragmatic side. In fact, we said in Chapter 5 that both name variations and name variants cannot be tackled automatically, given their dependency on language and context, rather we need to take advantage of named dictionaries for both of them. Therefore, there is no difference in the approach to tackle the issue between name variants and variations — it's only a difference in what the name dictionaries contain, i.e., pseudonyms or translations. Thus, this distinction would make sense for the user (clearly distinguishing a nickname from a translation), but not for the system, that would be forced to keep them apart, but has to use the same resources (e.g., name dictionaries).

Then, similarly to YAGO2 name attribute 'IsCalled', we need to merge 'Alternative Name' and 'Other' in *Other Names*, that will contain both name variations and name variants (plus any translation of name variants, e.g. 'la Grande Mela'[it] vs. 'the Big Apple'[en]), while still keeping 'Preferred' to represent the most salient name, following the saliency criterion used for alternative names in the tables from schema.org types in section 7.1. Yet, while all multilingual name variations and name variants should be kept with the entity, along with their target language, monolingual name variations can be treated automatically via algorithms at runtime[5] — thus, it would not be viable to store all of them. Although some format variations are fixed, e.g., bibliography styles, some of the algorithms available in the literature are especially designed to address format variations, e.g., Longest Common Substring [7].

[5] See the overview of algorithms in Section 6.1

To sum up, ‘Preferred’ and ‘Other Names’ should be used in the architecture following this guidelines:

Preferred This represents the most salient name of the entity, i.e., the most used name in as many context as possible; nevertheless, it does not favour name variants. For instance:

- CreativeWork: *Frankenstein*[en]; *Pride and Prejudice*[en]
- Event: *Second World War*[en]; *2008 Olympics*[en]
- Organization: *UN*[en]; *Inter*[it]
- Person: *Enrico Bignotti*[it]; *Freddie Mercury*[en]
- Place: *Roma*[it]; *Lago di Garda*[it]

Other Names This entry represent both name variants and multilingual name variations, while it leaves the monolingual ones to be treated automatically. For instance, showing both name variations and name variants:

- CreativeWork:[6] *Frankenstein; or, The Modern Prometheus*[en] (less salient name); *Orgoglio e Pregiudizio*[it] (full translation)
- Event: *WWII*[en] (less salient name); *Olimpiadi 2008*[it] (full translation)
- Organization: *United Nations*[en] (less salient name); *InterMilan*[it] (full translation)
- Person: *Bigno*[it] (nickname); *Farook Bulshara*[en] (less salient name)
- Place: *La Città Eterna*[it] (nickname); *Garda Lake*[en] (part-of translation)

7.3 Solutions Implementation

Having discussed in detail our taxonomy of name variations and the entity architecture to accommodate them, we will give a deeper overview of our scenario, and how we propose to solve the issue of matching name variations and name variants with our solutions.

Before that, let us briefly recall our framework. As we presented in chapter 5, our scenario is a P2P network of users, consisting of three levels: local,

[6] Admittedly, ‘Other Names’ could be changed to ‘Other Titles’ in this case

community and global. Local level represents the user level, i.e., the private of a user, that stores the user's entities; these entities may be share on community level, i.e., the level of communities, groups of users that share certain entities (e.g., a community of music will share entities related to music). In addition, there is the global level, i.e., the global view of all the entities that exists in the network,

One way to index the entities is via a *distributed hash table* (DHT), i.e., "a hash table which is distributed among a set of cooperating computers, [called] *nodes*[...] It contains key/value pairs, [called] *items.* The main service provided by a DHT is the *lookup operation*, which returns the value associated with any given key" [19]. Usually, a user may want to find the value of a key, so the DHT "provides the key to any one of the nodes, which then performs the lookup operation and returns the value associated with the provided key" [19]; in addition, "a DHT also has operations for managing items, such as inserting and deleting items" [19]. In our case, the DHT does not act as a global repository for entities (as one cannot 'store' them in it), but rather as a*global index*, holding the identifiers (i.e., keys) of the entities.

Moreover, we can distinguish two types of identifiers: a local one and a global one (GUID). As stated in Chapter 5, every user has a repositories of entities at local level, thus holding a *partial view* of the entities; this view is represented by a link to their position in the repository, i.e., a local URL. On the other hand, on global level the GUID acts as a rigid designator,[7] i.e., denotes the entity in every possible context (or possible word, in philosophy), thus representing the *real world* entity.

For instance, Fausto Giunchiglia is my advisor, and he has children, therefore he is their father; consequently, I call him 'Prof. Giunchiglia, while his daughter calls him 'Dad. Therefore the entity representing him will have two local URLs, redirecting either to my repository in the case of 'Prof. Giunchiglia or to his daughter otherwise, and one GUID identifier at global level, which will represent Fausto Giunchiglia *per se.*

As we stated in Chapter 5, the different scenarios where the task of matching is to be done is also affected, in addition to name variations and name variants, by the different levels and their interplay between themselves. Now that we provided a clear view of the framework where to implement our solution, we will proceed to explain how to implement our solutions.

[7] See Kripkes definition of rigid designator in the philosophical debate surrounding names in Section 2.2.4.

7.3.1 Local Level Implementation

First of all, the system should provide the user with local name matching and local search. In our case, we should aim to address the problems of Section 4.3 and 4.4 of our taxonomy of name variations from chapter 4, i.e., misspellings and format, in addition to name variants. Multilingual name variations (and translations of name variants) are not considered because it is unlikely for a single user to capture all the possible variations of a name; therefore, the issue of multilingual name variations concerns the global level.

As noted in section 6.1, the usual approach is to deploy algorithms[8] tailored to the type of names, i.e., strings, the system is to match. Given the many possible cases of monolingual variations, as illustrated in Section 4.3 and 4.4, most of the times name matching does not rely on a single algorithm, rather a combination of techniques, e.g., "filtering using bag distance [a bag is a multi-set of the characters in a string, smaller or equal to the edit distance] followed by a more complex edit distance based approach" [7]. On the other, speed is an important factor, and techniques like Jaro[9] or Winkler[10] are well suited if one keeps this factor in mind.

Moving to name variants, they are indexed at this level (unlike monolingual variations), but their presence relies either on manual insertion by a user or import from another user or global level; thus, they are stored in the name entry 'Other Names'.

As for our scenario, unlike many other systems, we do not have different fields for parts of name (e.g., first name and surname for Person entity type)[11], but speed is of paramount importance. Given these needs, we believe that the Winkler algorithm is the best choice; in fact, as [7] notes, this algorithm is overall useful for improving the matching, regardless of the other techniques employed. In addition, other suitable techniques to be coupled with the Winkler algorithm are, e.g., LCS, to address format issues and long unparsed names. One exception are the format changes due to bibliography styles[12], which show a certain fixity, requiring, in addition to Winkler and LCS, tailored algorithms that convert the format of a certain style to another one (e.g., from 'ama' to 'IEEE' style), thus treating the problem at runtime and avoiding storing them.

Overall, the sequence of matching at local level should follow these steps:

[8] See section 6.1 for an overview on name matching algorithm we are comparing in this section. For a deeper and more throughout comparison, see [42] or [7].

[9] See equation 6.1

[10] See equation 6.2

[11] See [23] for an overview of the methods to parse names.

[12] See Table 4.1 and overall Section 4.4 for an overview on the matter.

1. A name is queried.

2. The system checks both ‘Preferred’ and ‘Other Names’ entries.
   In addition, it considers the entity type:

   - If the entity *belongs* to ‘Person’ or ‘Landform’ entity types,[13] perform Winkler and LCS.
     If not matching and the entity belongs to ‘Person entity type, perform tailored bibliography style algorithms.
   - If the entity *does not belong* to ‘Person’ or ‘Landform’ entity types, perform Winkler
     If not matching, then perform LCS.

3. If the name *matches*, the process ends and the entity is retrieved.

4. If the name *does not match*, the search moves to the global level.

7.3.2 Global Level Implementation

If name matching moves to this level, either the local search is failed or a user is looking for more information about an entity. Nevertheless, now the issue are multilingual variations. In fact, we proposed in Chapter 5 and Section 4.2 that the solution to the changes name variations cause is using named dictionaries, i.e., lists of names. Relying on name dictionaries, from name matching at local level, i.e., finding whether an entity is present or not, we now have name *searching*, i.e., searching for matching candidates among lists of names.

The implementation of name dictionaries in our scenario is based on one factor: whether the resource is importable or not. The reasons for not importing may vary (e.g., licences, required disk space, etc. . .), but they can generally be avoided by using *application programming interfaces* (API), i.e., interfaces between different software programs to facilitates their interaction, thus keeping the resource external and still using it at the same time, although, in the case of license, the data may need to be purchased. We will not go into details about the resources available here, but an overview can be found in Section 7.4.

In Section 4.2, we also stated that named dictionaries should also list part-of name (e.g., ‘Lake’) and their implementation should not just rely on named dictionaries, but rather storing them in a dedicated index. Keys and values should be the token of the trigger words, plus all their translation and

[13] As they are the only one with format variations, see Table 7.1

the proper nouns or proper names that go along with them, e.g.*key* ‘Lago’[it], *value* ‘Maggiore, Trasimeno’[it]. This way, when matching an entity type affected by this variation, the system searches this index, instead of considering the whole DHT, speeding up the process of matching.

Now that we explained the implementation of name dictionaries for multilingual variations, we can illustrate the steps of name matching at global level:

Case 1 A name is queried at local level and no match is found.

Case 2 A user is searching for an entity.

1. The system searches through the DHT for a match, *if no match is found*, move to step 2.

   Case 1 If a match is found, the variation enriches the user’s entity name.

   Case 2 If a match is found, the entity is imported in the user’s repository.

1a If the entity belongs to the entity types of either ‘Organization’, ‘Person’, and ‘Place’, the dedicated index is searched instead.

2. The system searches the name dictionaries, following two criteria: i) The entity type to which the entity name belongs and ii) The language of the user, to further narrow down the list of candidates. *If no match is found*, move to step 3.
3. The system searches the external resources via APIs. *If no match is found*, move to step 4.
4. *If no match is found yet*, a new entity is created.

7.4 Entity Type Resources and Named Dictionaries

Now that we illustrated how to implement our solutions, we want to conclude this chapter by turning our attention to entities and entity types. In fact, we showed that tackling the issues of name variants and name variations require the aid of named dictionaries. Therefore, we need to provide information about which resources (e.g., named dictionaries) are to be used, indicating where to find them and how to implement them. Firstly, we will illustrate the resources available (to our knowledge) for each of the entity types listed in Section 7.1, then all the information about the resources are shown in Table 7.2 at the end of this section.

Before introducing the resources, we wish to illustrate the criteria by which we present (and thus chose) the resources.

Size we consider the amount of information (primarily names, of course) the resource provides

API or Import we consider if the resource can be imported and indexed or if an API (application programming interface), i.e., an interface between different software programs to facilitates their interaction, may be viable, thus keeping the resource external.

Downloadable we consider if the whole resource is downloadable or not.

Type of Data we consider whether the resource provides simply names or more structured information, e.g., attributes.

Language we consider what and how many languages the data are available in.

License we consider if the resources is free or under copyright.

7.4.1 CreativeWork

First of all, the available resources for every entity type of *CreativeWork* are:

Book the HeiNER database is a good starting point, but sites like 'ISBNdb'[14], with 6,954,137 books title (as of 03/08/2012), may be imported or, as in this case, used via API.

Movie the HeiNER database is a good starting point, as good film related sites like 'IMDb'[15] have copyright issues, since they don't allow any use of their data, whereas the MEDIA Film Database[16] is a free searchable database of 6325 European films.

MusicAlbum the HeiNER database is a good starting point (excluding name variants and name variations), but there are free databases like MusicBrainz,[17] which contains information 12 million recordings.

[14] http://isbndb.com/

[15] http://www.imdb.com/

[16] http://eacea.ec.europa.eu/media/films/

[17] http://musicbrainz.org/

MusicRecording See MusicAlbum.

Painting the HeiNER database is a good starting point, and a good thesaurus, although under development and copyrighted, could be the Cultural Objects Name Authority (CONA),[18] so it will not be included in our current table, as we have no data with regard to its size or availability.

TVSeries See Movie, since sites that deal with movies usually deal with tv series.

7.4.2 Event

As for 'Event' entity types, in addition to the HeiNER database, Eventseer[19] is a database of 19,740 events, which is both available by download or via API.

7.4.3 Organization

As for 'Organization' entity types, there are more resources than the HeiNER database.

Corporation the HeiNER database is a good starting point.

Organization the HeiNER database is a good starting point, but there are sites like 'devdir'[20] (for development organizations, listing 70,000 of them with also contact information) or 'NGO s and Directories'[21], to be used via API. In addition, government sites usually store names and contact information about national organizations.

SportsTeam the HeiNER database is a good starting point.

EducationalOrganization the HeiNER database is a good starting point; in addition, sites like 'univ'[22], a searchable of 8972 Universities in 204 countries (as of 22/07/2012) could be useful via API.

[18] http://www.getty.edu/research/tools/vocabularies/cona/index.html

[19] http://eventseer.net/

[20] http://www.devdir.org/

[21] http://www.gdrc.org/ngo/ngo-s.html

[22] http://univ.cc/

7.4.4 Person

The Person entity type resources may take advantage of the following resources.

Names behindthename[23] contains 18,086 names and their conversion to different languages, but seems easier to be searched via API, whereas a collection of 9,353 English names, plus their equivalents in 12 different languages, downloaded from rootsweb[24] could be imported as tokens, while the website itself could be used via API to take advantage of its links to multiple free searchable databases.
Moreover, the Union List of Artist Names (ULAN),[25] contains 638,900 artist names, but, because of copyrights and the fee needed for the data, seems to be accessible only via API, while ISBNdb also contains 1,998,869 names of authors, although in English only. In addition, MusicBrainz stores about 660,000 artists names, and IMDb contains 4,780,533 names of movie and tv series related people.

Religious Figures and Monarchy the HeiNER database is a good starting point, but for Cristian saints 'catholic-saints'[26] could be imported, as it also consist of biographic information about the saint.

7.4.5 Place

Finally, having already downloaded more than 7 million entities from GeoNames[27], the HeiNER database is but a small addition to names for *Place* entity types, considering also the Getty Thesaurus of Geographic Names (TGN),[28] that, because of copyright, could be accessible via API.

7.4.6 General Resources

In addition to these entity type specific resources, there are some free resources that cover the whole range of entity types we deal with in this work, containing general information of all types of entities.

Here are some of them:

[23] http://www.behindthename.com/

[24] http://www.rootsweb.ancestry.com/

[25] http://www.getty.edu/research/tools/vocabularies/ulan/index.html

[26] http://www.catholic-saints.info/

[27] http://www.geonames.org/

[28] http://www.getty.edu/research/tools/vocabularies/tgn/index.html

Freebase [29] A collaborative knowledge base built on structured data harvested from many sources, including individual wiki contributions, containing more than 23 million entities.

DBpedia [30] A community effort to extract structured information from Wikipedia and to make this information available on the Web, whose dataset describes more than 3.64 million entities.

YAGO2 A knowledge base containing information harvested from Wikipedia and linked to WordNet, containing more than 10 million entities.

TheDataHub [31] A community-run catalogue of useful sets of data on the Internet

To sum up, Table 7.2 shows the criteria (listed at the beginning of this section) about the resources in the columns, while the rows represent the resources and the relative information about the criteria. of course, this list is not to be considered exhaustive, but it can be used as a starting point for finding further resources for names.

[29] http://www.freebase.com/

[30] http://dbpedia.org

[31] http://thedatahub.org

Resources	Size[32]	API or Import	Downloadable	Type of Data	Language	License
BEHINDTHENAME	18,086	API	No	Names only	English, plus variations	Free
CATHOLIC-SAINTS	Unknown	API	No	Names and biography	English	Free
DBPEDIA	3.64 million	Both	Yes	Entities	English	
DEVDIR	72,432	Import	Pdf files	Names, contact info	English, French, Spanish	Free
EVENTSEER	19,740	Both	No	Event names and info	English	Free
FREEBASE	> 23 million	API	No	Entities	English	CC
HEINER	> 1.5 Million	Import	Full Download	Named Entities	253	Free
IMDB (MOVIES)	2,280,354	API	No	Movie names and info	English	Fee
IMDB (PEOPLE)	4,780,533	API	No	Movie names and info	English	Fee
ISBND (authors)	1,998,869	API	Some data, but requires a specific software	Names and published books	English	GPL 2.0
ISBND (books)	7,034,202	API	Some data, but requires a specific software	Titles, authors, publishers	English	GPL 2.0
MEDIA	6325	API	No	Film names and production year	European Languages	Free
MUSICBRAINZ (MUSIC)	12 million recordings	API	No	Recording names and info	English	CC
MUSICBRAINZ (PEOPLE)	660,000	API	No	Names and discography	English	CC
ROOTSWEB	> 640 Million	API	9,353 sample names	Names only	At least 12 languages	Free
THEDATAHUB	4346	Import	Yes	Datasets	22	Free[a]
TGN	1,711,110	API	No	Names only	Mostly English	Fee
ULAN	638,900	API	No	Names only	Mostly English[33]	Fee
UNIV.CC	8,989	API	No	Names, links to universities website	English[34]	Free
YAGO2	> 10 million	Both	Yes	Entities	253	Free

Table 7.2: Entity Types Resources

[a] Although some datasets may be not open links.

[a32] As of 15/09/2012

[a33] However, the structure of the ULAN (and TGN) supports multilinguality

[a34] Some universities names are translated in English, while others are not, without any clear criterion

Chapter 8

Conclusions

In this work, we aimed to tackle a complicated issue for name matching, i.e., matching name variations and name variants. In order to do so, we avoided relying completely on automatic approaches, taking a multidisciplinary approach instead.

We started from philosophy, and the debate surrounding names that took place during the last century. The two themes of this debate revolved around reference (i.e., the process that allows the speakers to identify an entity) and meaning (i.e., the possibility for a name to have a semantic value). Theories like descriptivism support the idea that names do have a meaning (i.e., the definite description associated with it) and explain reference as the process of associating the true description an entity, whereas other theories like causal theories and Millianism deny that names have any meaning outside their reference. While both the approaches have their strengths and weaknesses, there is no current theory exhaustive enough to be accepted; in fact, accepting any of the approaches described carries different problems.

Moving to sociology, one can see that a name, rather than indicating a single modality of reference, is actually more like a 'class' of types of names, i.e., *name variants* (more commonly, pseudonyms), varying according to different features (e.g., social contexts). Because of this, it is hard to distinguish them both in terms of usage (as social context borders tend to be blurred, if not overlapping) and in terms of definition (consider the various definitions of pseudonyms in [1], [35] and [11]), leading to confusion when referring to an entity.

Finally, in geography, we saw that name variations, be they multilingual (e.g., translations) or monolingual (e.g., misspellings), affect names of geopolitical entities, adding linguistic and cultural factors to social ones to the complexity of names. Furthermore, we saw that these factors are not limited to geography, but they are present and persistent in many other fields, even in daily life, too.

All these issues affect name matching, i.e., matching two strings to see if they refer to the same entity, and any area of computer science that deals with names at large. In our scenario of a P2P, entity-based network of users, name matching does not simply rely on strings when operating, but also considers other factors, e.g., the type of entity that is being matched. In addition, our scenario consists of three levels: local level (the users), community level (groups of users), and global level (all the entities). Entities at local level are a partial view of the real word entity, represented at the global level. Thus, name matching operates with different challenges with each level. At local level, there are name variants and monolingual name variations; while the variations are accounted for by algorithms, name variants must be tackled by using name dictionaries, since the vary on various factors. Similarly, multilingual name variations dominate the global level, and, like variants, must be tackled by using name dictionaries. Finally, both variants and variations should be represented in the structure of the entity name, so to provide a more fine-grained view of the entity itself

Thus, we proposed a taxonomy of name variations (full translation, part-of translation, misspellings, format and variants) to understand and predict the variations and variants of different entity names, validating the taxonomy by testing it on schema.org, a lightweight ontology; thus, we obtained a way to recognize patterns in the variants and variations behaviour. As for the entity name architecture, we divided it in two entries: 'Preferred' (i.e., the most salient name) and 'Other Names (name variants and variations)'; although simple enough, it allows to give a more extensive overview of the entity name. Both solutions are due to our multidisciplinary approach, taking advantage both from various fields (i.e., philosophy, sociology and geography), importing terms and views not found in computer science, and also drawing from areas close to name matching, building from their findings and expanding them.

Finally, we must note that our solutions are intended to be hints and guidelines for a future implementation, which was outside the scope of this work, while we focused on the more theoretical aspect of the issue of matching name variations and name variants in a distributed scenario. Although the scope of name matching may seem less concerned with names from a theoretical point of view, we believe that providing a more robust multidisciplinary background helps in clarifying issues such as the one discussed in this work. In additions, our work shows how apparently distant fields have much more in common that it seems or one wants to believe.

Appendix A

Tables from Schema.org

In this chapter we have a list of 22 tables, which are used to back up our intuitions of Chapter 4. In fact, while Table 7.1 shows the general trend of name variations, these tables illustrate the trend for every single entity type.

A.1 Notes to the Tables

Full Translation: this column represent all the translations are based on Wikipedia[1], i.e., we checked our translation by relying on the pages in other languages of every entity.

Part-of translation: in this column, because of space and formatting reasons, we just put the trigger word of the name underlined.

Alternative Names: this column was added to also take into account pseudonyms for the sake of completeness. Moreover, although they are name variants and not name variations, they may be translated.
Yet, it could be that 'Alternative Names'may be the official name of the entities (i.e., the entity original name). In fact, since we opted for salience as the main criterion for deeming the name 'Alternative' or not, as in section 3.1, it may be that the official name is less used that another one, although this is not a fixed occurrence.
To prove our claim, we decided to submit both names as queries to Google and Wikipedia, then labelling as 'Alternative' the one with the fewest occurrences. For instance, although 'Alice's Adventures in Wonderland' is the original name of the book by Carroll, 'Alice in Wonderland' is by far the most used (69.900.000 vs. 912.000 results

[1] http://en.wikipedia.org/wiki/Wikipedia

on Google); since ‘Alice’s Adventures in Wonderland’ is less used, it becomes a value of ‘Alternative Name’. On the other hand, ‘The Big Apple’ cannot substitute in salience ‘New York’ (1,9 billions vs. 3 billions results, respectively).

Misspellings: because of space and formatting reasons, we just put one of the possible misspelled parts of the name.

Format: because of space and formatting reasons, we only listed one example of format variation, as the whole list can be easily obtained automatically, which may be if, e.g., one follows a certain formatting style for authors’ name.[2]

Furthermore, the language considered for all the instances in the tables, thus for both Table 7.1 and those in Appendix B, are the EU official ones, i.e., English, German, French, and Spanish, plus Italian and Danish; nevertheless, the main languages considered are English and Italian. If an entity comes from either of the languages, we use the other one for translation (if available). On the other hand, if the entity comes from any other language than English and Italian, English becomes the language for translation; otherwise, we switch to Italian.

Further notes related to one table only are added accordingly.

[2] See Section 4.4 for an overview of the possible bibliography style, and how they change the name.

Book Type Instances	Class of Name Variations											
			Misspellings									
	Full Translation	**Part-of translation**	Punctuation	Capitalization	Spacing	Omissions	Additions	Substitution	Phonetic variations	Switching letters	**Alternative Title**	**Format**
Alice in Wonderland	Alice nel Paese delle Meraviglie[it]			alice	nel paese	dele	Merraviglie	im		Aliec	Alice's Adventures in Wonderland[en], Le avventure di Alice nel Paese delle Meraviglie[it]	
Ansichten eines Clowns[de]	Opinioni di un Clown[it]			clown	di un	Clown	Anssichten	Clowb		Clonw		
Moby Dick[en]			Moby-Dick	moby dick	Moby Dick	Dic	Mobby	Dicl		Moyb	The Whale[en], La Balena[it]	
La Divina Commedia[it]	Divine Comedy[en]			divine	Divine Comedy	Comedia	Commedy	Dibina		Dviina	Commedia[it]	
Kinkakuji[jp]	Il Padiglione d'Oro[it]			padiglione	d' Oro	Padilione	Orro	Oto		Padiglinoe		
Harry Potter and the Goblet of Fire[en]	Harry Potter e il Calice di Fuoco[it]			harry potter	of Fire	Poter	Harrry	Jarry		Fier		
Of Mice and Men[en]	Uomini e Topi[it]			uomini	Of Mice	an	Miice	Nice		Toip		
Paradise Lost[en]	Paradiso Perduto[it]			lost	Paradise Lost	Paradis	Perdutto	Kost		Perdtuo		
Pride and Prejudice[en]	Orgoglio e Pregiudizio[it]			pride	pride and	Orgolio	Pridde	Prejudics		Prejudiec		
Slaughterhouse-Five[en]	Mattatoio n. 5[it]		Mattatoio n 5	mattatoio	n 5	Matatoio	Slaughterhous	Mattatoiu		Mattatooi	The Childrens Crusade: A Duty-Dance with Death[en], La Crociata dei Bambini[it]	
The Catcher in The Rye[en]	Il Giovane Holden[it]			holden	in the	Hlden	Giovanee	Giovanr		Holdne		
The Lord of the Rings[en]	Il Signore degli Anelli[it]			lord	of the	Aneli	Rinngs	Anekli		Rnigs		
Der Process[de]	Il Processo[it]			processo	Il Processo	Proceso	Proccesso	Provesso		Porcesso	Der Prozess, Der Prozeß[de]	
The Picture of Dorian Gray[en]	Il ritratto di Dorian Gray[it]			dorian gray	Dorian Gray	Ritrato	Graay	Riyratto		Doiran		
Ubik[en]				ubik		Ubi	Ubbik	Ubij		Ubki		

Table A.1: Book Type Instances

Movie Type Instances	Class of Name Variations										
			Misspellings								
	Full Translation	**Part-of translation**	Punctuation	Capitalization	Spacing	Omissions	Additions	Substitution	Phonetic variations	Switching letters	**Alternative Title**	**Format**
$8\frac{1}{2}$			8 1/2	otto	8 $\frac{1}{2}$	Oto	Mezzzo	Mezzp		Otot	8 e Mezzo, Otto e Mezzo[it]	
Blade Runner[en]				blade	Blade Runner	Runer	Bladde	Bkade		Runnre		
Das Leben der Anderen[de]	Le Vite degli Altri[it]			leben	Le Vite	Alti	Vitte	Vire		Andreen		
Divorzio all'Italiana [it]	Divorce, Italian Style[en]		Divorce Italian Style	divorzio	All' Italiana	Italana	Stylee	Italiama		Divroce		
Eternal Sunshine of the Spotless Mind[en]	Se Mi Lasci Ti Cancello[it]			eternal	Of The	Spotles	Canccello	Minf		Sunshnie		
Forrest Gump[en]				forrest	Forrest Gump	Forest	Gummp	Forresr		Gmup		
Gone with the Wind[en]	Via Col Vento[it]			vento	Via Col	Gon	Vennto	Widn		Teh		
Hævnen[dk]	In un Mondo Migliore[it]			hævnen	Un Mondo	Hævnn	Hæevnen	Migliorw		Modno		
La Vita È Bella[it]	Life Is Beatiful[en]			vita	Life Is	Bela	Liffe	Beatuiful		Beautifuk		
Mar Adentro[es]	Il Mare Dentro[it]			mar	Il Mar	Adntro	Marr	Adnetro		Marw		
No Country for Old Men[en]	Non è un Paese Per Vecchi[it]				For Old	Vechi	Oldd	Cointr	E un	Vechci		
The Lion King II[en]	Il Re Leone II[it]		Il Re Leone 2	re	Re Leone	Kin	Leonne	Leonw		Lino	The Lion King 2: Simba's Pride[en], Il Re Leone II- Il Regno di Simba[it]	
Rambo[en] [3]			Rambo, First Blood[en]	rambo	First Blood	Blod	Rammbo	Ramno		Frist	First Blood[en], Rambo: First Blood[en]	
Pride and Prejudice[en]	Orgoglio e Pregiudizio[it]			pride	pride and	Orgolio	Pridde	Prejudics		Prejudiec		
Runaway Bride[en]	Se Scappi ti Sposo[it]			bride	Se Scappi	Scapi	Sposso	Scapip		Sposp		
Terminator 2[en]			Terminator II	terminator	Il Giorno	Giono	Dayy	Dat		Terminatro	Terminator 2: Judgment Day[en],Terminator 2- Il Giorno del Giudizio[it]	
The Little Mermaid[en]	La Sirenetta[it]			mermaid	The Little	Litle	Sirrenetta	Sirenrtta		Sirenetat		
The Silence of the Lambs[en]	Il Silenzio degli Innocenti[it]			lambs	The Silence	Inocenti	Slienzzio	Lambd		Slience		

Table A.2: Movie Type Instances

[3] In this case Rambo may also refer to the fourth movie of the serie. On the other hand, the last Rambo film has a different 'Alternative Title' (i.e, John Rambo, Rambo IV[en]), which can be used for disambiguation.

MusicAlbum Type Instances	Class of Name Variations										
	Full Translation	**Part-of translation**	**Misspellings**							**Alternative Title**	**Format**
			Punctuation	Capitalization	Spacing	Omissions	Additions	Substitution	Phonetic variations	Switching letters		
A Rush of Blood to the Head[en]				rush	A Rush	Blod	Russh	Bloos		Rsuh		
Back in Black[en]				back	Back In	Bck	Blacck	Bkack		Bakc		
Bridge Over Troubled Water				bridge	Over Troubled	Brige	Troubkled	Ocer		Wtaer		
Gold: Greatest Hits[en]				gold	Greatest Hits	Greatst	Hitts	Greayst		Hist		
Innuendo[en]				innuendo		Inuendo	Innnuendo	Innuendp		Innuedno		
Le Dimensioni del Mio Caos[it]				dimensioni	Del Mio	Dimnsioni	Caoss	Caod		Coas		
Low[en]				low		Lo	Loow	Leo		Lwo		
Madman Across the Water[en]				madman	Across The	Acros	Maddman	Madmsn		Watre		
Rosenrot[en]				rosenrot		Rosnrot	Rossenrot	Rosentot		Roesnrot		
Stagioni[en]				stagioni		Stagoni	Staggioni	Staigioni		Stagiomi		
Storia di un Impiegato[en]				storia	di un	Impegato	Storria	Impiehato		Impeigato		
The Beatles[en]				beatles	The Beatles	Betles	Beattles	Beatkes		Baetless	The White Album[en]	
The Dark Side of the Moon[en]				dark	The Dark	Sid	Darck	Nlack		Sied		
Thriller[en]				thriller		Thriler	Thrriller	Thrillwr		Thrillre		
Pot Luck with Elvis[en]				pot	Pot Luck	Luc	Poot	Poy		Lukc	Pot Luck[en]	

Table A.3: MusicAlbum Type Instances

MusicRecording Type Instances	Class of Name Variations											
	Full Translation	**Part-of translation**	**Misspellings**								**Alternative Title**	**Format**
			Punctuation	Capitalization	Spacing	Omissions	Additions	Substitution	Phonetic variations	Switching letters		
9. Sinfonie[de]	Nona Sinfonia[it]		9 Sinfonie	sinfonia	La Nona	Sinfona	Nonna	Sindonia		Sinfoina	9. Sinfonie in d-Moll op. 125[de], Nona Sinfonia in Re minore, Op. 125, La Nona (di Beethoven)[it]	
Bohemian Rhapsody[en]				bohemian	Bohemian Rhapsody	Rapsody	Bohemmian	Bohwmian		Rhaspody		
Canzone del Maggio[it]				maggio	Del Maggio	Magio	Cannzone	Canzonw		Caznone		
Die Zauberflöte[de]	Il Flauto Magico[it]			flauto	Il Flauto	Fluto	Maggico	Flite		Maigco		
Du Hast[en]				hast	Du Hast	Has	Hasst	Dy		Hats		
Eye of the Tiger[en]				eye	Of The	Tigr	Eyee	Ete		Tiegr		
Hey Jude[en]				jude	Hey Jude	Jud	Heyy	Judw		Jdue		
Innuendo[en]				innuendo		Inuendo	Innnuendo	Innuendp		Innuedno		
L'Oiseau de Feu[fr]	L'Uccello di Fuoco[it]			uccello	L' Uccello	Ucello	Fuocco	Fei		Oisaeu		
La Traviata[it]				traviata	La Traviata	Travita	Travviata	Travuta		Travaiata		
Le quattro stagioni[it]	The Four Seasons[en]			season	Le Quattro	Quatro	Staggioni	Quattrp		Seasno		
Macarena[es]				macarena		Macarna	Maccarena	Macarwna		Macaerna		
Messiah[en]	Le Messie[fr]			messiah	Le Messie	Mesie	Messiiah	Messiaj		Messei		
We Are the World[en]				we	We Are	Word	Arre	Worls		Wolrd		
Y.M.C.A.[en]			YMCA	ymca		YMC	YMMCA	TMCA		YCMA		

Table A.4: MusicRecording Type Instances

Painting Type Instances	**Class of Name Variations**											
			Misspellings									
	Full Translation	**Part-of translation**	Punctuation	Capitalization	Spacing	Omissions	Additions	Substitution	Phonetic variations	Switching letters	**Alternative Title**	**Format**
Le Café de Nuit[fr]	Il Caffè di Notte[it]			caffè	Il Caffè	Note	Caffé	Nottw		Acffè		
Campbell's Soup[en]				campbell's	Campbell' s	Campbel's	Cammpbell's	Spup		Capmbell's		
Le Déjeuner sur l'Herbe[fr]	Colazione sull'Erba[it]			erba	La Colazione	Herb	Colazzione	Herve		Coalzione		
El Sueño de la Razón Produce Monstruos	Il Sonno della Ragione Genera Mostri[it]			mostri	Il Sonno	Sono	Raggione	Geners		Ragoine		
Guernica[es]				guernica		Guernia	Guerrnica	Guerbica		Guerinca		
L'Empire des Lumières[fr]	L'Impero delle Luci[it]			luci	Impero Delle	Dele	Immpero	Luco		Emprire		
Monna Lisa[it]	Mona Lisa[en]			mona lisa	Mona Lisa	Mona[it]	Giocconda	Moba		Lias	Portrait of Lisa Gherardini, wife of Francesco del Giocondo[en], La Gioconda[it]	
Moulin Rouge: La Goulue[fr]				moulin	La Goule	Ruge	Moulinn	Rougw		Guolue	Moulin Rouge- La Goulue[fr]	
Les Joueurs de Cartes[fr]	I Giocatori di Carte[it]			giocatori	Di Carte	Jouer	Carrtes	Joieurs		Catre		
Nascita di Venere[it]	Birth of Venus[en]			venus	Birth Of	Nasita	Birrth	Nadcita		Veenre		
Maestà di Ognissanti	Ognissanti Madonna[en]			madonna	di Ognissanti	Madona	Maddonna	Ognissanyi		Maodnna	Madonna Enthroned[en]	
La Persistencia de la Memoria[es]	La Persistenza della Memoria[it]			memoria	La Persistenza	Blanos	Blanndos	Persistwnza		Memoira	Los Relojes Blandos[es]	
San Patrizio Vescovo d'Irlanda[it]	Saint Patrick, Bishop of Ireland[en]	Saint[en], San[it]		Patrick Bishop	San Patrizio	Bisop	Pattrick	Irkanda		Partick		
Série des Cathédrales de Rouen[fr]	La Cattedrale di Rouen in Pieno Sole[it]	Cattedrale di[it]		cattedrale	Di Rouen	Catedrale	Rouenn	Royen		Catterdale	La Cattedrale di Rouen[it]	
Vocazione di San Matteo[it]	The Calling of Saint Matthew[en]	Saint[en], San[it]		calling	San Matteo	Caling	Vocazzione	Mattwo		Callign		

Table A.5: Painting Type Instances

MusicEvent Type Instances	Class of Name Variations											
	Full Translation	**Part-of translation**	**Misspellings**								**Alternative Name**	**Format**
			Punctuation	Capitalization	Spacing	Omissions	Additions	Substitution	Phonetic variations	Switching letters		
A Bigger Bang Tour[en]				bigger	A Bigger	Ban	Banng	Tout		Biggre		
Africa Oyé 2006[en]				africa	Africa Oyé	Afrca	Affrica	Afwica	Oye	Afirca		
Ferrara Buskers Festival 2008[it]				buskers	Ferrara Buskers	Buskrs	Busskers	Buskwrs		Buskres		
Diamond Jubilee Concert[en]				diamond	Diamond Jubilee	Diamnd	Jubillee	Concery		Diamodn		
Dangerous World Tour[en]				dangerous	Dangerous World	Dangrous	Worrld	Dangerpus		Wordl		
Evolution Festival 2008[it]				evolution	Evolution Festival	Evlution	Evollution	Evolutoon		Evoltuion		
Futuresonic 2006[en]				futuresonic	Futuresonic 2006	Futursonic	Futurresonic	Futurwsonic		Futuersonic		
Gods of Metal 2011[it]				gods	Gods of	Metl	Mettal	Metsl		Meatl		
Grange Park Opera 1998[en]				grange	Grange Park	Grane	Parrk	Opwra		Gragne		
Grimeborn 2007[en]				Grimeborn	Grimeborn 2007	Grimborn	Grimeborrn	Grimwborn		Griemborn		
Live Aid[en]				live	Live Aid	Liv	Aidd	Livw		Adi		
Roskilde Festival 2006[dk]				roskilde	Roskilde Festival	Roskild	Rosskilde	Rosjilde		Rosklide		
Projekt Revolution 2006[en]				projekt	Projekt Revolution	Projek	Revvolution	Projwkt		Reovlution		
Summer 1963 UK Tour[en]			U.K.	summer	UK Tour	Summr	Touur	Summwr		Tuor		
The Works Tour[en]				works	The Works	Work	Worrks	Wprks		Wroks		
The Wall Tour[en]				wall	The Wall	Wal	Waall	Wsll		Wlal		

Table A.6: MusicEvent Type Instances

SocialEvent Type Instances	Class of Name Variations										
			Misspellings								
	Full Translation	Part-of translation	Punctuation	Capitalization	Spacing	Omissions	Additions	Substitution	Phonetic variations	Switching letters	Alternative Name	Format
34th Golden Globe Awards[en]	Golden Globe 1977[it]			globe	34th Golden	Goldn	Globbe	Globr		Goledn	34th Golden Globe, Golden Globe[it]	
4th July[en]	4 Luglio[it]			july	4th July	Indipendnce	Fourrth	Lugkio		Juyl	Fourth of July, Independence Day[en]	
40th Grammy[en]	Grammy 1998[it]			grammy	40th Grammy	Gramy	Grrammy	Grammt		Gramym	40th Grammy Awards, Grammy[en]	
65th British Academy Film Awards[en]	BAFTA 2012[it]		B.A.F.T.A.	bafta	65th BAFTA	Britsh	Awwards	Btitish		Acaedmy	65th BAFTA[en], Premi BAFTA 2012, BAFTA[it]	
84th Academy Awards[en]	Oscar 2012[it]			oscar	Oscar 2012	Acadmy	Awarrds	Academt		Osacr	Premi Oscar 2012, Oscar[it]	
9/11[en]	11 Settembre[it]			settembre	9/ 11	Atentati	Septembber	Attentsti		Sepetmber	September 11 Attacks[en], Attentati dell'11 Settembre 2001[it]	
David di Donatello 2006[it]				david	David di	Donatelo	Davvid	Dsvid		Doantello	David di Donatello[it]	
French Presidential Elections 2012[en]	Elezioni Presidenziali Francesi del 2012[it]			elezioni	Elezioni Francesi	Presidnziali	Franncesi	Frebcj		Eletcions	French Elections 2012, French Elections[en]	
IJCAI 2001[en]			I.J.C.A.I.	ijcai	IJCAI 2001	IJCI	IJJCAI	IJCSI		IJACI	International Joint Conference on Artificial Intelligence[en]	
Elezioni Politiche Italiane 2006[it]	Italian General Election 2006[en]			italian	Election 2006	Eletion	Ittalian	Generak		Poiltiche	Italian Elections 2006,Italian Elections[en]	
MUC-7			M.U.C.	MUC	MUC 7	MC	MUUCI	Understanfing		Conefrence	Message Understanding Conferences 7[en]	
The Civil War[en]	Guerra di Secessione[it]			war	The Civil	Secesione	Guerrra	Civik		Ameircan	American Civil War[en], Guerra di Secessione Americana[it]	
TREC 2007[en]			T.RE.C.	trec	TREC 2007	TRC	TREEC	TRWC		TERC	Text REtrieval Conference 2007[en]	
USA Elections 2008[en]	Elezioni Stati Uniti 2008[it]		U.S.A.	elections	Stati Uniti	208	Unitti	Presidenyial		Eletcion	United States Presidential Election 2008, United States Presidential Election[en]	
World War II[en]	Seconda Guerra Mondiale[it]			guerra	WW 2	Guera	Monndiale	Gyerra		Secodna	Second World War, WWII, WW2[en]	

Table A.7: SocialEvent Type Instances

SportEvent Type Instances	Class of Name Variations											
			Misspellings									
	Full Translation	Part-of translation	Punctuation	Capitalization	Spacing	Omissions	Additions	Substitution	Phonetic variations	Switching letters	Alternative Title	Format
14th FINA World Championships[en]	Mondiali di Nuoto 2011[it]		F.I.N.A.	mondiali	14th FINA	Championship	Chammpionships	Chanpionships		FNIA	2011 World Aquatics Championships, FINA World Championships[en], Campionati Mondiali di Nuoto 2011[it]	
1999 World Netball Championships[en]				netball	World Netball	Ntball	Nettball	Netbsll		Nebtall	World Netball Championships[en]	
RoboCup 2010 Singapore[en]				roboCup	Robo Cup	RobCup	RobboCup	RoboCyp		RoobCup	RoboCup 2010, RoboCup[en]	
2010 Winter Olympics[en]	XXI Giochi Olimpici Invernali[it]			invernali	Giochi Olimpici	Vancuver	Vanncouver	Vancouvwr	Olimpics	Wintre	Vancouver 2010, Winter Olympics[en]	
2006 Commonwealth Games[en]	XVIII Giochi del Commonwealth[it]			commonwealth	2006 Commonwealth	Comonwealth	Commonweallth	Commobwealth		Commownealth	Commonwealth Games[en]	
2008 New York City Marathon[en]	Maratona di New York 2008[it]		N.Y.C.	new	New York	Maraton	Marrathon	Maratjon		Martahon	2008 NYC Marathon, NYC Marathon, New York City Marathon[en], Maratona di New York del 2008, Maratona di New York[it]	
2008 Olympics[en]	Olimpiadi 2008[it]			pechino	2008 Olympics	Beijin	Pecchino	Olimpisdi	Olimpics	Olympisc	2008 Summer Olympics, Olympics[en], Pechino Giochi della XXIX Olimpiade, Olimpiadi[it]	
2008 Paralympics[en]	XIII Giochi Paralimpici[it]			paralimpici	2008 Paralympics	Parlympics	Parallympics	Paealympics	Paralimpics	Giohci	2008 Summer Paralympics, Paralympics[en], XIII Giochi Paralimpici estivi, Giochi Paralimpici[it]	
El Clàsico[es]				clàsico	El Clàsico	Clàsco	Clàssico	Clasico		Clàscio	El Derby Español[es]	
Italia 90[it]	1990 FIFA World Cup[en]			italia	Italia 90	Copa	Worlld	Mondp		Campioanto	Campionato Mondiale di Calcio 1990, Coppa del Mondo FIFA del 1990, Coppa del Mondo 1990, Coppa del Mondo[it], FIFA World Cup[en]	
Masters Tournament 2003[en]	Masters[it]			masters	Masters Tournament	Master	Massters	Mastwrs		Matsers	The Masters 2003, 2003 U.S. Masters, The Masters, U.S. Masters[en]	
Super Bowl XLIII[en]				super	Super Bowl	Supr	Bowll	Supwr		Supre	Super Bowl 43, Super Bowl[en]	
Tour de France 2011[fr]	Tour de Francia 2011[es]			france	Tour de	Franc	Frrance	Tout		Turo	Tour de France[en]	
Wimbledon 2011[en]				wimbledon	2011 Wimbledon	Wimbldon	Wimbleddon	Wimblwdon		Wimbeldon	2011 Wimbledon Championships, Wimbledon Championships[en],Torneo di Wimbledon 2011, Torneo di Wimbledon[it]	
2011 World Series[en]				world	World Serie	Seris	Serries	Seroes		Serise	World Series[en]	

Table A.8: SportEvent Type Instances

Festival Type Instances	Class of Name Variations											
			Misspellings									
	Full Translation	Part-of translation	Punctuation	Capitalization	Spacing	Omissions	Additions	Substitution	Phonetic variations	Switching letters	Alternative Title	Format
Capital Pride 2008[en]				capital	Capital Pride	Capitl	Pridde	Capitsl		Pried		
Coney Island Mermaid Parade 2009[en]				coney	Island Mermaid	Parad	Islland	Mernaid		Paraed		
Festival di Venezia 2010[it]	Venice Film Festival 2010[it]			venezia	Festival di	Venzia	Fesstival	Mostrs		Cineam	Mostra Internazionale d'Arte Cinematografica,Festival del Cinema di Venezia[it]	
Filmfestspiele Berlin 1999[de]	Berlin Film Festival 1999[en]			berlin	Film Festival	Berlinal	Berlinnale	Bwrlinale		Filmfsetspiele	Berlin International Film Festival[en], Internationale Filmfestspiele Berlin, Berlinale[de]	
Festival de Cannes 1978[fr]	Cannes Film Festival 1978[en]			cannes	Festival de	Canes	Caannes	Cannws		Internatioanl	Cannes International Film Festival [en], Le Festival International du Film de Cannes[fr]	
Festivalfilosofia 2003[it]				festivalfilosofia	Festivalfilosofia 2003	Festivlfilosofia	Fesstivalfilosofia	Festivalfilsofia		Festivlafilosofia		
Festivaletteratura 2004[it]				Festivaletteratura	Festivaletteratura 2004	Festivaleteratura	Festivalletteratura	Festivalettwratura		Festivaletteartura		
Hebridean Celtic Festival 2011[en]				hebridean	Celtic Festival	Hebrdean	Celltic	Fwstival		Hebirdean		
ISFiT 2013[no]			I.S.F.i.T.	isfit	i Trondheim	Trondhim	Tronndheim	Trobdheim		Trodnheim	International Student Festival in Trondheim[en], Internasjonale Studentfestivalen i Trondheim[no]	
London International Hot Air Balloon Festival 2005[en]				hot	Hot Air	Balon	Lonndon	Festivsl		Londno		
Stratford Shakespeare Festival 2002[en]				stratford	Stratford Shakespeare	Shakespeare	Shakespeare	Shakespeare		Strtaford		
Sundance Film Festival 2008[en]				Sundance	Sundance Film	Sundanc	Sundannce	Sunfance		Sundnace		
The River To River Festival 1999[en]				river	River To	Rivr	Rivver	Rivwr		Rivre		
Toronto International Film Festival 2011[en]			T.I.F.F.	tiff	Toronto Film	Toroto	Torontto	Torpnto		Torotno	TIFF[en]	
Vinitaly 2012[it]				Vinitaly	Vinitaly 2012	Vinitly	Vinnitaly	Vinitsly		Viniatly		

Table A.9: Festival Type Instances

SportTeam Type Instances [4]	**Class of Name Variations**											
			Misspellings									
	Full Translation	**Part-of translation**	Punctuation	Capitalization	Spacing	Omissions	Additions	Substitution	Phonetic variations	Switching letters	**Alternative Name**	**Format**
F.C. Barcelona[es]			FC	barça	FC Barcelona	Blaugrana	Blaugrranas	Blaigranas		Blaguranas	Fùtbol Club Barcelona, Barça, Barcelonistas, FCB, Blaugranas, Culés, el Barça[es]	
Chicago Bears[en]				bears	Chicago Bears	Cicago	Bearrs	Chicsgo		Beras	Da Bears, The Monsters of the Midway[en]	
Dodgers[en]			LA	dodgers	Los Angeles	Dodgrs	Doddgers	Dodgwrs		Dogders	Los Angeles Dodgers, The Bums, The Boys in Blue, The Blue Crew[en]	
England cricket team[en]	Nazionale di Cricket Inglese[it]			Cricket	di Cricket	Crickt	Criccket	Crickwt		Crikcet	England[en]	
Inter[it]	Inter Milan[en]		FC	inter	FC Internazionale	Intr	Innter	Intwr		Intre	Football Club Internazionale Milano,Internazionale[it]	
Lakers[en]			LA	Lakers	LA Lakers	Lakrs	Lakerrs	Lakwrs		Lakres	Los Angeles Lakers, L.A. Lakers[en]	
Manchester United[en]			FC	manchester	United FC	Manchestr	Unitted	Manvhester		Manchseter	Manchester United F.C., Manchester United Football Club, The Red Devils[en]	
Nazionale di Calcio Italiana[it]	Italy National Football Team[en]			calcio	di Calcio	Naionale	Ittaliana	Calcip		Nazionael	Azzurri, Italia[it]	
Nederlands Voetbalelftal[nl]	Netherlands Football Team[en]			oranje	Football team	Nethrlands	Netherrlands	Netherkands		Nethrelands	Oranje, Nederlands[nl], Clockwork Orange, Holland, The Flying Dutchmen[en]	
New York Rangers[en]				rangers	New York	Ranger	Ranngers	Rangwrs		Rangres		
Selección de fútbol de Paraguay[es]	Paraguay National Football Team[en]			Paraguay	de Paraguay	Paragay	Parraguay	Paragiay		Paragauy	Los Guaranìes, La Albirroja, Paraguay[es]	
San Francisco Giants[en]				giants	San Francisco	Giant	Giannts	Gisnts		Ginats		
Mets[en]				mets	New York	Mts	Metts	Mwts		Mest	New York Mets, The Amazin's, The Metropolitans[en]	
Scuderia Ferrari[it]				ferrari	Scuderia Ferrari	Ferari	Ferrarri	Ferrqri		Ferarri	Scuderia Ferrari Marlboro[it]	
Trentino Volley[it]				trentino	Trentino Volley	Trntino	Volleyy	Trentono		Ovlley	Itas Diatec Trentino, Trentino PlanetWin653, Gialloblù[it]	

Table A.10: SportTeam Type Instances

[4] In the cases of teams named after a place, we chose to consider the second best results, because the simple query of the city (e.g., L.A.) would return also results of the city itself.

EducationalOrganization Type Instances	Class of Name Variations											
			Misspellings									
	Full Translation	**Part-of translation**	Punctuation	Capitalization	Spacing	Omissions	Additions	Substitution	Phonetic variations	Switching letters	**Alternative Name**	**Format**
Accademia di Brera[it]	Brera Academy[en]			brera	Brera Academy	Academia	Accaddemy	Brwra		Brear	Academy of Fine Arts of Brera[en], Accademia di Belle Arti di Brera[it]	
Accademia dei Lincei[it]				lincei	Accademia dei	Lince	Linncei	Libcei		Licnei	Lincean Academy[en], Accademia Nazionale dei Lincei[it]	
Columbia University[en]	Universidad de Columbia[es]			columbia	de Columbia	Colmbia	Collumbia	Cplumbia		Colubmia	Columbia University in the City of New York[en], Universidad de Columbia en la Ciudad de Nueva York[es]	
Cornell University[en]	Universidad Cornell[en]			cornell	Cornell University	Cornel	Corrnell	Univerdity		Cronell	Cornell[en]	
Dartmouth College[en]				dartmouth	Dartmouth College	Datmouth	Collegge	Dartmputh		Darmtouth	Dartmouth[en]	
École Normale Supérieure[fr]				école	Normale Supérieure	Normle	Norrmale	Normdle		Nomrale	Normale sup', Normale, ENS[fr]	
Massachusetts Institute of Technology[en]			M.I.T.	mit	Massachusetts Institute	Massachusets	Massacchusetts	Massachysetts		Instiutte	MIT[en]	
Pratt Institute[en]				pratt	Pratt Institute	Prat	Prratt	Prstt		Prtat		
Cambridge University[en]	Universidad de Cambridge[es]			cambridge	of Cambridge	Cambrige	Cammbridge	Cambtidge		Camrbidge	University of Cambridge[en]	
Università Cattolica[it]	Université Catholique du Sacré-Cœur[fr]		U.C.S.C.	cattolica	Università Cattolica	Catolica	Cattollica	Cattplica		Cattloica	Università Cattolica del Sacro Cuore, UCSC[it]	
Universidad Nacional de Asunción[es]				nacional	de Asunción	Univrsidad	Naccional	Asunvión	Asuncion	Asunción	UNA[es]	
Oxford University[en]	Universidad de Oxford[es]										University of Oxford, Oxford[en]	
Università di Padova[it]	University of Padua[en]			padova	di Padova	Padva	Padovva	Padpva		Pauda	Università degli Studi di Padova[it]	
Università del San Raffaele[it]	San Raffaele University[en]	San Raffaele[en]	Vita Salute	san	San Raffaele	Rafaele	Raffaelle	Raffarle		Raffeale	Vita-Salute San Raffaele University[en], Università Vita–Salute San Raffaele[it]	
Università di Trento[it]	Trento University[en]			trento	di Trento	Univrsità	Trennto	Universotà		Univesrità	Università degli Studi di Trento[it]	

Table A.11: EducationalOrganization Type Instances

Country Type Instances	Class of Name Variations											
			Misspellings									
	Full Translation	**Part-of translation**	Punctuation	Capitalization	Spacing	Omissions	Additions	Substitution	Phonetic variations	Switching letters	**Alternative Name**	**Format**
Australia[en]				australia	of Australia	Austrlia	Australlia	Austrslia		Austarlia	Commonwealth of Australia[en]	
Brasil[pt]	Brazil[en]			brasil	do Brasil	Brsil	Brassil	Brasol	Republica	Braisl	República Federativa do Brasil[pt], Federative Republic of Brazil[en]	
Canada[en]				canada		Canda	Cannada	Cansda		Candaa		
Zhongguó[ch]	China[en]		P.R.C.	prc	of China	Repblic	Repubblic	Republoc		Chian	People's Republic of China, PRC[en]	
Côte d'Ivoire[fr]	Ivory Coast[en]		Côte-d'Ivoire	coast	ivory Coast	Ivoir	Ivoiire	Ivoirw		Cosat	Republic of Côte d'Ivoire[en], République de Côte-d'Ivoire[fr]	
Misr[ar]	Egypt[en]			Egypt	of Egypt	Jumhuriya	Eggypt	Egylt		Egpyt	Arab Republic of Egypt[en], Jumhuriyya Misr al-'Arabiyya[ar]	
Deutschland[de]	Germany[en]			germany	of Germany	Deutscland	Deutschlland	Deutschlsnd		Deutshcland	Federal Republic of Germany[en], Bundesrepublik Deutschland[de]	
Nippon[jp]	Japan[en]			japan		Jpan	Jappan	Japsn		Japna	The State of Japan[en], Nihon, Nippon-koku, Nihon-koku[jp]	
Italia[it]	Italy[en]			italia	Italian Republic	Itlia	Itallia	Itslia		Iatlia	Italian Republic[en], Repubblica Italiana[it]	
México[es]	Mexico[en]			mexico	Estados Unidos	Mexic	Mexxico	Mexoco		Meixco	United Mexican States[en], Estados Unidos Mexicanos[es]	
Nederland[nl]	Netherlands[en]			nederland		Nedrland	Nederrland	Nedwrland		Nederalnd	Holland[en]	
South Africa[en]	Sudafrica[it]			sudafrica	South Africa	Afric	Afrrica	Afroca		Afirca	Republic of South Africa[en], Repubblica del Sudafrica[it]	
Daehan Minguk[hl]	South Korea[en]		R.O.K.	rok	South Korea	Kora	Korrea	Korwa		Koera	Republic of Korea, ROK[en]	
U.S.[en]	USA[it]		U.S.A.	usa	of America	Amrica	Amerrica	Ametica		Ameirca	United States of America,United States, USA, America,States[en], Stati Uniti D'America[it]	
Uruguay[es]				uruguay	of Urugay	Uruguay	Urugguay	Urugyay		Urugauy	Répública Oriental del Uruguay[en], Oriental Republic of Uruguay, Eastern Republic of Uruguay[it]	

Table A.12: Country Type Instances

Landform Types Instances	Class of Name Variations											
			Misspellings									
	Full Translation	Part-of translation	Punctuation	Capitalization	Spacing	Omissions	Additions	Substitution	Phonetic variations	Switching letters	Alternative Name	Format
Appalachian Mountains[en]	Appalachi[it]	Mountains[en]		appalachi	Appalachi Mountains	Apalachi	Appallachi	Appalschi		Appalachain	Appalachian Mountains[en]	
Dead Sea[en]	Mar Morto[it]	Sea[en], Mar[it]		dead	Dead Sea	Ded	Deadd	Sdlt		Satl	Salt Sea[en]	
Death Valley[en]	Valle della Morte[it]	Valley[en],Valle[it]		death	della Morte	Valey	Valleey	Vallwy		Valely		
Europe[en]	Europa[it]			europe	il Vecchio	Europ	Eurrope	Eurppe		Euorpe	the Continent[en],il Vecchio Continente[it]	
Lago di Garda[en]	Lake Garda[it]	Lake, Lake of[en], Lago, Lago di[it]		garda	Lake Garda	Grda	Garrda	Gatda		Gadra	Lake of Garda[en], Benaco[it]	Garda Lake
Lake Ontario[en]	Lago Ontario[it]	Lake[en], Lago[it]		ontario	Lake Ontario	Ontaro	Onttario	Ontsrio		Ontraio		Ontario Lake
Mediterranean[en]	Mediterraneo[it]	Mar[it], Sea[en]		mediterranean	Mediterranean Sea	Mediteranean	Meditterranean	Mediterranwan		Mediterrnaean	Mediterranean Sea[en], Mar Mediterraneo[it]	
Mont Blanc[fr]	Monte Bianco[it]	Mont[fr], Monte[it]		mont	Mont Blanc	Mnt	Montte	Mpnt		Blacn	La Dame Blanche[fr], Il Bianco[it]	
Monte Etna[it]	Mount Etna[en]	Mount[en], Monte[it]		etna	Monte Etna	Etn	Etnna	Etn		Enta	Etna[it]	
Everest[en]		Mount[en], Monte[it]		Everest	Mount Everest	Everst	Everrest	Everwst		Everset	Mount Everest[en], Monte Everest[it]	
Campidoglio[it]	Capitoline Hill[en]	Hill[en], Monte[it]		hill	Capitoline Hill	Campdoglio	Campiddoglio	Campidpglio		Campiodglio	Monte Capitolino[it]	
Vesuvio[it]	Vesuvius[it]	Mount[en], Monte[it]		vesuvio	Monte Vesuvio	Vesuvo	Vessuvio	Vesuvuo		Vesvuio	Mount Vesuvius[en], Monte Vesuvio[it]	
Niagara Falls[en]	Cascate del Niagara[it]	Falls[en],Cascate[it]		niagara	Niagara Falls	Niagdra	Niaggara	Niagsra		Niagraa		
Pacific[en]	Oceano Pacifico[it]	Ocean[en], Oceano[it]		pacific	Pacific Ocean	Pacfic	Paciffic	Pacifoc		Pacifci	Pacific Ocean[en], Pacifico[it]	
Río Paraguay[es]	Paraguay River[en]	River[en],Río[es]		río	Río Paraguay	Paragay	Parraguay	Paragyay		Pargauay		

Table A.13: Landform Types Instances

LOrHBs Type Instances	Class of Name Variations											
			Misspellings									
	Full Translation	Part-of translation	Punctuation	Capitalization	Spacing	Omissions	Additions	Substitution	Phonetic variations	Switching letters	Alternative Name	Format
Alcatraz[en]		Island[en],Isola[it]		isola	Alcatraz Island	Alctraz	Alccatraz	Alcatrsz		Alcatarz	Alcatraz Island, The Rock[en], Isola Alcatraz[it]	
Cristo Redentor[pr]	Christ the Redeemer[en]			cristo	Christ the	Redemer	Reddentor	Redwntor		Redeemre		
Brooklyn Bridge[en]	Ponte di Brooklyn[it]	Bridge[en], Ponte di[it]		bridge	Ponte di	Broklyn	Brookklyn	Brooklin		Brooklny		
Colosseo[it]	Colosseum[en]	Amphitheatre[en], Anfiteatro[it]		flavio	Anfiteatro Flavio	Colosswo	Collosseo	Colosswo		Cloosseo	Coliseum, Flavian Amphitheatre[en], Anfiteatro Flavio[it]	
Den Lille Havfrue[dk]	The Little Mermaid[en]			little	The Little	Havfru	Havefrue	Havftue		Havfure		
Tour Eiffel[fr]	Eiffel Tower[en]	Tour[fr], Tower[en]		tour	Tour Eiffel	Eifel	Eiiffel	Eiffwl		Eiffle	La Dame de Fer[fr]	
Great Wall[en]	Grande Muraglia[it]	Wall[en], Muraglia[it]		great	Great Wall	Gret	Murro	Grandw		Muralgia	Great Wall of China[en], Grande Muraglia Cinese[it]	
Mausoleo di Augusto[it]	Mausoleum of Augustus[en]	Mausoleum of[en], Mausoleo di[it]		mausoleo	Mausoleum of	Mausleum	Maussoleum	Augpstus		Augsutus		
Palazzo Pitti[it]	Pitti Palace[en]	Palace[en], Palazzo[it]		pitti	Pitti Palace	Piti	Pallace	Pittu		Palazoz		
Palace of Whitehall[en]	Palazzo di Whitehall[it]	Palazzo di[it], Palace of[en]		di	Palace of	Whitehal	Whittehall	Whitwhall		Whiethall		
Pyramid of Khafre[en]	Piramide di Chefren[it]	Pyramid of[en], Piramide di[it]		khafre	Pyramid of	Piramid	Pirammide	Piramode		Piraimde	Pyramid of Chefren[en]	
Piazza San Pietro[it]	Saint Peter's Square[en]	Square, Saint Peter[en], Piazza, San[it]		pietro	San Pietro	Squre	Squarre	Squsre	S. Peter	Squrae		
Statue of Liberty[en]	Statua della Libertá[it]	Statue of[en], Statua della[it]		statua	Statua della	Librty	Libberty	Libwrty		Librety	Liberty Enlightening the World[en], La Libertá che Illumina il Mondo[it]	
Tiananmen Square[en]	Piazza Tiananmen[it]	Piazza[it], Square[en]		piazza	Piazza Tiananmen	Tianamen	Tiannanmen	Tiansnmen		Tianamnen		
Torre di Pisa[it]	Tower of Pisa[en]	Torre[it], Tower[en]		torre	Torre di	Tore	Pissa	Torte		Pias	Leaning Tower of Pisa[en], Torre Pendente di Pisa[it]	

Table A.14: Landmark Or Historical Buildings (LOrHBs) Type Instances

City Type Instances	Class of Name Variations											
			Misspellings									
	Full Translation	Part-of translation	Punctuation	Capitalization	Spacing	Omissions	Additions	Substitution	Phonetic variations	Switching letters	Alternative Name	Format
al-Qahira[ar]	Cairo[en]		al Qahira	cairo		Caro	Cairro	Cairp		Cario		
Amsterdam[nl]				amsterdam		Amstrdam	Amsterrdam	Amstwrdam		Amstredam		
Beijing[ch]				Beijing		Beijin	Beijiing	Beijung		Beijign	Peking[en]	
Brasília[pt]	Brasilia[it]			Brasília		Brsília	Brassília	Brdsília		Barsília		
Cape Town[en]	Cittá del Capo[it]			cape	Cape Town	Cap	Cappe	Towm		Twon		
London[en]	Londra[it]											
Ciudad de México[es]	Mexico City[es]										City of Mexico[en], México D.F., D.F.[es]	
Mogadishu[en]	Mogadiscio[it]			mogadiscio		Mogdiscio	Mogaddiscio	Mohadiscio		Mogadsicio		
New Delhi[en]	Nuova Delhi[it]			new	New Delhi	Deli	Dellhi	Delho		Dehli		
New York[en]			N.Y.C.	nyc	New York	Ne	Yorrk	Citu		Yrok	New York City, The City of New York, NYC, The Big Apple[en], La Grande Mela[it]	
Paris[fr]	Parigi[it]		La Ville Lumiére	paris	La Ville	Pari	Parris	Paros		Parsi	La Ville-Lumiére[fr]	
Roma[it]	Rome[en]			roma	La Cittá	Rma	Romma	Rona		Rmoa	la Cittá Eterna, Urbe[it]	
Seoul[hl]	Seul[it]			seoul	Seoul Special	Seol	Seooul	Sroul		Soeul	Seoul Special City[en]	
Krung Thep[th]	Bangkok[en]			krung	Thep Maha	Bangok	Banngkok	Bamgkok		Bangokk	Krung Thep Maha Nakhon[th]	
Washington, D.C[en]	Washington[it]		DC	dc	District of	Columba	Washinngton	Washongton		Washignton	District of Columbia Washington, “the Distric”, D.C.[en]	

Table A.15: City Type Instances

Organization Type Instances	Class of Name Variations											
	Full Translation	Part-of translation	Misspellings								Alternative Name/Acronym	Format
			Punctuation	Capitalization	Spacing	Omissions	Additions	Substitution	Phonetic variations	Switching letters		
AA[en]			A.A.	aa	Alcoholics Anonymous	Alcolsti	Anonnimi	Alcplisti		Anoinmi	Alcoholics Anonymous[en], Alcolisti Anonimi[it]	
American Federation of Teachers[en]			A.F.T.	aft	of Teachers	Fedration	Teacherrs	Federstion		Fedeartion	AFT[en]	
Amnesty International[en]	Amnistía Internacional[es]		A.I.	ai	Amnesty International	Amnsty	Amnnesty	Amnwsty		Anmesty	Amnesty, AI[en]	
Confindustria[it]				confindustria	dell' Industria	Cnfindustria	Confinndustria	Confindusstria		Cnofindustria	Confederazione Generale dell'Industria Italiana[it]	
European Union[en]	Unione Europea[it]		E.U.	ue	European Union	Euroean	Unnion	Europwan		Unoin	EU[en], UE[it]	
FIFA[fr]			F.I.F.A.	fifa	Federation of	FFA	FIFFA	FIDA		FIAF	Fédération Internationale de Football Association[fr], International Federation of Association Football[en]	
FIOM[it]			F.I.O.M.	fiom	Impiegati Operai	Fedrazione	Imppiegati	Opwrai		Metallugrici	Federazione Impiegati Operai Metallurgici[it]	
Greenpeace[en]				greenpeace		Grenpeace	Greenppeace	Greenleace		Greepneace		
MSF[fr]			M.S.F.	msf	Doctors Without	Doctor	Witthout	Bprders		Bodrers	Doctors Without Borders[en], Médecins Sans Frontières[fr]	
Oxfam[en]				oxfam	Oxfam International	Oxfm	Oxffam	Oxfsm		Oxfma	Oxfam International[en]	
Red Cross[en]	Croce Rossa[it]			croce	Red Cros	Crross		Crpss		Crcoe		
Save the Children[en]				save	Save the	Childrn	Savve	Childrwn		Svae	The Save the Children Fund[en]	
UN[en]	ONU[it]		U.N.	un	Nazioni Unite	Unitd	Unnited	Unitwd		Natoins	United Nations[en], Nazioni Unite[it]	
World Wide Fund for Nature[en]	Fondo Mundial para la Naturaleza[es]		W.W.F.	wwf	Fund For	Worl	Widde	Funf		Natuer	WWF[en]	

Table A.16: Organization Type Instances

Corporation Type Instances	Class of Name Variations											
	Full Translation	Part-of translation	Misspellings								Alternative Name/Acronym	Format
			Punctuation	Capitalization	Spacing	Omissions	Additions	Substitution	Phonetic variations	Switching letters		
Allianz[de]		SE	S.E.	se	Allianz SE	Alianz	Alliannz	Allidnz		Alilanz	Allianz SE[de]	
Apple[en]		Inc.	Inc	apple	Apple Inc.	Aple	Applle	Spple		Aplpe	Apple Inc.[en]	
AT& T[en]		Corp.	A.T.& T.	at& t	& T	At	AT&& T	AY& T		TA& T	AT& T Corp.[en]	
BP[en]		p.l.c.	B.P.	bp	BP p.l.c.	P	BPP	NP		PB	BP p.l.c.[en]	
Eni[it]		S.p.A.	SpA	eni	Eni S.p.a.	Ei	Enni	Eno		Ein	Eni S.p.A.[it]	
Fiat[it]		S.p.A.	F.i.a.t.	fiat	Fabbrica Italiana	Fit	Fiaat	Fist		Fait	Fiat S.p.A.,Fabbrica Italiana Automobili Torino[it]	
GM		Company	G.M.	gm	General Motors	Genral	Mottors	Genwral		Motros	General Motors, General Motors Company[en]	
IBM[en]		Corporation	I.B.M.	ibm	Business Machines	Busines	Macchines	Machinws		Machiens	International Business Machines Corporation[en]	
McDonald’s[en]		Corporation	McDonalds	mcDonald’s	McDonald’ s	McDonld’s	MacDonald’s	McDonsld’s		McDonlad’s	McDonald’s Corporation[en]	
Nestlé[fr]		S.A.	SA	sa	Nestlé S.A.	Nstlé	Nesstlé	Nedtlé	Nestle	Nestlé	Nestlé S.A.[fr]	
Nintendo[jp]		Co., Ltd.[en]	Nintendo	Nintendo	Nintendo Co.,Ltd.	Nintndo	Ninttendo	Nintwndo		Nintedno	Nintendo Co., Ltd.[en]	
Sony[jp]		Corporation[en]		sony	Sony Corporation	Sny	Sonny	Spny	Soni	Snoy	Sony Corporation[en]	
Starbucks[en]		Corporation		starbucks	Starbucks Corporation	Starbuck	Starrbucks	Starbycks		Starbcuks	Starbucks Corporation[en]	
Vodafone[en]		Plc	P.l.c.	plc	Vodafone Plc	Vodafon	Voddafone	Vodafpne		Vodfaone	Vodafone Group Plc[en]	
Walmart[en]		Inc.	Wal Mart	walmart	Stores Inc.	Walmrt	Wallmart	Walmsrt		Walmrat	WalWalmart-Mart Stores, Inc.[en]	

Table A.17: Corporation Type Instances

Name Type[5]	Person Type Instances	Class of Name Variations											
		Full Translation	**Part-of translation**	**Misspellings**									
				Punctuation	Capitalization	Spacing	Omissions	Additions	Substitution	Phonetic variations	Switching letters	**Alternative Name**	**Format**
Common Names	**Alex Catania Zampieri[it]**			AC	ac	Alex Catania	Zamperi	Allex	Catanoa		Aelx	Alex Catania, Alex Zampieri, A. Catania, A. Zampieri, A. C., A. Z., A. C. Z.[it]	Alex Zampieri Catania
	Enrico Bignotti[it]		Sig.[it]	EB	enrico	E. B.	Bignoti	Bignnotti	Bignptti		Bingotti	E. Bignotti, Enrico B., Bigno, Chicco, E. B., Sig. Bignotti Enrico[it]	Bignotti Enrico, Sig. Enrico Bignotti
	Fausto Giunchiglia[it]		Prof.[it]		prof	Fausto Giunchiglia	Giunchilia	Giunnchiglia	Giynchiglia		Fasuto	Papi,Prof. Giunchiglia[it]	Giunchiglia Fausto
Pen Name	**Freddie Mercury[en]**				freddie	Freddie Mercury	Mrcury	Merrcury	Mercyry	Freddy	Merucry	Farookh Bulsara[en]	Bulsara Farookh
	Mark Twain[en]				mark	Mark Twain	Twin	Twainn	Twsin		Tawin	Samuel Clemens, Samuel Langhorne Clemens, Samuel L. Clemens[en]	Clemens Samuel
	Slash[en]				saul	Saul Hudson	Slsh	Slassh	Slssh		Slsah	Saul Hudson, S. Hudson, Saul H., S. H.[en]	Hudson Saul
Spanish Format	**Franco[es]**				franco	Francisco Franco	Caudilo	FRancissco	Francp		Franicsco	Francisco Franco y Bahamonde,, Generalísimo Franco, Caudillo de España[es]	Franco y Bahamonde Francisco
	José Chilavert[es]				josé	José Luis	Féli	Chillavert	Gonzálwz		Joés	José Luis Félix Chilavert González[es]	Chilavert José
	Juan Pane[es]				juan	Juan Ignacio	Pan	Panne	Panw		Jaun	Juan Ignacio Pane Fernández, Juanito[es]	Pane Juan
Middle Name	**Alberto Asor Rosa[it]**				asor	Asor Rosa	Albert	Assor	Rosd		Albreto		Rosa Asor Alberto
	Philip Dick[en]			K	philip	Philip Dick	Pilip	Phillip	Pholip		Dikc	Philip K. Dick, Philip Kindred Dick[en]	Dick Philip
	William Carlos Williams[en]				C	William Carlos	Wiliam	Williamm	Willism		Willaim	William C. Williams[en]	Williams C. William

Table A.18: Person Type Instances

[5] See section 3.1 for a definition of Pen Name, and section 4.4 for a definition of Spanish Format and Middle Name

Name Subtype	Name Type Instances	Class of Name Variations											
				Misspellings									
		Full Translation	Part-of translation	Punctuation	Capitalization	Spacing	Omissions	Additions	Substitution	Phonetic variations	Switching letters	Alternative Name	Format
Religious Figures and Monarchy[6]	**King Henry VII[en]**	Re Enrico VIII[it]	Re[it],King[en]	Henry xviii	henry	Henry VIII	Kin	Hennry	Enricp		Kign	Henry VIII of England, King Henry VIII[en], Enrico VIII d'Inghilterra, Re Enrico VIII[it]	
	Dronning Margrethe II[dk]	Queen Margrethe II[en]	Dronning[dk],Queen[en]	Margrethe ii	margrethe	Queen Margrethe	Margrete	Marrgrethe	Margrwthe		Margerthe	Margaret II, Margrethe II[en], Margrethe II, Margrethe Alexandrine dórhildur Ingrid[dk]	
	John Paul II[en]	Giovanni Paolo II[it]	Pope,Blessed[en],Papa,Beato[it]		papa	Paul II	Karl	Giovvanni	Karpl		Paloo	Karol Józef Wojtyła[po], Beato Giovanni Paolo II, Papa Giovanni Paolo II[it], Pope John Paul II, Blessed Pope John Paul II[en]	
	Antonio da Padova[it]	Anthony of Padua[en]	Sant', S.[it], Saint, St[en]	St.	st	St Anthony	Antony	Anthonny	Antjony	St Anthony	Antohny	Anthony of Lisbon, Saint Anthony of Padua, Saint Anthony of Lisbon, Saint Anthony, St Anthony of Lisbon, St Anthony of Padua, Fernando Martins de Bulhes[en], [it]	
	Saint Henry[en]	Sant'Enrico[it]	Sant', S., Re[it], Saint,St, King[en]	Henry ii	henry	Henry II	Enric	Ennrico	Enroco	St Henry	Enrcio	Henry II, King Henry II, St Henry[en], Re Enrico II, Enrico II[it]	
Translated Name	**Julius Caesar[en]**	Giulio Cesare[it]			giulio	Giulio Cesare	Cesre	Cessare	Cessre		Ceasre	Gaius Julius Caesar[en], Gaio Giulio Cesare[it]	Cesare Giuglio Gaio
	Napoléon[fr]	Napoleon[en]			napoleon	Bonaparte	Bonapart	Napolleon	Bonsparte		Bonaprate	Napoleon Bonaparte, Napoleon I[en], Napoléon Bonaparte, Napoléon I[it]	Bonaparte Napoleon
	Aristotéles[gk]	Aristotle[en]			aristotle		Aristotl	Arristotle	Aristptle		Aritsotle		

Table A.19: Religious Figures and Monarchy Person Instances

[6] See section 7.1 for a definition of this subtype

Appendix B

Schema.org Full Type List

The types in italics are the five main entity types, whereas the bold one are those chose as entity types that undergo multilingual name variations

CreatieWork
Article
BlogPosting
NewsArticle
ScholarlyArticle
Blog
Book
ItemList
Map
MediaObject
AudioObject
ImageObject
MusicideoObject
VideoObjec
Movie
MusicPlaylist
MusicAlbum
MusicRecording
Painting
Photograph
Recipe
Reiew
Sculpture
TvEpisode
TvSeason
TvSeries
WebPage
AboutPage
CheckoutPage
CollectionPage
ImageGallery
VideoGallery
ContactPage
ItemPage
ProfilePage
SearchResultsPage
WebPageElement
SiteNaigationElement
Table
WPAdBlock
WPFooter
WPHeader
WPSideBar
Event
BusinessEvent
ChildrensEvent
ComedyEvent
DanceEvent
EducationEvent
Festial
FoodEvent
LiteraryEvent

MusicEvent
SaleEvent
SocialEvent
SportsEvent
TheaterEvent
UserInteraction
UserBlocks
UserCheckins
UserComments
UserDownloads
UserLikes
UserPageisits
UserPlays
UserPlusOnes
UserTweets
VisualArtsEvent
Intangible
Enumeration
BookFormatType
ItemAailability
OfferItemCondition
JobPosting
Language
Offer
AggregateOffer
Quantity
Distance
Duration
Energy
Mass
Rating
AggregateRating
StructuredValue
ContactPoint
PostalAddress
GeoCoordinates
GeoShape
NutritionInformation
Organization
Corporation
EducationalOrganization
CollegeOrUniersity
ElementarySchool
HighSchool
MiddleSchool
Preschool
School
GovernmentOrganization
LocalBusiness
AnimalShelter
AutomotieBusiness
AutoBodyShop
AutoDealer
AutoPartsStore
AutoRental
AutoRepair
AutoWash
GasStation
MotorcycleDealer
MotorcycleRepair
ChildCare
DryCleaningOrLaundry
EmergencyService
FireStation
Hospital
PoliceStation
EmploymentAgency
EntertainmentBusiness
AdultEntertainment
AmusementPark
ArtGallery
Casino
ComedyClub
MovieTheater
NightClub
FinancialService
AccountingService
AutomatedTeller
BankOrCreditUnion
InsuranceAgency
FoodEstablishment
Bakery

BarOrPub
Brewery
CafeOrCoffeeShop
FastFoodRestaurant
IceCreamShop
Restaurant
Winery
GovernmentOffice
PostOffice
HealthAndBeautyBusiness
BeautySalon
DaySpa
HairSalon
HealthClub
NailSalon
TattooParlor
HomeAndConstructionBusiness
Electrician
GeneralContractor
HACBusiness
HousePainter
Locksmith
MoingCompany
Plumber
RoofingContractor
InternetCafe
Library
LodgingBusiness
BedAndBreakfast
Hostel
Hotel
Motel
MedicalOrganization
Dentist
Hospital*
MedicalClinic
Optician
Pharmacy
Physician
VeterinaryCare
ProfessionalService
AccountingService*
Attorney
Dentist*
Electrician*
GeneralContractor*
HousePainter*
Locksmith*
Notary
Plumber*
RoofingContractor*
RadioStation
RealEstateAgent
RecyclingCenter
SelfStorage
ShoppingCenter
SportsActiityLocation
BowlingAlley
ExerciseGym
GolfCourse
HealthClub*
PublicSwimmingPool
SkiResort
SportsClub
StadiumOrArena
TennisComplex
Store
AutoPartsStore*
BikeStore
BookStore
ClothingStore
ComputerStore
ConenienceStore
DepartmentStore
ElectronicsStore
Florist
FurnitureStore
GardenStore
GroceryStore
HardwareStore
HobbyShop
HomeGoodsStore

JewelryStore
LiquorStore
MensClothingStore
MobilePhoneStore
MovieRentalStore
MusicStore
OfficeEquipmentStore
OutletStore
PawnShop
PetStore
ShoeStore
SportingGoodsStore
TireShop
ToyStore
WholesaleStore
TelevisionStation
TouristInformationCenter
TravelAgency
NGO
PerformingGroup
DanceGroup
MusicGroup
TheaterGroup
SportsTeam
Person
Place
AdministrativeArea
City
Country
State
CivicStructure
Airport
Aquarium
Beach
BusStation
BusStop
Campground
Cemetery
Crematorium
Eventenue
FireStation*
GovernmentBuilding
CityHall
Courthouse
DefenceEstablishment
Embassy
LegislatieBuilding
Hospital*
MovieTheater*
Museum
Musicvenue
Park
ParkingFacility
PerformingArtsTheater
PlaceOfWorship
BuddhistTemple
CatholicChurch
Church
HinduTemple
Mosque
Synagogue
Playground
PoliceStation*
RPark
StadiumOrArena*
SubwayStation
TaxiStand
TrainStation
Zoo
Landform
BodyOfWater
Canal
LakeBodyOfWater
OceanBodyOfWater
Pond
Reseroir
RierBodyOfWater
SeaBodyOfWater
Waterfall
Continent
Mountain
Volcano

LandmarksOrHistoricalBuildings
LocalBusiness*
AnimalShelter
AutomotieBusiness
AutoBodyShop
AutoDealer
AutoPartsStore*
AutoRental
AutoRepair
AutoWash
GasStation
MotorcycleDealer
MotorcycleRepair
ChildCare
DryCleaningOrLaundry
EmergencyService
FireStation*
Hospital*
PoliceStation*
EmploymentAgency
EntertainmentBusiness
AdultEntertainment
AmusementPark
ArtGallery
Casino
ComedyClub
MovieTheater*
NightClub
FinancialService
AccountingService*
AutomatedTeller
BankOrCreditUnion
InsuranceAgency
FoodEstablishment
Bakery
BarOrPub
Brewery
CafeOrCoffeeShop
FastFoodRestaurant
IceCreamShop
Restaurant
Winery
GovernmentOffice
PostOffice
HealthAndBeautyBusiness
BeautySalon
DaySpa
HairSalon
HealthClub*
NailSalon
TattooParlor
HomeAndConstructionBusiness
Electrician*
GeneralContractor*
HACBusiness
HousePainter*
Locksmith*
MovingCompany
Plumber*
RoofingContractor*
InternetCafe
Library
LodgingBusiness
BedAndBreakfast
Hostel
Hotel
Motel
MedicalOrganization
Dentist*
Hospital*
MedicalClinic
Optician
Pharmacy
Physician
eterinaryCare
ProfessionalService
AccountingService*
Attorney
Dentist*
Electrician*
GeneralContractor*
HousePainter*

Locksmith*
Notary
Plumber*
RoofingContractor*
RadioStation
RealEstateAgent
RecyclingCenter
SelfStorage (***)
ShoppingCenter
SportsActiityLocation
BowlingAlley
ExerciseGym
GolfCourse
HealthClub*
PublicSwimmingPool
SkiResort
SportsClub
StadiumOrArena*
TennisComplex
Store
AutoPartsStore*
BikeStore
BookStore
ClothingStore
ComputerStore
ConvenienceStore
DepartmentStore
ElectronicsStore
Florist
FurnitureStore
GardenStore
GroceryStore
HardwareStore
HobbyShop
HomeGoodsStore
JewelryStore
LiquorStore
MensClothingStore
MobilePhoneStore
MovieRentalStore
MusicStore
OfficeEquipmentStore
OutletStore
PawnShop
PetStore
ShoeStore
SportingGoodsStore
TireShop
ToyStore
WholesaleStore
TelevisionStation
TouristInformationCenter
TravelAgency
Residence
ApartmentComplex
GatedResidenceCommunity
SingleFamilyResidence
TouristAttraction
Product

Bibliography

[1] M. Aceto, *Ethnic Personal Names and Multiple Identities in Anglophone Caribbean Speech Communities in Latin America*, Language in Society, Vol. 31, pp. 577–608, (2002)

[2] M. Bilenko, W. Cohen, S. Fienberg, R. Mooney, P. Ravikumar, *Adaptive Name-Matching in Information Integration*, IEEE Intelligent Systems, Vol. 18, No. 5, pp. 16-23, (2003)

[3] L. K. Branting, *Name-Matching Algorithms for Legal Case-Management Systems*, The Journal of Information, Law and Technology (JILT), Vol. 1, (2002)

[4] L. K. Branting, *Name Matching in Law Enforcement and Counter-Terrorism*, ICAIL 2005 Workshop on Data Mining, Information Extraction, and Evidentiary Reasoning for Law Enforcement and Counter-Terrorism, Bologna, Italy (2005)

[5] R. Bunescu, *Using Encyclopedic Knowledge for Named Entity Disambiguation*, In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 9–16, (2006)

[6] P. Chen, *The Entity-Relationship Model: Toward a Unified View of Data*, ACM Transactions on Database Systems, Vo. 1, pp. 9-36, (1976)

[7] P. Christen, *A Comparison of Personal Name Matching: Techniques and Practical Issues*, In *Workshop on Mining Complex Data (2006)*, pp. 290–294, (2006)

[8] W. Cohen, P. Ravikumar, S. Fienberg, *A Comparison of String Metrics for Matching Names and Records*, KDD Workshop On Data Cleaning And Object Consolidation, In *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web*, pp. 73–78, (2003)

[9] Department of Economic and Social Affairs Statistic Division, United Nations Group of Expert on Geographical Names, *Glossary of Terms for the Standardization of Geographical Names*, United Nations, New York, (2000)

[10] M. Devitt, *Designation*, New York: Columbia University Press, (1981)

[11] P. Driscoll, D. Yarowsky, *Disambiguation of Standardized Personal Name Variants*, In *Proceedings of Multisource, Multilingual Information Extraction and Summarization, RANLP*, (2007)

[12] G. Evans, *The Varieties of Reference*, Oxford University Press, Oxford, (1982)

[13] A. Fader, S. Soderl, O. Etzioni, *Scaling Wikipedia-based Named Entity Disambiguation to Arbitrary Web Text*, in *Proceedings of WikiAI*, (2009)

[14] D. Farmakiotou, V. Karkaletsis, J. Koutsias, G. Sigletos, C. Spyropoulos, P. Stamatopoulos, *Rule-Based Named Entity Recognition For Greek Financial Texts*, In *Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries*, pp. 75–78, (2000)

[15] I. Fellegi, A. Sunter, *A Theory for Record Linkage*, Journal of the American Statistical Association Vol. 64, No. 328, pp. 1183–1210, (1969)

[16] G., Frege, *Über Sinn und Bedeutung* In *Zeitschrift für Philosophie und Philosophische Kritik*, pp. 25–50, 1892. Translation by M. Black In *Translations from the Philosophical Writings of Gottlob Frege*, P. Geach, M. Black, Blackwell, Oxford, (1980)

[17] W. Gao, K. Wong, W. Lam , *Phoneme-based Transliteration of Foreign Names for OOV Problem*, In *Proceedings of the First International Joint Conference on Natural Language Processing*, pp. 110–119, (2005)

[18] A. Gentile, Z. Zhang, L. Xia, J. Iria, *Graph-based Semantic Relatedness for Named Entity Disambiguation*, Serdica Journal of Computing, Vol. 2, pp. 28–34, (2010)

[19] A. Ghodsi, *Distributed k-ary System: Algorithms for Distributed Hash Tables*, KTH-Royal Institute of Technology, PhD Thesis, (2006)

[20] N. Habash, *Four Techniques for Online Handling of Out-of-Vocabulary Words in Arabic-English Statistical Machine Translation*, Annual

Meeting of the Association of Computational Linguistics, pp. 57–60, (2008)

[21] L. Hill, *Georeferencing: The Geographic Associations of Information*, Digital Libraries and Electronic Publishing W. Arms, (2006)

[22] J. Hoffart, F. Suchanek, K. Berberich, G. Weikum, *Yago2. A Spatially and Temporally Enhanced Knowledge Base from Wikipedia*, Technical Report, (2010)

[23] G. Holloway and M. Dunkerley,*The Math, Myth and Magic of Name Search and Matching*, SearchSoftwareAmerica, (1999)

[24] F. Huang, *Multilingual Named Entity Extraction and Translation from Text and Speech*, Phd Dissertation, Carnegie Mellon University Pittsburgh, (2006)

[25] R. Huddleston, *English Grammar. An Outline*, Cambridge University Press, (1988)

[26] R. Huddleston, G. Pullum, *A Student's Introduction to English Grammar*, Cambridge University Press, Cambridge, (2005)

[27] A. Hume, *Distributed Entity Search*, Technical Report no. DISI-11-462, (2011)

[28] P. Jordan, *Language and Place Names in National and Regional Atlases. Methodological Considerations and Practical Use Exemplified by New Atlases From the Eastern Part of Europe*, International Cartographic Conference, (2005)

[29] J. Katz, *Names without Bearers*, Philosophical Review, Vol. 103, pp. 1–39, (1994)

[30] D. Kladnik, *Characteristics of Exonym Use in Selected European Languages*, Acta Geographica Slovenica, Vol. 47, pp. 199–222, (2007)

[31] S. Kripke, *Naming and Necessity*, Harvard University Press, Cambridge, (1980)

[32] R. Larson, *Information Retrieval: Searching in the 21st Century; Human Information Retrieval*, J. Am. Soc. Inf. Sci. Technol., Vol. 61, pp. 2370–2372, (2010)

[33] D. Lawrie, J. Mayeld, P. McNamee, D. Oard, *Creating and Curating a Cross-Language Person-Entity Linking Collection*, In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, (2012)

[34] K. Lindén, *Multilingual Modelling of Cross-Lingual Spelling Variants*, Information Retrieval, Vo. 9, pp. 295–310, (2006)

[35] C. Lucock, M. Yeo, *Naming Names: The Pseudonym in the Name of the Law*, University of Ottawa Law and Technology Journal, Vol. 3, No. 1, (2006)

[36] T. Mandl , C. Womser-Hacker, *Proper Names in the Multilingual CLEF Topic Set*, In *Proceedings of the CLEF 2003 Workshop*, pp. 21–28, (2003)

[37] S. Mann and D. Yarowsky, *Multipath Translation Lexicon Induction Via Bridge Languages*, In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pp. 1–8, (2001)

[38] P. McNamee, J. Mayfield, and C. Piatko, *Processing Named Entities in Text*, John Hopkins APL Techical Digest, Technical Report, Vol. 30, No. 1, (2011)

[39] I. Michna, *Generalization of Geographical Names on Atlas maps*, in *Polish Cartographical Review*, Vol. 40, No. 1, pp. 28–45, (2008)

[40] J. Mill, *A System of Logic, Ratiocinative and Inductive*, University Press of the Pacific, Honolulu, (2002)

[41] S. Monahan, *Cross-Lingual Cross-Document Coreference with Entity Linking*, in *Proceedings of the 17th ACM Conference on Information and knowledge Management*, pp. 1359–1360, (2008)

[42] E. Moreau, F. Yvon, O. Cappé, *Robust Similarity Measures for Named Entities Matching*, In *Proceedings of the 22nd International Conference on Computational Linguistics*, Vol. 1, pp. 593–600, (2008)

[43] G. Navarro, *A Guided Tour to Approximate String Matching*, ACM Computing Surveys, Vol. 33, pp. 31–88, (2001)

[44] O. Piton, T. Grass, D. Maurel, *Linguistic Resource for NLP: Ask for Die Drei Musketiere and Meet Les Trois Mousquetaires*, in *Natural*

Language Processing and Information Systems, 8th International Conference on Applications of Natural Language to Information Systems, Vol. 29, pp. 200–213, (2003)

[45] U. Pfeifer, T. Poersch, N. Fuhr, *Retrieval Effectiveness of Proper Name Search Methods*, Information Processing and Management, pp. 667–679, (1996)

[46] B. Pouliquen, R. Steinberger, C. Ignat, I. Temnikova, A. Widiger, W. Zaghouani, J. Žižka, *Multilingual Person Name Recognition and Transliteration*, CoRR, Vol. 3, No. 2, pp. 115–123. (2006)

[47] D. Rao, P. McNamee, M. Dredze, *Entity Linking. Finding Extracted Entities in a Knowledge Base*, In *Multi-source, Multi-lingual Information Extraction and Summarization*, pp. 93–115, (2011)

[48] J. Raukko, *A Linguistic Classification of Exonyms. With a Case study of the Names of 100 European Cities in Eight European Languages*, In *Exonyms and the International Standardisation of Geographical Names. Approaches Towards the Resolution of an Apparent Contradiction* edited by A. Jordan, P. Jordan, M. Oroen Adamic, LIT Verlag Berlin-Hamburg-Münster, pp. 240, (2007)

[49] M. Rössler, *Corpus-based Learning of Lexical Resources for German Named Entity Recognition*, In *Proceedings of LREC*, (2004)

[50] B. Russell, *On Denoting*, Mind, New Series, Vol. 14, No. 56, pp. 479–493, (1905)

[51] J. Searle, *Speech Acts: An Essay in the Philosophy of Language*, Cambridge University Press, Cambridge, (1969)

[52] C. Snae, *A Comparison and Analysis of Name Matching Algorithms*, in *International Journal of Applied Science. Engineering and Technology*, Vol. 9, pp. 252–257, (2007)

[53] A. Spear, *Searle and Kripke on the Description Theory of Proper Names*, Draft, University at Buffalo, (2009)

[54] A. Valarakos, R. Valarakos, G. Paliouras, V. Karkaletsis, G. Vouros, *A Name-Matching Algorithm for Supporting Ontology Enrichment*, In *Proceedings of SETN, 3rd Hellenic Conference on Artificial Intelligence*, (2004)

[55] N. Wacholder, Y. Ravin, M. Choi, *Disambiguation of Proper Names in Text*, In *Proceedings of the 5th Applied Natural Language Processing Conference*, pp. 202–208, (1997)

[56] W. Wentland, J. Knopp, C. Silberer, M. Hartung, *Building a Multilingual Lexical Resource for Named Entity Disambiguation, Translation and Transliteration*, Language Resources and Evaluation Conference, (2008)

[57] W. Zhang, J. Su ,C. Lim, T. Wen, T. Wang, *Entity Linking Leveraging Automatically Generated Annotation*, In *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 1290-1-298, (2010)