Concept and Rule Based Naming System

Chakkrit Snae and Michael Brückner
Naresuan University, Phitsanulok, Thailand

chakkrits@nu.ac.th michaelb@nu.ac.th

Abstract

Names are important in many societies, even in technologically oriented ones which use e.g. ID systems to identify individual people. There are many elements of personal names which vary in different cultures. Names such as surnames are the most important as they are used in many processes, such as identifying of people and genealogical research. On the other hand variation of names can be a major problem for the identification and search for people, e.g. web search or security reasons. We show name variations for different cultures to guide the implementation of a rule based naming system, currently worked out for Thai names. We characterize the LIG (Levenshtein, Index of Similarity Group (called ISG), and Guth) algorithms which help to find reasonable variants of names and use an ontology of names to capture the meaning of the variants which are based on Thai naming methodologies and rules. A further benefit of this process is an optimized name searching.

Keywords: personal names, name variations, name matching, ontology, rule based system

Introduction

Names are used for identifying persons, places, things and even ideas or concepts. Names serve for labelling of categories or classes and for individual items. They are properties of individuals which are of greater importance in most communities. In technologically oriented societies such as modern Western the reference between a name as a label and the person is not as obvious as in small tribal societies. This is especially true where names are stored within large information systems. This includes government, medical, educational and even commercial records which are kept about individuals. Names are the most important referrer to a person even if there are numbering systems like ID numbers because such systems are not universal. Names are often queried in a different way than they were entered. Personal names lead to many problems with regard to data retrieval because names are also subject to multiple variations not only between different cultures and writing systems but in a specific culture as well. Names represent complex lexical structures which have to be handled systematically for data entry, storage and retrieval in order to get sufficient recall or precision in the retrieval process.

In this paper we present a first account of our findings on elements of personal names in different cultures with their respective conventions and variations, ontology of names, rules for constructing Thai names, and algorithms of name matching to overcome name variations. Here we present a hybrid name matching procedure which is based on probabilistic phonetic and sound variation recognition with the help of an expert

system and which can deal with multicultural names as well. This procedure is used in the Thai naming system.

This paper is organized as follows: Section 2 contains a description of names and their variation. In Section 3 we outline the basic term of ontology which is used in Section 4 for the conceptualisation of names. In Section 5 we present some basic characteristics of the name matching algorithms of our choice, the LIG algorithms, as well as the rule based system for naming applied to Thai names. Section 6 shows the conclusions of our study and further work which has to be performed.

What is a Name?

Names for individuals are often called proper names, for humans sometimes also anthroponyms. Names for places are called toponyms, for bodies of water hydronyms, for ethnic groups ethnonyms, for metaphors metonyms and so on. Names are more than just strings of characters. Names show important information such as titles, gender, marital status, and even birthplace. For this reason names provide different elements, which may differ amongst cultures. They also undergo variations such as phonetic and alternate spellings. This problem can be made clear in this way: how do we know that differently spelled or pronounced names belong to the same person?

Each culture has a set of conventions which govern the appearance and function of names as well as a range of permitted variations in its naming system, e.g. surname inheritance. Some cultures show the marital status very clearly, others do not refer to it, cf. Thai and German differences in stating the marital status for female. The categories of name variation are generally the same across all cultures; the difference is in the realization of the variation.

Naming for Identity and Security

From the technical point of view we want to link and match as many names as possible with the correct individuals. If we deal with individuals of the same name, e.g. John Smith, we have to establish a second identifier at least. This can be – and is in many cases – a temporal element, like the date of birth, which is an individual and unchanging property of the person. Another way to circumvent the problem is to establish numbering systems, like ID numbers. Systems of numbers or other ciphers can be generated within individual organizations. It is not likely that the resulting ID numbers will be the same in different organizations. The numbering may have limitations as well, e.g. the individual health care setting (e.g. within a hospital or district) or, in principle, more widely (e.g. the National Health Service number). In the past, the National Health Service number in England and Wales had serious limitations as a matching variable, and it was not widely used on health-care records. With the allocation of the new ten-digit number throughout the NHS all this has been changed (Gill, 1997).

Although numbering systems are simple to implement they can lead to different errors in recording, transcription, and keying. So we have to take into account methods which reduce these errors and facilitate good quality of data entry and retrieval. One such method uses a checking device such as check-digits (Wild, 1968; Hamming, 1986). When we are not able to use unique numbers or ciphers, natural matching variables are the person's name, date of birth, sex and perhaps other supplementary variables such as the address with postal code and place of birth, which are used in combination for matching. Recently, it has been suggested that this simple code could be extended for security critical places (e.g. airports, checkpoints etc.) with biometric marker information extracted from person identifier information e.g. fingerprints/iridograms.

Elements of Personal Names

Figure 1 shows typical elements of personal names together with potential variations and sources of choices, e.g. dictionary of given names.

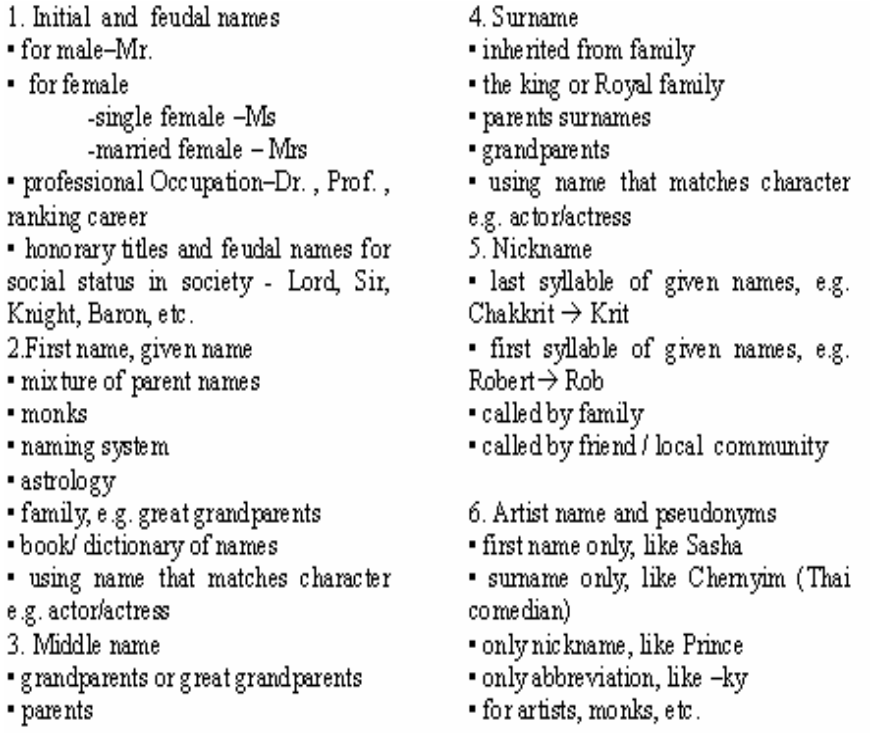

Figure 1: The elements of names

Ontology of Names

The term ontology has been widely used in recent years in the field of Artificial Intelligence, computer and information science especially in domains such as, cooperative information systems, intelligent information integration, information retrieval and extraction, knowledge representation, and database management systems (Andrade & Saltz, 1999, 2000; Guarino, 1998). Many different definitions of the term are proposed. One of the most widely quoted and well-known definition of ontology is Gruber's (Gruber, 1993): An ontology is an explicit specification of a conceptualization.

The term is borrowed from philosophy, where an ontology is a systematic account of existence. Here in this paper we adopt the following definition: Ontology is the study or concern about what kinds of things exist - what entities or things are there in the universe (Blackburn, 1996). Our work on ontologies will comprise: a terminological component where we lay down the concepts and an assertional component (or Knowledge Base) which contains the individual instances (entities).The level of description will be taxonomies with hierarchically related terms and controlled vocabularies (thesaurus) with the help of semantic networks.

In this research we use ontologies for different reasons. Firstly, we want to find out the gender for a specific Thai given name. We captured the meaning of approximately 10,000 Thai given names in a database and can therefore derive the gender to which it belongs with the help of an ontology. Secondly, we use the same strategy for naming baby girls and baby boys. Here the users can check for appropriate given names according to the gender of their babies and for names with a helpful meaning and good sound. An ontology can help to combine different names, e.g. parents'

names, into a new name with a meaning. For example father's name Suchat, mother's name Wipada, and so a baby girl can take the syllables Sucha- from father and -da from mother to become Suchada. This new name means "beautiful girl". Thirdly we want to capture given names belonging to the same person and which have many variations. For example, we have the name Chakkrit with the spelling variations Jakkrit, Chakrid, Chakkrid, Jagrid, Chakkit, etc. In Thai these different spellings result in the same sound. Ontologies are used when we Romanize Thai characters into Latin alphabet. Ontologies help the user to choose the appropriate Romanized version of the Thai name. This works also the other way. Consider the name Peter and its short form Pete. By using an ontology we find that both names can refer to the same person. In this case ontologies work like a thesaurus for given names.

An ontology of names can be worked out in many different forms, but every ontology will include a dictionary, some definition of the terms (semantics), and indications how they depend on each other, e.g. in hierarchies and semantic networks. For example, an ontology of names can be defined as what kinds of names exist, e.g. first name, surname, nickname, etc. This typically comprises definitions of different names, the elements of names and their structures. In this section we show how an ontology of names can be captured and defined.

An ontology can also be used to establish the network of synonyms, e.g. using spelling norms to determine whether two names are the same/similar or not. For example, two names: Totie and Totiey can be defined based on assumption that they are the same as Totty. This attempts to tackle the seemingly irreducible conventions of surname. In compositional semantics let us consider the name "Gutensohn". This name will be used to illustrate the various semantic considerations in German naming. The name is a composition of the two strings Godith and Sohn, which have unambiguous, meaningful interpretations. The interpretation of Godith is god or good battle and Sohn is interpreted as a male child in relation to his parent. The composition Gutsohn, Gudzon, or in other cultures: Guditson, Godyeson and Godithson and Godison (Reaney & Wilson 1997).

Identifying and searching result in a list of many names with variations and meanings. In order to find the correct person with a name we have to adopt ontologies of names, e.g. based on place of birth or relationship of people. The typical origins of surnames which can be a basis for ontologies of names can be classified as follows: local surnames - surnames of relationship - surnames of occupation or office.

Local surnames, that are most widely used, stem from toponyms, we can call them toponymic. They reflect land owners, place of birth, or the center of life. For example, Richard de Tonebridge was named after his castle of Tonbridge, but he was also called Richard de Clara from the Suffolk Clare, which became his chief seat and the family's definitive surname. Also Richard de Hadestoke, a London alderman, had left Hadstock (Essex) and settled in London (Reaney & Wilson, 1997). These local surnames derive (with occasional exceptions) from English, Scottish or French places (e.g. de, at, in). Toponymic Thai names are derived from Thai places and took originally a preposition na, for example, Prapas na Ranong is a person from a Southern province in Thailand called Ranong.

Surnames which come from family relation are often called patronymic, but we have to introduce a more elaborate term, because we encounter names from females and other relations than just father, such as Gilbert Fathevedsteppeson, Richard Hannebrothir, America Ibbotdoghter, and John Prestebruther.

Surnames of occupation and office refer to actual office holders like clergy names or state offices. Some of these, such as steward, constable, marshal, etc., became hereditary and gave rise to hereditary surnames, but the terms were also commonly used of lesser offices, whilst marshal was a common term for a farrier and such names frequently denoted the actual occupation. However,

Nuns, Abbots, Priors, Monks and other clerical people were bound by vows of celibacy and thus did usually not have families which adopted their respective surname.

We have to bear in mind that a full and accurate classification of surnames is impossible as there is a significant number of overlap between the different classes of names. People did not follow special rules in the naming process rather they wanted to solve an immediate problem: identifying and characterizing themselves or others. Nicknames, in particular, are even today often the result of a spontaneous reaction in a particular situation. Local surnames may be occupational, for example, John atte Gate may have lived near the town-gate, or he may have been a gate-keeper or porter. Surnames of office, such as Abbot, Bishop, Cardinal and King, are often nicknames. Ralph Vicar was a glass worker, not a clergyman, and is also called Verrer. A single modern name may belong to more than one class. For instance, Mew may be a patronymic, a nickname from the sea-mew, or occupational, either metonymic for Mewer, keeper of the hawks, or from a local surname, with the same occupational meaning. It is impossible to fit surnames into a straitjacket (Reaney & Wilson, 1997).

Variation of Names

Name variation is one of the major problems in identifying people, because it is not easy to determine whether a name variation is a different spelling of the same name or a name for a different person. Most of these variations can be mainly categorized as mentioned in the Sections below (Bouchard and Pouyez, 1980; Branting, 2002; Dematteis, Lutz & McCallum-Bayliss, 1998).

From searching on the Internet for some personal names, e.g. in Thailand, we have found many variants of them which refer to the same names. Table 1 shows the results of variants of the personal name called “Somchai” using Google search in many search processes.

Table 1: Variants of ”Somchai” using Google search

Personal name	Variation	Number of results
Somchai	Somchai	200,000
	Som Chai	149,000
	Somchay	1,640
	Somshai	231
	Somchia	77
	Somchair	48
	Somchaiy	35
	Somcai	17

As can be seen from Table 1, personal name like Somchai, Som Chai, and Somchay are the most similar but names like Somchair, Somchaiy, and Somcai exhibit the least similarity. With the help of name matching methods we find reasonable alternatives of the original name, e.g. Somchai. Then all alternative names of Somchai can be used in one single search process which covers all variants at once.

Spelling variations

Spelling variations rely on the assumption that the source and target names are strings which differ because of errors or transcription differences (e.g. different pronunciation). Spelling error patterns can be taken into consideration and single-error misspellings (mistyping) can be categorized as follows (Jurafsky & Martin, 2000): (1) insertion, e.g. BROWN as BROWMN, and MCMANUS as MACMANUS; (2) deletion or omission, e.g. BROWN as BOWN, and ROBBIN as ROBIN; (3) substitution, e.g. BROWN as BTOWN, and SMYTH as SMITH; (4) transposition, e.g. BROWN as BRONW, and BREADLEY and BRAEDLEY. Generally such variations do not

affect the phonetic structure of the name but still cause problems in matching names. These variations mainly arise from misreading or mishearing, by either a human or an automated device. These can include interchanged or misplaced letters due to vowel replacement (EVANTUREL as EVENTUREL), consonant replacement (LEBRE as LETRE), consonant doubling (MAUFET as MAUFFET), very different spellings (LEWIS as LOUIS), and problematic transcription (GARWOOD as YARWOOD) (Winchester, 1973).

Phonetic variations

Phonetic variations depend on the dialect or pronunciation conventions of the speaker. For example, the nickname Pooh, as it is spelled in English, would be spelled in German as Puh. Where the phonemes of the name are modified, e.g. through mishearing, the structure of the name may be substantially altered. MAXIME and MAXIMIEN are related names but their phonetic structure is very different. Indeed, phonetic variations in first names can be very large as illustrated by ADELINE and its shortened form LINE.

Other variations

Character Variation. The problem created by capitalization, punctuation, spacing, qualifiers and abbreviations (Branting, 2001) can be shown as follows:

- Capitalization, e.g. brown and Brown; SMITH and Smith
- Punctuation, e.g. WILL SMITH and WILL-SMITH; SMIT and S.M.I.T
- Spacing, e.g. YOUNGSMITH and YOUNG SMITH
- Qualifiers, e.g. WILL SMITH and WILL SMITH YOUNG
- Abbreviations, e.g. ROB and ROBBIN; BOB and BOBBY

Double Names. There are some cases where surnames and first names are composed of two elements but both are not always shown. For example, a double surname such as PHILIPS-MARTIN may be given in full or sometimes as single names, such as PHILIPS and MARTIN. Double first names although are not common in the English language, when considering German, for example, given names such as Klaus-Dieter can be transformed like Klaus and Dieter.

Alternate First Names. It can cause major problems for identifying people when a person changes the name during the course of life or is called by one of the first names during a period of life and another later on. An example: in German culture you can get more than one first name, but only one is the caller name (Rufname). This name must not be the first given name and it can even change in the course of life. A name like Jochen Peter Spohn consists of two given names and the family name. During a certain period of time the person may have adopted Jochen as Rufname, later may be changed to Peter. In this situation an algorithm that recognizes simple variations in spelling or phonetics would not be able to identify two such names as referring to the same person. Another example is the Christian tradition of wives taking her husband's surname such as "Mrs. Totty the former Miss Penket" or "Mrs. Totty nee Haggart".

Culture of Names / Naming Conventions and Customs in Different Cultures

The issue of name variations becomes more problematic when dealing with names from other cultures because the sorts of variation that are permitted may not be the same as those permitted in English. Names vary between cultures which for a long time has been an obstacle for creating a single method for automatic name processing.

For example, within each of the following cultures (Korean, Arabic, Hispanic, and Hungarian) all the names given are permitted variants of the same name except the last one (Dematteis et al., 1998). How to find out which of the examples below are given names and surnames?

> PARK DOE REE / PAG TO NI / TO NI PAG (Korean),
>
> MOHAMMAD ALI ABD EL NADIR NUR EL DIN / IMHEMED ABDUNADEER NOOREDDINE / MHMD NUR ABD AL NADER (Arabic),
>
> ENRIQUE CESAR VELEZ ARGUETA ENRIQUE BELES, QUIQUE VELEZ A. E. C. ARGUETA (Hispanic),
>
> Eoetvoes Lorant / Roland Eoetvoes / Eoetvoes Roland (Hungarian)

In each case, the final name would be considered an unacceptable variation of the name under consideration; it would be another name (Dematteis et al., 1998).

Another example, in French names you sometimes cannot differ between male and female variations relying only on phonetics. Consider the male first name MICHEL which pronounced the same way as the female MICHELLE, whereas German uses two different names both in spelling and pronunciation: MICHAEL and MICHAELA.

Naming Methodology for Given Names

The way of naming can vary, e.g. naming by monks, grandparents. Since antiquity names have been very important to people. Naming from the past to the present has been continuously developed and has evolved into a variety of patterns. Each pattern has its own rules depending on local belief and language that has been developed until the present. The basic goal of naming is to provide a good fortune and progress during life. Most first names have a meaning. Three methodologies are briefly described in the following.

- Principal naming using Thai astrology is widely used since antiquity as it involves the birth day in order to form the name. This is a belief that the individual has a set of 8 attributes called name of the angles referred to in Thai astrology. These attributes influence each person's livelihood, fortune, etc. The attributes refer to Servant > Age > Power > Honor > Property > Diligence > Patron > Misfortune. Each attribute has it own letters which can be used for constructing names.
- Principal naming using numeric methodology: Each letter has distinct numbers which can be added and have according values. These values represent low or high characteristics. The method can always be used along with naming both first names and surnames by increasing the value of first name and surname. Thus using numeric methodology can be used to increase "power" in names and check for better names.
- Principal naming which uses the traditional calendar is considered by Thai fortune tellers as the best method of anticipating the horoscope or destiny of people. This methodology takes day, month and year of birth including the time of birth to calculate the personality according to astrology. The results of this prediction are defined to tell the fate and personality thoroughly in the future.

In many cultures naming is not only important because every individual needs to have a name but to have helpful names or names with a good sound. Thai parents always try to choose names which they feel will bring good luck to their offsprings and to the family. The choice of appropriate names bases on the rules of available letters that can influence the destiny of the individuals as described in the following. Letters and days refer to Thai astrology. In that process the fortune depends on the day of birth and the related letters shown in Figure 2. Note that Wednesday occurs

twice. This is because counting of the Thai dates is different from the Western style. Thai people start a week on Sundays and a new day at 6 a.m. Thus naming using dates has to consider timing as well, especially on Wednesdays, which is the middle of the week and can be divided into daytime (from 06:00 to 17:59) and nighttime (18:00 to 05:59 the next day).

Figure 2: Naming rules using letters of the angles referred to in Thai astrology

The method for naming according to the available letters shown in Figure 2 is as follows. Starting with the weekday of birth we have a set of letters available which refer to 8 basic properties.

Servant > Age > Power > Honor > Property > Diligence > Patron > Misfortune

Related concepts with these attributes as they are used in Thai astrology are shown in Table 2.

Table 2: Personal attributes and related concepts

Servant	Children, Husbands, wives, including people who we support within family
Age	Life, livelihood, including the way of living
Power	Destiny, honor, fame, position, including education and love
Honor	Asset, money, appliance fortune which can be gained in the future
Property	Properties that inherited and still exist in the present including status of relatives
Diligence	Diligence, success from working, including creative and hard working
Patron	Supporters, such as parents , teachers, bosses and helpers
Misfortune	Evil, enemies, sins, including any obstacles

For example, in Figure 3 people who are born on a Sunday and need to add fortune in the attribute Servant need to use a set of vowels in naming process. For a boy the first letter must only be one of the letters in the attribute "Power" e.g. จ ฉ ช ฌ ญ. If they want a long life they need to choose one of those letters in the attribute "Age". The same applies for the attribute "Property" if the offsprings should have properties in their life. However the parents will not use letters from the set of letters in "Misfortune", in the example that is ศ ษ ส ห ฬ อ in order to avoid evil, enemies, sins including any obstacles.

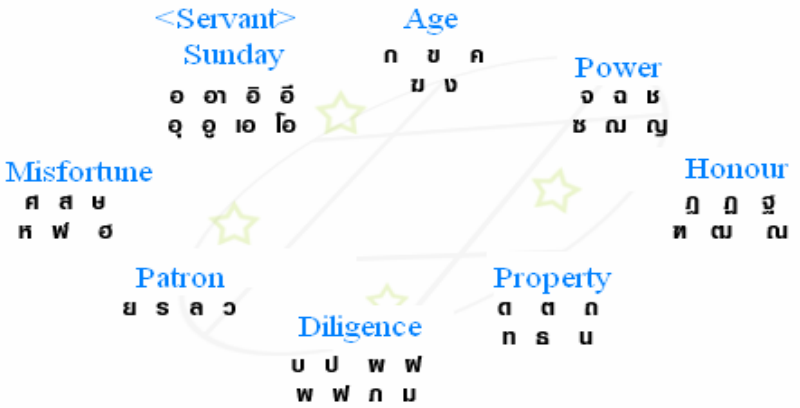

Figure 3: An example for people who were born on Sunday according letters according to Thai astrology

An example of naming using an attribute "Servant" from Thai astrology as is illustrated and described in Table 3.

Table 3: Possible letters in attribute "Servant" according the Birthday

Date of birth	Servant attribute
Sunday	All vowels
Monday	**ก ข ค ฆ ง**
Tuesday	**จ ฉ ช ช ฌ ญ**
Wednesday (daytime)	**ฏ ฏ ฐ ฑ ฒ ณ**
Thursday	**บ ป ผ ฝ พ ฟ ภ ม**
Wednesday (nighttime)	**ย ร ล ว**
Friday	**ศ ษ ส ห ฬ อ**
Saturday	**ด ต ถ ท ธ น**

In the past, the Thai culture widely used letters in "Power" interval (attribute) for boys' names and letters in "Honor" for girls' names. Nowadays there is no limit in choosing letters for genders but letters from "Misfortune" attribute are generally avoided in the naming process.

Order of Names

Name order is crucial; although the spelling variants of the name elements in the final name are acceptable, their order is not. In Eastern naming system, e.g. Korean, Japanese, and Chinese, the family name appears in the leftmost position and cannot move to the rightmost position. This applies to Hungarian personal names as well. In Western system, e.g. English, German, French, Spanish, Italian, and American, the family name appears in the rightmost position.

For example, in Hispanic names, the family name is the next to the last element (VELEZ); the rightmost name (ARGUETA) may be dropped, but not the family name. ARGUETA would therefore refer to another family, if it occurs alone (Dematteis et al., 1998).

As the function of personal names is not only to distinguish between individuals but also to serve as a help for indexing we have to mention the different conventions for ordering names. In Western order we change the position of given names and family names for indexing. In Eastern we do not have to change because the order for family names is already there. Thai ordering of names is distinct to these because Thai names are sorted according to the first names. There is no need to change the order of first name and family name as well.

Transcription of Names

English spelling with its many-to-many sound/letter correspondences contributes to the problem of Romanization of non Western names. Dialectal differences, historical and phonetic spellings make the English names somewhat unpredictable. The latter is even the case for English names. The following examples are given by (Reaney & Wilson, 1997).

> COLWELL, COLWILL, COLLWELL
>
> LEA, LEE, LEGH, LEIGH, LEY, LEYS, LAY, LAYE, LYE
>
> THOMPSON, THOMSON, TOMSEN, TOMSON
>
> WORCESTER, WORSTER, WOOSTER, WOSTEAR

Spelling variations are especially prominent in names from non-Roman writing cultures when such names have been transcribed to Roman characters, e.g. the Romanized form of an Arabic name: NOOR EL DIN, NURELDIN, NUREDDINE.

For the Thai writing system which is very complicated there exists an official standard called Royal Thai General System of Transcription (Wikipedia, 2005) which is used for rendering Thai names into the Roman alphabet. It uses only straight letters for vowels, diphthongs and aspirated consonants. It does not indicate the length of a vowel and the five different tones. From Chulalongkorn University, Bangkok, there is also an automated tool available (<http://www.arts.chula.ac.th/%7Eling/tts/>) which transcripts Thai names or terms into Roman letters. It is called “Thai Romanization.”

In many cultures, available standard transcription systems are not used or are used inconsistently. The range of variation found in distinct instances of the same name is therefore not fully predictable from such systems. In Arabic, for example, although there are transcription systems used by libraries and other official agencies, transcription tends to be far less predictable and highly inconsistent, even with a single individual. For example, an individual whose name is “ABD EL NADIR” may Romanize the name on one occasion as ABDUL NADEER and on another as ABDUNNADIR. Both name representations are “correct” and can be said to be accurate Romanization of the same Arabic name. Even in cultures in which transcription systems provide a reliable standard, personal interpretation, accommodation to the spelling of another culture or perceptual confusion can cause the spelling to deviate from the standard. Thus, for example, the Thai name GOFF will vary with KOFF, because G and K are Romanization alternatives from different tran-

scriptions systems. An observed variant of GOUGH, however, are GOFF and GOFFE, representing the influence spelling of English surnames by Reaney and Wilson (1997). The Thai Romanization tool gives KOP, on the other hand KOP is a correct transliteration of two different Thai names: กอฟ and กอล์ฟ.

Implementation of Hybrid Algorithms and System Concept

Many techniques have been used to cope with the important problem of matching variant names. However, most of these techniques were developed for general word matching and as a result they are not optimized for personal names matching. Spelling as well as phonetic variations combined with cultural aspects are the more challenging problems for automated multicultural name matching systems.

Most systems today use specific techniques for name matching rules and specific variations, e.g. Guth, 1976 and Levenshtein, 1965 algorithms are spelling/string analysis based, whereas Soundex (Winchester, 1970) and Phonex (Lait & Randell, 1998) are phonetic/sound based algorithms. However, most researchers have tried to implement a method which can deal with the culture of names and naming system which are meant to overcome the ethnic problems, e.g. NameX (<http://www.namethesaurus.com/Thesaurus/FAQ.htm>) and Varispell (<http://www.alphabic.com>).

Matching Algorithms for Names

The difficulty of the name matching task and the requirements for an effective algorithm to perform this task, both depend on the type and degree of name variations which occur. More recently published name matching techniques are either of the composite or hybrid form (Snae & Diaz, 2002) and several novel hybrid algorithms (e.g. LIG2, and LIG3) have been developed for specific purposes. All the name matching algorithms encountered in the literature and presented in this paper are based on alphabetic and/or phonetic similarity and/or name transformations (e.g. forename abbreviations) but may use a variety of distance and other metrics for representing the match. From an initial search of the literature, we distinguished four types of algorithms and implemented them using the C programming language:

1. spelling/string analysis based algorithms (e.g. Guth and Levenshtein),
2. phonetic/sound based algorithms (e.g. Soundex, Metaphone, NYSIIS, and Phonex),
3. composite methods (spelling or sound, e.g. SIMPLEX, and ISG),
4. hybrid approaches (spelling and sound, e.g. LIG algorithms).

A hybrid algorithm combines phonetic and spelling based approaches using similarity measure as probability is called LIG1, LIG2, and LIG3 (LIG algorithms). The LIG algorithms are a combination of three name matching methods: Levenshtein, Index of Similarity Group (called ISG), and Guth. The LIG algorithms have the best performance in term of producing most accurate true matches, overcoming name variations, and increasing the hit rate. They have proved to be more accurate than other methods in the literature (Snae & Diaz, 2002). The advantageous characteristics of these algorithms can be summarized as follows: (1) simple design, which can lead to accuracy improvements without decreasing the performance, (2) use of probabilistic similarity measures based on distance and weight, (3) increase correct positive and reduce negative matches to maximize the overall accuracy, (4) provide phonetic tuning to address multi-cultural names without depending on the language.

For these reasons we will use LIG algorithms for name variations and matching in the proposed naming systems, e.g. Thai naming system. For that we use a dictionary database of more than

8.000 Thai names which contains not only the spelling, but also the meaning and correct pronunciation. In situations where names follow the rules but do not have a meaning we compare the name with similar names in a dictionary database and check for similarity above a specific threshold. Then the user can select the best name from the resulting list of names.

Rule Based Naming System for Thai Names

In this section we will show the basic rules for Thai given names.

Thai Transcription

For consonants, the transcription is different depending on the location of the letter within the syllable. In the column showing the vowels in Figure 4 a dash ("–") indicates the relative position of the initial consonant belonging to the vowel.

Consonants

Letter	Initial	Final
ก	k	k
ข	kh	k
ฃ	kh	k
ค	kh	k
ฅ	kh	k
ฆ	kh	k
ง	ng	ng
จ	ch	t
ฉ	ch	t
ช	ch	t
ซ	s	t
ฌ	ch	t
ญ	y	n
ฎ	d	t
ฏ	t	t
ฐ	th	t
ฑ	d	t
ฒ	th	t
ณ	n	n
ด	d	t
ต	t	t
ถ	th	t
ท	th	t
ธ	th	t
น	n	n
บ	b	p
ป	p	p
ผ	ph	p
ฝ	f	p
พ	ph	p
ฟ	f	p
ภ	ph	p
ม	m	m

Consonants

Letter	Initial	Final
ย	y	-
ร	r	n
ฤ	rue, ri, roe	-
ฤๅ	rue	-
ล	l	n
ฦ	lue	
ฦๅ	lue	-
ว	w	-
ศ	s	t
ษ	s	t
ส	s	t
ห	h	-
ฬ	l	n
อ	h	-

Vowels

Letter	Romanisation
–ะ, –ั, รร (with final), –า	a
รร (without final)	an
–ำ	am
–ิ, –ี	i
–ึ, –ื	ue
–ุ, –ู	u
เ–ะ, เ–็, เ–	e
แ–ะ, แ–	ae
โ–ะ, –, โ–, เ–าะ, –อ	o
เ–อะ, เ–ิ, เ–อ	oe
เ–ียะ, เ–ีย	ia
เ–ือะ, เ–ือ	uea
–วะ, –ว, –ว–	ua
ใ–, ไ–, –ัย, ไ–ย, –าย	ai
เ–า, –าว	ao
–ุย	ui
โ–ย, –อย	oi
เ–ย	oei
เ–ือย	ueai
–วย	uai
–ิว	io
เ–็ว, เ–ว	eo
แ–็ว, แ–ว	aeo
เ–ียว	iao

Figure 4 Thai Transcription according to the Royal Thai General System, (Wikipedia, 2005)

Basic rules of forming syllables

Syllables are aggregated to names which sound good or aimed at good fortune according to the three methodologies mentioned above. As a consonant can not stand alone in Thai language and personal names we consider rules for vowels only. (See examples in Table 4.) The order is:

- Rule 1: Vowels can come first or can be followed by a first consonant, e.g. Ek
- Rule 2: Vowels can follow a first consonant without a final consonant, e.g. Ka
- Rule 3: Vowels that can not have final consonant, e.g. Tua
- Rule 4: Vowels that need final consonant, e.g. Kak

Table 4 Examples of Classification of Thai letter

Rules of vowels	Classification of vowels	Romanization	Examples
Rule 1	เ–, โ, แ–	E, o, ae	Ek
Rule 2	(รร), (–า), (ิ–, ี–), (ึ–, ื–) (ุ–, ู–),	(An), (a), (I), (ue), (u)	Ka
Rule 3	ย, –ำ, เ–ย, –ัว	A, am, e, ua	Tua, Tam
Rule 4	–ั, เ็–,	A, e	Kek

Example of construction of Thai syllables using Thai Romanization 1.10 Unicode (CU 2004) according to Figure 5: กา (Ka) = CV, เอก (Ek) = VC, กก (Kok) = CF, กัก (Kak) = CVF, เอก (Ek) = VF.

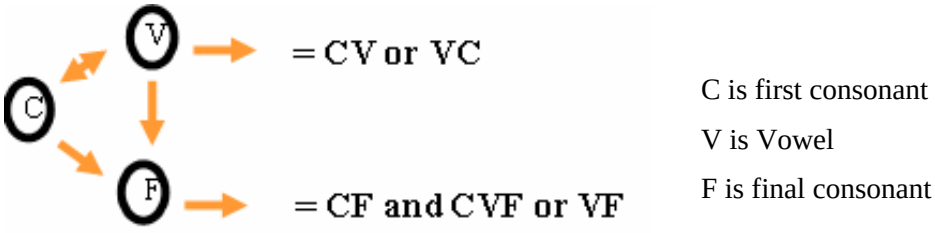

Figure 5 Forming of Thai syllables

Thai names are built from one or more syllables, which may or may not have a meaning. In the following it is shown how to construct Thai names with the help of ontologies that convey a meaning. Syllables are built from consonants (either C or F) and vowels. A name consists of one or more syllables. One syllable can have a meaning of its own, which leads in case of two or more syllables in a name to more complex meanings.

The process of constructing names according to the naming rules and methodology begins with a leading consonant or vowel that can be the only letter in the name. If we continue to add more letters we come either to a valid name (a name which has a meaning) or to an invalid name (a name without a meaning). Invalid names will be discarded in such a way that the last letter will be replaced by another or will be added with more letters.

In Figure 4 it is shown that a name comprises *n* syllables with a reasonable number of letters. The meanings of the syllables as well as of the name are found with the help of an ontology of names.

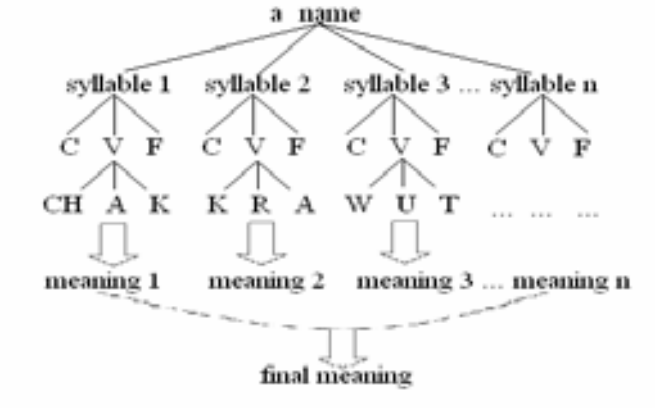

Figure 6: Representation and construction of Thai names

The meaning of the name chak-kra-wut in the example (see Figure 6) is “a man who carries a discus as a weapon”.

Currently we are constructing and implementing a web-based naming expert system (see Figure 7) which offers two basic ways to come to acceptable Thai names according to the first naming methodology mentioned above. The system will first display the letters for each date of birth (as we then know the day of the week). The system uses these letters to construct names based on the basic rules (see Figure 5). It will display the names in a user readable list, so that the user will be able to choose from it taking into account their respective meaning and sound.

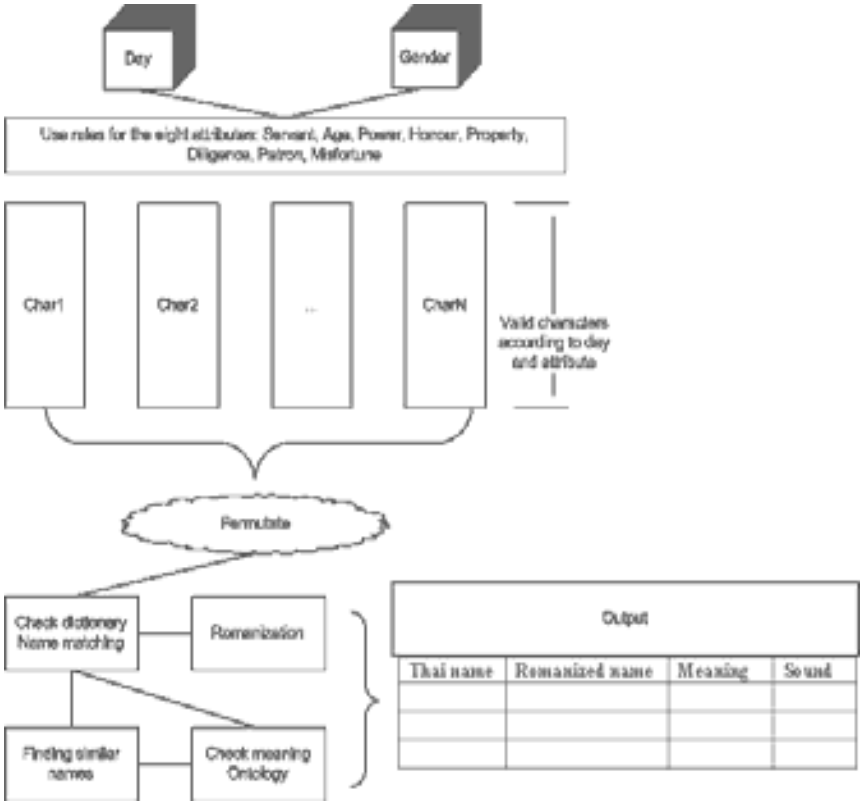

Figure 7. Architecture of Thai Naming Expert System

A second way to come to names is by using ontologies instead of basic spelling rules which are used according to the gender and date of birth. For this we check the different names against the date of birth by implementing an indexed database system of names from Thai dictionary for every day of a week.

Conclusion and Further Work

We have used Thai astrology as a naming methodology, an ontology of names, the LIG algorithms for personal name matching and the basic rules for forming syllables in Thai to construct the rule based naming system. Our proposed system will use name matching algorithms to return the variants of names from a dictionary with the relative probability of their similarity. The advantage of this process is to improve searching algorithms for multicultural names in databases as well as in the internet.

Currently we are developing a system called IT-TELLS (Interactive Transcription Tools for Explanation Level Language System) which will be a Romanization tool for names in Thai characters into Roman fonts.

A primary objective here would be to study how ontologies and algorithms can help in deciding which rules of naming system have to be implemented. This will also require an investigation into how ontologies which cover the different elements of names can be merged.

We want to extend our work to cover naming according to the birth day of the person to name. If we would know that a given name stems from the naming system using Thai astrology we can even derive the day of birth of a person with the help of ontology.

References

Andrade, H & Saltz, J. (1999). Towards a knowledge base management system (KBMS): An ontology-aware database management system (DBMS). *Proceedings of the 14th Brazilian Symposium on Databases*, Florianopolis, Brazil.

Andrade H. & Saltz, J. (2000). Query optimization in Kess – An ontology-based KBMS. *Proceedings of the 15th Brazilian Symposium on Databases* (SBBD'2000). João Pessoa, Brazil.

Blackburn, S. (1996). *The Oxford dictionary of philosophy.* Oxford: OUP.

Bouchard, G. & Pouyez, C. (1980). Name variations and computerised record linkage. *Historical Methods, 13*(2), 119-125.

Branting, L. K.. (2002). Name-matching algorithms for legal case-management systems. *The Journal of Information, Law and Technology (JILT).*

Dematteis, K., Lutz, R., & McCallum-Bayliss, H. (1998). Whose name is it: Names, ownership and databases. Originally written for: 1998 Annual Meeting American Name Society San Francisco, CA.

Gill, L. E. (1997). OX-LINK: The Oxford medical record linkage system, Complex linkage made easy, Record Linkage Techniques. *Proceedings of an International Workshop and Exposition*, 15-33.

Gruber, T. R. (1993). A translation approach to portable ontology specification. *Knowledge Acquisition, 5*(2), 199-220.

Guarino, N. (1998). Formal ontology and information systems. In N. Guarino (Ed.), *Proceedings FOIS'98*, Amsterdam.

Guth, G. J. A. (1976). Surname spellings and computerized record linkage. *Historical Methods Newsletter, 10*(1), 10-19.

Hamming, R. W. (1986). *Coding and information theory* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Jurafsky, D. & Martin, J. H. (2000). *Speech and language processing.* Prentice Hall.

Lait, A. J. & Randell, B. (1998). An assessment of name matching algorithm. *Society of Indexers Genealogical Group, Newsletter Contents, SIGGNL issues 17.*

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR. 163*, 845-848. [Translation: *Soviet Physics Doklady, 10*, 707-710].

Reaney, P. H. & Wilson, R. M. (1997). *A dictionary of English surnames.* Oxford: OUP.

Snae, C. & Diaz, B. M. (2002). An interface for mining genealogical nominal data using the concept of linkage and a hybrid name matching algorithm. *Journal of 3D-Forum Society*, *16*(1), 142-147.

Wikipedia. (2005). <http://en.wikipedia.org/wiki/Royal_Thai_General_System_of_Transcription> (Nov. 19, 2005)

Wild, W. G. (1968). The theory of modulus *N* check digit systems. *The Computer Bulletin*, *12*, 308-311.

Winchester, I. (1973). On referring to ordinary historical persons. In E. A. Wrigley (Ed.), *Identifying People in the Past* (pp. 17-40).

Winchester, I. (1970). The linkage of historical records by man and computer: Techniques and problems. *Journal of Interdisciplinary History*, *1*, 107-124.

Biographies

Chakkrit Snae. Ph.D. in Computer Science, University of Liverpool, England. M.Sc. in Computer Science, University of Newcastle Upon Tyne , England. B.Sc. in Mathematics, Naresuan University , Phitsanulok, Thailand. Since 2005 Lecturer at Department of Computer Science and Information Technology, Naresuan University, Thailand, mainly for Software Engineering and Artificial Intelligence. Research interests: Web Based Technologies, Hybrid Name Matching Algorithms, Data Mining, Intelligent System, and Expert System

Michael Brückner. Diploma in Physics from Technical University Munich, worked for many companies like Siemens and Deutsche Telephonwerke Berlin in SW Quality Assurance. Since 2004 Lecturer at Department of Computer Science and Information Technology, Thailand, mainly for Information Science. Research interests: Personal Information Management, Information Literacy, Thesaurus construction, Intelligent web search tools