Chemical compound names remain the primary method for conveying molecular structures between chemists and researchers. In research articles, patents, chemical catalogues, government legislation, and textbooks, the use of IUPAC and traditional compound names is universal, despite efforts to introduce more machine-friendly representations such as identifiers and line notations. Fortunately, advances in computing power now allow chemical names to be parsed and generated (read and written) with almost the same ease as conventional connection tables. A significant complication, however, is that although the vast majority of chemistry uses English nomenclature, a significant fraction is in other languages. This complicates the task of filing and analyzing chemical patents, purchasing from compound vendors, and text mining research articles or Web pages. We describe some issues with manipulating chemical names in various languages, including British, American, German, Japanese, Chinese, Spanish, Swedish, Polish, and Hungarian, and describe the current state-of-the-art in software tools to simplify the process.

INTRODUCTION

Chemical nomenclature forms a small but economically significant specialization of technical document translation. The requirement for pharmaceutical and biotechnology companies to file patents in multiple territories or for European customs legislation to be published in the native languages of member states generates demand for translations of chemical compound names. However, several technical aspects of chemical naming complicate the task for conventional human translators or machine-translation software.

The linguistic morphology of chemical names is often very different from the host language. For example, in English, the components of a chemical name such as “chlorobenzene” are not separated by spaces as are words in English sentences. Systems that assume that English text can be divided at word boundaries (spaces) and the resulting words looked up via a lexicon or indexed are often confused by chemical names. For example, English search engines such as Google and Yahoo! are unable to find “chlorobenzene” by searching for “benzene”. Interestingly, in other languages such as Chinese, Japanese, or Korean (CJK languages), this is less of a problem, where for example the Japanese “クロロベンゼン” (chlorobenzene) can usually be found by querying for “ベンゼン” (benzene).

Another complicating factor is that whitespace itself is significant in chemical nomenclature. The molecule “phenyl acetate” is different from “phenylacetate” (see Figure 1).

Capitalization is also sometimes significant in chemical naming. The chemical name “N-butyrsulfiminidoylactic acid” represents a different chemical structure to name “n-butyrsulfiminidoylactic acid” (see Figure 2).

* Corresponding author e-mail: roger@eyesopen.com.
Although computer software for interpreting chemical names has been an area of active research since the 1960s, almost all of this effort to date has concentrated solely on English as the source language.

The amount of chemical information available on the Internet can be estimated by searching with Google using different languages. The results of such a survey are shown in Table 1 which used “benzoic acid” and its translations as the search term.6

Although perhaps skewed by the word boundary issue described previously, these results show that a significant amount of information concerning chemistry is available on the Internet in languages other than English.

**HISTORICAL INFLUENCE**

An interesting aspect of chemical name translation is the influence of history on the conventions used by different languages. As knowledge of chemistry has developed and evolved in parallel with that of human languages, the similarities and differences between the chemical terms (words) used in different languages frequently reflect the geopolitics at the time that class of chemical compounds was first discovered or synthesized. To a first approximation, the history of chemical nomenclature can be divided into three important periods. These are chronologically the prehistoric and alchemic era, chemistry’s renaissance, and the modern IUPAC era.

**Prehistoric.** Since prehistoric times, all human languages have had a word for water. Although many human civilizations and cultures never developed an advanced understanding of the physical sciences, the need to express the most primitive of needs is universal. Indeed, unlike most other forms of chemistry, the word for “water” is probably better associated with the point at which language was acquired, rather when it was discovered. It is interesting to observe that both Chinese and Japanese use the same word (character) for water, “水”, predating their divergence 3000 years ago, as do English and Dutch (reflecting their shared Anglo-Saxon heritage). Similarly, the word for water is pretty much the same in the Latin-based languages: aqua (Latin), agua (Spanish), água (Portuguese), acqua (Italian), and apa (Romanian).

As civilizations developed metallurgy, terms for easy to refine metals such as iron, copper, and gold were added to languages. Taking a word such as “mercury”, the similarities between the German “quecksilber” and the Swedish “kvicksilver” are apparent, as are the similarities between the English and the French “mercurie” and the Spanish “mercurio”. The Japanese for the metal mercury is “水銀” quite literally water (“水”) silver (“銀”).

By the Middle Ages, chemistry had become the domain of the alchemists. The compounds and elements found were named by their properties. The element oxygen is recognized as being associated with acidity, both “sauerstoff” in German (“säure” meaning acid, related to the English word “sour”) and “酸素” in Japanese (literally acid “酸” element “素”). The element hydrogen is associated with water, “wasserstoff” in German, “waterstof” in Dutch, “водород” in Russian (water is “водя” Russian) and “水素” in Japanese (literally water “水” element “素”). This relationship is even preserved in some modern synthetic languages, such as Klingon where water is “חוכב” (bIQ) and hydrogen is “חכם” (bIQ-SIP).

It should be noted that the introduction of a specialized chemical nomenclature by alchemists was originally intended as a form of obfuscation. It was deliberately intended to be incomprehensible to the general public and to keep results secret from competing alchemists.

**Chemistry’s Renaissance.** The birth of modern chemistry can be attributed to the pioneering work of Antoine-Laurent de Lavoisier, the father of modern chemistry, and his colleagues in the 1780s and 1790s.8 Lavoisier was the first to publish a list of elements. He, Guyton de Morveau, Bertholet, and de Fourcroy also published the first list of naming recommendations. They introduced the terms “succinic acid” and “malic acid”. This era standardized what are today considered traditional names. Competing chemists in different countries named compounds differently but shared some underlying principles. The system of naming “formic acid”, “acetic acid”, “propionic acid”, “butyric acid”, etc. was formalized.

For example, in English “formic” is derived from “formica” the Latin word for “ant”. The French, Spanish, and Romanian forms, “acide formique”, “ácido fórmico”, and “acid formic”, respectively, all follow the same Latin root. In German it is called “ameisensäure” (the German for ant is “ameisen”), in Polish it is “kwas mrókwkiowy” (from the Polish word for ant “mrówka”), in Hungarian it is “hangyasav” (from the Hungarian for ant “hangyák”), in Swedish it is “myrsyr” (from the Swedish word for ant “myra”), and so on. The English name “butyric acid” is derived from “butter”, leading to “butter säure” in German and “smörsyr” in Swedish. And the English word “lactic acid” is derived from “milk” leading to “milchsäure” in German and “mjölksyr” in Swedish.

During this same period, the Swedish Chemist Jöns Jacob Berzelius in 1813 proposed that chemical symbols be based on the Latin names of the elements, such as the symbol “Fe” for iron. This convention was generally adopted by the mid-19th century and is still in universal use today.

**The Modern IUPAC Era.** In the last 75 years or so, huge strides have been made in the field of systematic nomenclature standardization. This built upon the work of Hoffman in 1865 to arrange hydrocarbons into series by their formula, which introduced the series “methane”, “ethane”, “propane”, and “butane”. The International Commission on Chemical Nomenclature was first organized in 1889, and by 1930 IUPAC published its 68 “Liege rules”. Today known as the International Union of Pure and Applied Chemists (IUPAC), this standards body has published revisions and refinements to organic chemical nomenclature in 1957, 1965, 1971, 1979,9

---

**Table 1. Number of Web Pages Found by Searching Google for the Query “benzoic acid” in Different Languages**

| Language | Query      | Page Hits  |
|---------|------------|------------|
| English | benzoic acid | 1,680,000  |
| Chinese | 苯甲酸     | 1,040,000  |
| Japanese| 安息香酸     | 524,000    |
| German  | benzesäure | 331,000    |
| Spanish | ácido benzoico | 329,000 |
| French  | acide benzoique | 203,000 |
| Russian | besincev kiselota | 56,600  |
| Dutch   | benzoëzuur  | 9,940      |
and 1993. A new "blue book" is expected to be ratified in the near future. Chemical Abstracts Service (CAS) has also made significant contributions to the effort of systematizing chemical nomenclature.

Modern chemical nomenclature, as standardized by IUPAC, provides the underlying foundations for naming chemical compounds around the world. The original standards are published in English, and then each national body (National Adhering Organization in IUPAC terminology) is responsible for publishing the localized versions of the standards. One result of this approach is that English chemical nomenclature tends to be the most developed, with different languages advancing by the publication date of the most recent translation. For example, some terms introduced in the most recent IUPAC 1993 standard have not yet been officially translated into some languages. Examples of translated standards documents or chemical nomenclature texts in French, German, Hungarian, Polish, Russian, Spanish, and Swedish are given in the bibliography. A more complete list of official national translations of IUPAC's "blue book" is maintained on the IUPAC Web site.

Despite significant differences in their written forms, one remarkable property is that systematic IUPAC names are pronounced similarly around much of the world. IUPAC's efforts at international standardization are particularly phonetic, such that a speaker presenting at a conference in one language may be understood fairly well by an attendee in a second language. Much like British and Americans may disagree on the use of "aluminium" and "aluminum", respectively, these differences are frequently perceived as an accent or dialect.

Figure 3. IUPAC names for the same compound in different languages.

Table 2. Differences in IUPAC Naming between English Dialects

| British  | American | International |
|----------|----------|---------------|
| sulphur  | sulfur   | sulfur        |
| aluminium| aluminum | aluminium     |
| caesium  | cesium   | cesium        |

Languages

Although an article such as this cannot aspire to teach the reader the details of how to translate chemical nomenclature between languages, the following sections provide an overview of some of the issues with chemical names in English, Japanese, and Chinese.

English.

With 380 million people using English as their first language, English is the third most spoken language in the world, ranked by the number of native speakers. Despite being in third place, English has become the "de facto" standard in the scientific community, with current evolution in chemical nomenclature being dictated by international standards written in English and then independently translated into other languages. Although, in the following sections, English is used as a baseline for comparison, the differences between the British and American forms of chemical names is informative.

There are only three significant differences in systematic chemical naming between American and traditional British names. The element "sulfur" in American English has traditionally been spelled as "sulphur" in British English (more precisely in the United Kingdom and the Commonwealth of Nations), and the elements "aluminum" and "cesium" are spelled "aluminium" in American English and "caesium" in British English.

IUPAC recommendations are expressed in International English, which is a hybrid of British and American spellings, as a compromise to both sides of the Atlantic. Officially, "sulfur" is now used for element #16 (following the American spelling), while "aluminium" and "cesium" officially have their British spellings (see Table 2).

The "sulfur" vs "sulphur" issue affects not only the element itself but also many uses of "sulf" as a stem in IUPAC names. This gives rise to names such as "benzene-
Table 3. Some Japanese Katakana Characters and Their Romanization

| ア | イ | ウ | エ | オ |
|---|---|---|---|---|
| カ | キ | ク | ケ | コ |
| サ | シ | ス | セ | ソ |
| ダ | デ | ド | ノ |
| ハ | ヒ | フ | ヘ | ホ |
| マ | ミ | ム | メ | モ |
| ラ | リ | ル | レ | ロ |
| ハ | ヒ | フ | ヘ | ホ |

The influence of IUPAC’s 1990 decision to adopt “sulfur” as the preferred spelling resulted in the Royal Society of Chemistry deciding to officially change to “sulfur” in 1992 and the Qualifications and Curriculum Authority in England and Wales to recommend its use in 2000. Given that the spelling “sulphur” is no longer being taught in schools, the distinction between British English chemical nomenclature and International English chemical nomenclature is likely to affect only legacy documents in coming years.

The history of the American spelling “aluminum” is also interesting. The root of the confusion dates back to its original isolation and discovery by Humphry Davy in 1808, who first called it “aluminum” but by 1812 had decided upon the name “aluminum” in his article in the Journal “Chemical Philosophy” reporting the discovery. However, it was quickly pointed out that there was a developing precedent for using the suffix “-ium” for naming new metallic elements, and later articles referred to it as “aluminum”. This spelling was adopted universally, with even the 1828 edition of Webster’s American dictionary spelling it “aluminium”. All this changed in 1892, when American businessman Charles Martin Hall developed a new method for producing the metal by electrolysis and marketing it under the (trade)name “aluminum”. While it has been suggested that this may have been a spelling mistake, Hall’s domination of the metal’s market led to it commonly being spelled as “aluminum” in North America and ultimately the American Chemical Society adopting it as the official spelling in 1926. Today, the Canadian Oxford Dictionary follows the American spelling, while the Australian Macquarie Dictionary follows the British.

Japanese. Japanese is the ninth ranked language in the world, with 130 million native speakers.

The Japanese writing system consists of three main scripts or types of characters, kanji (logographic characters of Chinese origin) and hiragana and katakana (both syllabaries whose characters reflected how parts of a word should be pronounced). Kanji characters were borrowed from Chinese over 1000 years ago and are used to denote the most common concepts and words in Japanese. Hiragana is used to write words more recently added to Japanese for which there are no kanji characters or occasionally in place of rare or difficult kanji characters. Finally, katakana is used to write foreign words and names loaned from other languages and as a consequence for technical and scientific words.

Each katakana character denotes a syllable, much like the characters of the English alphabet denote phonemes. An abbreviated table of katakana along with their romanization is shown in Table 3. This table may be used to romanize many of the Japanese examples in this article.

As an example of katakana transliteration, the Japanese word for “methane” is “メタン” which is composed of three symbols, the first denoting “me”, the second “ta”, and the final one “n”, giving literally “me-ta-n”. A slightly more complicated example is the word “methyl” which in Japanese is “メチル”, which literally would be “me-ti-ru”. In the transliteration process the English letter/phoneme “y” is treated like “i”, the letter/phoneme “l” is treated like “r” (the origin of jokes about “flied lice”), and finally the u-form is used for a terminal consonant. This demonstrates that a major task of translating from Japanese to English is disambiguating these characters, by making use of their context.

In the periodic table, the prehistoric metals known since antiquity, such as gold (Au), silver (Ag), lead (Pb), iron (Fe), copper (Cu), and tin (Sn), all have kanji characters, while more recent elements such as sodium (Na) and uranium (U) are expressed in katakana, which is consistent with their time of addition to the language.

Also of note is that the Japanese word for sodium is “ナトリウム”, which is a transliteration of “natrium” (literally “na-to-ri-u-mu”), showing that the word was acquired from Latin, perhaps via Italian or German, rather than from English.

There are also a small number of exceptions for relatively obscure compounds that use kanji characters, often natural products. Examples include chrysanthemic acid and mucic acid (see Figure 5).

Chinese. The Chinese (or Sinitic) languages are the most spoken language family in the world, with about 1 billion native speakers. Although Chinese consists of several spoken languages, including Mandarin, Wu, and Cantonese, they share the same written form and may be considered a single language for machine-translation of chemical names. However, since the 1950s, this single written language assumption is no longer entirely accurate. In an attempt to make Chinese characters easier to learn and faster to write, the People’s Republic of China attempted to reform the complex “Traditional Chinese” characters, making many obsolete and replacing others with simpler forms creating “Simplified Chinese”. Hence written Chinese consists of two forms, “Simplified Chinese” which is used in Mainland China, Singapore, and Malaysia and “Traditional Chinese” which is still used in Hong Kong, Macau, and Taiwan.

All chemical elements in Chinese are represented by their own character. Indeed this requirement combined with ongoing discoveries of new heavy elements means that symbols for new elements are among the newest symbols to enter Chinese dictionaries. For examples, elements above atomic number 104 were only added as traditional Chinese characters to the Unicode standard with version 3.1 (2001).
and therefore fail to display in some WWW browsers, while simplified Chinese characters for these elements (such as seaborgium) have yet to be added to the Unicode standard (as of v4.1 2005).

Chinese chemical names are perhaps the single exception to the rule that most natural languages use phonetic transliterations of English systematic names. Instead Chinese retains the same structure/ordering as other languages but uses its own symbols and rules for encoding the chemistry. For example, the sequence “methane”, “ethane”, “propane”, “butane”, and “pentane” is translated as "甲烷", "乙烷", "丙烷", "丁烷", and "戊烷", effectively the sequence “first-alkane”, “second-alkane”, “third-alkane” where the symbol “烷” indicates an alkane. Likewise the names “methyl”, “ethyl”, and “propyl” use the same idiom, becoming “甲基”, “乙基”, and “丙基” where the character “基” denotes alkyl. Even the carboxylic acids, “formic acid”, “acetic acid”, and “propionic acid” are translated as “甲酸”, “乙酸”, and “丙酸”. First, notice that the symbol for acid “酸” is the same symbol as the Japanese Kanji used for the word “acid”. Second, notice that this clever systematization is equivalent to the systematic (but not preferred) names “methanoic acid” and “ethanoic acid”. As with Japanese, because the names “acetic acid” and “ethanoic acid” are not distinguished in Chinese, translation from Chinese to English may need to disambiguate a preferred form, even though there is no chemical ambiguity in the name itself. A more significant complication involves the translation of esters between English and Chinese. For English, the currently preferred form is to write compound names such as “methyl acetate”. Chinese, on the other hand, does not appear to support this idiom, instead expressing this name as “乙酸甲酯” which literally translated would be (the equivalent) “acetic acid methyl ester”.

METHODS

Implementation. As explained previously, the common “structure” of chemical names allows a software implementation to translate between languages using simple substitution. Normally, in most natural language translation software it is necessary to perform sentence parsing and deep analysis in order to identify nouns, verbs, and adjectives and from there identify the subject, object, and tense. This allows resolution of ambiguities, such as which of the translations in a dictionary is required. But in the case of chemical names, there is very little ambiguity in word/token usage and almost no change in word/token ordering, allowing names to be translated by string replacement. This is the technique used to perform language translation in OpenEye Scientific Software’s Lexichem “structure-to-name” and “name-to-structure” products,23 described in detail below. A flowchart of this process is given in Figure 6.

The lexical string replacement is performed at the whole name level, identifying tokens or lexemes in an input string, translating them, and composing the results in an output string. This process is completely independent of machinery used to parse or generate names from English words. For example, the language translation functionality is able to translate some names that cannot be parsed or would not be generated by Lexichem or similar software. Indeed, the text-to-string translation process allows Lexichem’s translation software to be used in conjunction with other “name-to-structure” or “structure-to-name” software such as ACD Labs’ ACD/Name,23 CambridgeSoft’s Name=Struct,24 and MDL/Beilstein’s AutoNom25,26 (or others27) or to be used purely for the purposes of translation. One useful benefit of combining translation to English with conventional name-to-structure software is that the correctness of the translation can be automatically assessed by the ability of the parser to recognize the translation as a valid chemical connection table.

An alternative approach might have been to follow the usual method of “internationalizing” software, by providing alternate translations for each token read by a parser or written during name generation. The drawback with this approach is that a significant number of subtleties when translating between languages do not occur at the token boundaries found in English. For example, in Welsh, the multiplier “di” is usually translated as “deu” and the prefix “chloro” is translated as “cloro”. The problem is that these translations are context sensitive such that “dichloro” is actually translated as “deugloro”, an interaction not seen in English. Another example is from Japanese where katakana characters need not end between English tokens; words such as “ethanol” are represented as “エタノール” and the prefix “di” is usually translated as “デウロール”, straddles two tokens. The first significant difficulty encountered when translating between languages is the issue of character sets. Historically, while English documents have traditionally been stored in ASCII, the requirements and character sets of other languages have meant that they frequently use their own encodings. West European languages that use accented Latin characters use the ISO 8859-1 character set (also known as “Latin-1”) that encodes additional characters in a single byte by using the values between 128 and 255. In Russia, the typical encoding of Cyrillic characters is KOI-8. In Hong Kong, the usual encoding of traditional Chinese is called “Big-5”. And in Japan, many documents are stored in either

![Figure 6](image-url)
“Shift-JIS” or “EUC-JP”, depending upon the operating system being used. More recently, the UTF-8 encoding of Unicode characters has become more common, allowing the same representation of characters to be shared between languages and even allowing multiple languages to be used in a single document. To simplify having to deal with these numerous representations, Lexichem’s translation machinery internally uses a standard Unicode encoding, converting from (or to) the appropriate external character encodings on either input or output. In addition to the more common character encodings, a similar conversion allows support for HTML encoded characters, allowing strings such as “&amp;#589;&amp;#x606f;&amp;#x9999;&amp;#x9178”; (Japanese, in hexadecimal) and “&amp;#33519;&amp;#30002;&amp;#37240”; (Chinese, in decimal) to be handled.

To simplify software development and string pattern matching, the non-ASCII Unicode characters are represented internally as 7-bit clean ASCII characters using Unicode escape sequences similar to those used by the Java programming language.28 In this scheme, extended characters are encoded by the sequence “uXXXX” where XXXX is a four digit (lowercase) hexadecimal value that specifies the 16-bit Unicode character. This allows the code to work internally a single byte at a time on a single canonical character representation. Hence the Japanese symbol for “gold” is encoded as the string “u91d1” in both the strings being modified and the files used to specify substitution rules. By restricting the source code and rule files to simple 7-bit clean ASCII, we also avoid potential problems editing files or modifying/compiling source code on machines without suitable fonts or internationalization support. By using ‘\’ escapes, the need to use strange keyboards to enter symbols can be avoided.

Another technicality is the potential problem of mixed case. To minimize the number of substitution rules required for language translation, each compound name is converted to lower case prior to pattern matching. This allows the Spanish word “AGUA” to be treated and recognized identically to “agua”. Fortunately, Japanese and Chinese do not have a notion of uppercase or lowercase to represent capitalization, but alas the Russian Cyrillic alphabet and Greek alphabets do, as do many of the European accented characters in Latin-1. This requires the appropriate algorithms to perform the transliteration to English lowercase, with the appropriate chemistry-aware checks of which characters are case-sensitive in IUPAC names.

A curious corner case of note is the problem of dealing with Russian compound names such as “1Н-Пиррол” used to denote “1H-pyrrole”. The subtlety is that instead of inserting a Latin “H” for the indicated hydrogen, it is often easier when using a Russian keyboard to instead use the Cyrillic capital “ч” (Unicode character “u041d”), which looks indistinguishable from “H” in most fonts. Hence, the chemistry-aware case conversion needs to be able to transliterate “\u041d” with “1H” - and not “1u043d” - or the equivalent transliterated Latin “1n-”.

For generating names, the equivalent inverse function exists to capitalizing the chemical name correctly by determining the appropriate character to modify.

The core knowledge of the translation process is encoded by a rule file, containing a number of rules, that each specify the pattern string to match in the input string, and the replacement text to use in the output string. At run-time the translation algorithm proceeds left-to-right over the input string, identifying the longest matching pattern at the current position. If a suitable pattern is found, the replacement text is appended to the output string, and the input is advanced by the number of characters in the matching pattern. If no such pattern/rule is found, the current character from the input is appended to the output string, and the input advances a single character.

As a concrete example the entire rule file for translating American compound names to International English (for example for filing a Worldwide Patent application) is given by the two lines in Figure 7. In the Lexichem implementation, the pattern and replacement are separated by one or more TAB characters, allowing spaces to be used in both the pattern and replacement. As a convenience to rule file authors, blank lines are allowed, and lines beginning with ‘#’ are treated as comments.

A single rule from the Greek to English rule set (that shows escaped Unicode) is given in Figure 8.

Although it is theoretically possible to use a single set of rules (file) for converting from English to language X and from language X to English, in practice it has been found easier to treat translation directions separately and encode the rules independently. This asymmetry allows the Chinese rules to handle both Simplified and Traditional Chinese when translating to English (as a single rule set), but only Simplified Chinese is currently supported from English.

As text-mining of large data sets and interactive translate-as-you-type are significant target applications, translation performance is a potential issue. To improve the rate at which text can be processed, the rule files are treated like source code and are “compiled” to generate the efficient C++ source code that is used to perform the pattern matching. By performing the longest prefix lookup using tries, implemented by C++ switch statements, the time to determine the longest prefix can be made independent of the number of rules (and their complexity) in the rule set. This is particularly important as some languages require rule files with several thousands of patterns, and a naïve implementation might loop through each at each character. The result of compiling the rules into C++ is an extremely efficient translator with complexity linear in the size of the input. Running on a single 2 GHz AMD Opteron processor, Lexichem is able to translate 14Mbytes of compound names (250,251 names) from English to German in under 2 s. Although it is possible to make the translation process even faster using finite state machines, the current level of speed was considered more than acceptable.

In order to support the necessary complexities of language translation, the pattern and replacement text are extended beyond simple textual patterns, by adding “meta” characters similar to regular expressions. The character “$” at the end of an input pattern checks that the pattern text appears as the end of the input string. An example use is in the German

Figure 7. An example of a Lexichem translation rule file.

\u03bd\u03b5\u03c1\u03cc  water

Figure 8. A more complex translation rule demonstrating escaped Unicode.
to English rules where “chlor” appearing at the end of a name is assumed to mean the element “chlorine”, but in the middle of the name is assumed to mean the prefix “chloro”, so that “chlorobenzol” is translated as “chlorobenzene”. A minor subtlety is that Lexichem by convention allows multiple disconnected components to appear in the same name, separated by semicolons, such as “benzene; acetic acid”. As a result in addition to matching the end of string (the NUL character), the “$” meta-character also matches a semicolon.

To handle word reordering, if the first character of the replacement pattern is a “$”, the replacement text is considered to form a prefix of the resulting string, and if the first character is a “_,” the replacement text is considered to form part of the prefix. For example, to handle the frequent occurrence of the word “acid” being moved to the start of sentence, the English to Polish rules contain the pattern “ acid$” with the replacement “kwas “. And likewise, the Polish to English rules contain the pattern “kwas ” with the replacement “_ acid”. Notice that these examples also include appropriate spaces in both the pattern text and the replacement text.

A particularly chemistry specific meta-character that can often greatly simplify rule sets when translating to English is the substring “e?” which can be interpreted specially in the replacement text. Many languages, including German, Japanese, Russian, and Swedish, do not retain the “e” at the end of alkanes (and alcanes and alkylen) or ring names. So in German, the alkanes are “methan”, “ethan”, “propan”, etc., and the rings “pyrrole”, “pyridine”, and “pyrimidine” are written as “pyrrol”, “pyridine”, and “pyrimidin”. One great convenience this provides to the language is that it completely avoids IUPAC’s complex vowel elision rules. To a first approximation, the vowel elision rules allow the final “e” of an alkane or ring system name to be omitted when the first alphabetic character following it is another vowel. Hence in English, we have “methanol” and “methanamine” but “methanethiol” and “methanesulfonic acid”. Stripping out this “e” in the translation to German is trivial, but inserting it correctly when translating to English is far more complex, especially given the exceptions to this simplified rule. In names like “pyridin-2(1H)-one”, the indicated hydrogen “H” and possibly alphabetic characters on the ring locants are not considered. Another exception/variant is with compound names like “propanoectol” where the multiplier “octa” is an exception (a problem compounded by its “a” being elided). Rather than recode these rules repeatedly, it is far more convenient to write the German to English rule as “pyridin” becomes “pyridine” and allow a shared postprocessing pass, that implements the IUPAC rule in all its complexity, to correctly handle cases like “pyridine?-2-acetic acid”.

This relatively simple “pattern-replacement” rule syntax has shown itself to be adequate for translating a significant fraction of organic chemistry to and from various languages. To give some figures to the number of rules typically required, Table 4 shows the current number of substitution rules required/implemented in Lexichem’s language translation functionality. Separate numbers are given for the count of English-To-X rules, and for X-To-English rules. To provide a guideline for how fully implemented support for each language is (explained further below), the table below also includes an approximate classification of “quality” into five categories: “A” best through “E” worst. “A” may be considered production quality, while “E” should be considered investigative or experimental. A second indication of the rule set quality for several of these languages is given by the round-trip benchmarks given in the “Results” section.

Typically, though not always, translating a language to English requires more rules than translating to it from English. For those languages that use accented Latin characters, the Lexichem To-English rules often contain duplicates to allow both the accented and unaccented forms to be recognized when there is no ambiguity. For example, the French “acid benzoique” and “acid benzoïque” are both understood to denote “benzoic acid”. In some languages a translation may not be unique, in which case the English-To rules contain the single preferred translation, but the To-English rules may recognize multiple forms. Such an example is the support for both simplified and traditional Chinese characters mentioned previously. It is also frequently the case that lexemes used in English are perhaps better disambiguated than in other languages; for example when translating to German it is relatively straightforward to specify that both “ine” and “yne” should be replaced with “in”, but the return rules to determine which occurrences of “in” need to become “yne?” should be replaced with “yne?”, and which, such as in “indol”, need to be left alone. Additionally, the “closer” a language is to English the fewer rules it will require, as many words/tokens may not need to be modified between western European languages, but all tokens require explicit processing for Asian languages.

Another major factor is how comprehensive the translation support for a language is. Clearly, if a set of rules only covers the periodic table of elements (approximately “E” quality), it will be significantly smaller than a rule set that can fully handle organic, inorganic, organometallic, natural product, and “traditional” nomenclature as well as common drug names (approximately “A” quality). As mentioned elsewhere, Lexichem’s functionality is not restricted to just handling systematic IUPAC names. To be useful in practice, chemical machine translation software has to be able to handle common terms such as “water” or “caffeine”, even when IUPAC recommends using the terms “oxidane” and “1,3,7-trimethylpurine-2,6-dione”, respectively, instead. Naturally, when allowed by the source and destination languages,

| language | quality | English-To rules | To-English rules |
|----------|---------|------------------|-----------------|
| German   | A       | 292              | 831             |
| Japanese | A       | 742              | 1481            |
| Swedish  | A       | 190              | 403             |
| Spanish  | A       | 336              | 585             |
| Hungarian| B       | 390              | 756             |
| Polish   | B       | 506              | 562             |
| Chinese  | B       | 659              | 776             |
| Italian  | C       | 236              | 184             |
| Danish   | D       | 56               | 115             |
| Dutch    | D       | 83               | 90              |
| French   | D       | 91               | 77              |
| Romanian | E       | 114              | 132             |
| Russian  | E       | 232              | 201             |
| Slovak   | E       | 132              | 329             |
| Irish    | E       | 178              | 366             |
| Welsh    | E       | 170              | 186             |
common names in one language should be translated as common names in the other. Likewise, systematic names and other distinctions should be preserved. While Lexichem’s other functionality (not discussed here) allows the interconversion of different name styles, the role of language translation is to express the original name in the target language as faithfully as possible.

The current approach used by Lexichem is to provide translation rule files for each language to and from English, with the expectation that translation between two foreign languages will go via English as an intermediate. However, there is no reason why additional rule files, such as from German to Japanese, could not also be used.

**Examples**

To give some more concrete and nontrivial examples of compound name translation, translations of the scientific names of the three best selling drugs worldwide, Lipitor, Plavix, and Nexium, are shown in Figures 9–11. These are the names generated from the given structures by Lexichem, demonstrating the English-To rules applied to software generated names. Of note is that the (original English) names were generated with “typical usage” settings rather than any of the “strict standard adherence” settings, explaining the appearance of “2-pyridyl” instead of “pyridin-2-yl”. This demonstrates the need and ability to translate tokens such as “pyridyl” in addition to “pyridinyl”. These names have
been checked and confirmed correct/acceptable by native speakers of each of the languages shown.

BENCHMARKS

Round-Trip Benchmarks. One way of evaluating the quality of machine-translation software is by round-trip testing. A set of names is translated from a source language to the destination language and then retranslated back to the source language, and the results are compared to the original source language.

For Table 5, we used 250,251 machine-generated (English) IUPAC compound names from the National Cancer Institute’s screening database. A strict “string equality” test was used to determine whether the round-trip result was identical to the original. This standard is perhaps harsher than required in practice, as it is possible for the result to uniquely and unambiguously describe the original chemical structure but be named slightly differently due to native language preferences. The languages evaluated in Table 5 are the “A” and “B” quality languages given previously in Table 4.

Of course, one aspect that is not covered by round-trip benchmarking is whether the translation is valid in the foreign language. These tests only assess the degree of consistency between the (independent) “to” and “from” rules, over the range of chemistry (and naming) used in the test set. The correctness-validity of the current rule sets has additionally been checked by manual inspection of translated names by native speakers/chemists. Unfortunately, such evaluations are subjective and are only possible for samples of perhaps a few hundred names. Round-trip testing has the advantage of being automatable across data sets of many millions of compound names. These numbers also do not reflect the best possible values that are achievable but simply report the current performance given the effort the author has put into each language. Countries and languages that are existing Lexichem customers have had more time invested than those of more academic interest.

One conclusion that can be drawn from the round-trip fractions in Table 5 is that the quality of chemical translation is likely to exceed the ability of computer software to correctly interpret the chemical names. For comparison, Lexichem’s name-to-structure conversion rate is currently only 92.49% for the same benchmark set of names (i.e., it can convert 231,452 of the 250,251 names into connection tables). Hence in text-mining applications, such as extracting compounds from foreign patents, the failures due to mistranslation are likely to be rare compared to the failures due to poor name quality or complex chemistry.

An open question is what level of accuracy is desired or required of machine-translation software. For text-mining from foreign patents, any success rate is acceptable if previously the contained information could not to be extracted or indexed. However for authoring patents and legislative documents a very high level of accuracy is required of finished translation. However in such critical applications the role of machine-translation is often “machine assisted translation” where a native speaker proofreads or checks the result. In such usage, machine-translation is a significant productivity tool as it can significantly reduce the time taken to perform and type a manual translation from scratch. In practice, a Lexichem “failure” often caused by a missing translation rule typically results in the original term being retained untranslated (or just transliterated) in the output string, which while undesirable can still often be recognized and understood by a chemist.

An observation that can be made from analyzing some of the remaining failures is that not all of the differences are attributable to issues with the software or methodology. A number of mismatches are legitimately caused by ambiguities and inabilities to name compounds uniquely (with IUPAC
rules) in the target language. Such an example was given in the Introduction, where “phenylacetate” and “phenyl acetate” are different compounds but indistinguishable in Japanese and Chinese that do not retain a notion of whitespace. In these cases, alternate name styles such as translating the equivalent name “acetic acid phenyl ester” may potentially be used to resolve the ambiguity.

Another example failure with translations to and from Japanese is caused by the many-to-one mapping of katakana. Under Japanese chemical society rules the pair of compounds shown in Figure 12 has the same form in Japanese. Here the standardized transliteration rules result in the same name, and “d” and “z” are mapped to the same katakana character. In this case, one possible workaround might be to use alternate katakana variants to resolve the ambiguity. In Lexichem’s rule sets, the expected frequency of occurrence in pharmaceutical databases is used to choose between ambiguous translations. Additionally, contextual information such as an indicated hydrogen (which implies a conjugated ring system) can often be used to correctly disambiguate names like “enz”.  

Another related ambiguity in Japanese is shown by the pair of compounds in Figure 13. Unlike the previous example, there is no way to disambiguate “r” and “l” in Japanese katakana character. In this case, one possible workaround might be to use alternate katakana variants to resolve the ambiguity. In Lexichem’s rule sets, the expected frequency of occurrence in pharmaceutical databases is used to choose between ambiguous translations. Additionally, contextual information such as an indicated hydrogen (which implies a conjugated ring system) can often be used to correctly disambiguate names like “enz”.

Another related ambiguity in Japanese is shown by the pair of compounds in Figure 13. Unlike the previous example, there is no way to disambiguate “r” and “l” in Japanese katakana character.

Finally Figure 14 presents yet another Japanese ambiguity issue, but this time with “traditional” names. In English the two names “fluorescin” and “fluorescein” differ only by a single character, but when transliterated to Japanese this distinction is lost. In addition to round-trip benchmarking, Lexichem’s translation functionality has also been evaluated by native speaking chemists and used to translate names found in standards documents and chemical supplier catalogues. For example, the company ChemBlink provides searchable online databases, providing names of its products in both English and Chinese.  

Comparison to Existing Systems. To evaluate the described algorithms against existing machine-translation systems, a small benchmark set of chemical names were translated using SYSTRAN based software available via Altavista’s babelfish, Google translate, or Yahoo! Translate. The results are summarized in Figure 15. The left-hand column contains the input phrase, and the right-hand column contains the result. For the compound names shown in Figure 15, all three online translation tools returned identical translations indicating their common SYSTRAN heritage or the use of a standard dictionary/training set. The strange typography (inappropriate spaces) reflects the actual results returned. On this almost trivial test set, Lexichem perfectly translates all of the compound names giving the expected names given in the left-hand column.

As can be seen in Figure 15, existing state-of-the-art machine translation software performs poorly on chemical nomenclature. Presumably, the software assumes that English and German names are delimited by spaces and uses dictionary-based approaches to perform the actual translation. While this works fine for simple names such as “benzene” and “propane”, it breaks down completely when faced with the more usual compound names shown above. Even when translating from Japanese, the software suffers from the difficult vowel elision rules and unusual character composition of IUPAC names.

We finish with an example of economic value to the pharmaceutical industry. The Japanese Patent Office (http://www.jpo.go.jp/) makes the contents of filed patents available online electronically via the Industrial Property Digital Library (IPDL). Among the services provided is the ability to translate the textual contents of the filing into English. Taking as an example, a patent recently filed by Osterhout and Roschangar from GlaxoSmithKline, Japanese Patent Number 2008-50363, the exemplified compound being claimed in claim number 8 is given in Japanese as “5-(4-[3-chloro-4-(3-fluorobenzyloxy)anilino]-6-quinazolinyl)-furan-2-carbaldehyde”. The automatic English translation provided by the Japanese patent office is the less than helpful “5. -4-[aniline [ the 3-chloro-4-(3-fluoro benzyl oxy)-]-6- chinae-cortex ZORIN-IRU]- franc 2-carbaldehyde”. The correct translation (produced by Lexichem) is “Z.-4-[3-chloro-4-(3-fluorobenzyloxy)-anilino]-6-quinazolinyl]-furan-2-carbaldehyde” which has the structure shown in Figure 16. The correctness of the translation can be confirmed (in this case) by comparing to
spelled English names can be automatically corrected. The nomenclature several researchers have shown how incorrectly edge of the restricted grammar and lexemes of IUPAC also frustrates spelling checking software. By using knowl-
dictionary lookup difficult/impossible for chemical names, such as "lead" in the phrase "a lead compound". The deeper semantic analysis performed by traditional machine-transla-
caracteristics and is easily tripped up by ambiguous terms such as "lead" in the phrase "a lead compound". The deeper semantic analysis performed by traditional machine-transla-
tion software may significantly help the process.

One interesting area of investigation is the field of spelling correction. The same morphology differences that makes dictionary lookup difficult/impossible for chemical names, also frustrates spelling checking software. By using knowl-
dge of the restricted grammar and lexemes of IUPAC nomenclature several researchers have shown how incorrectly spelled English names can be automatically corrected. The techniques presented in this paper should allow such approaches to be extended to multiple languages.

Another important area of research is how best to integrate the special-purpose translation of chemical names, such as the expert system described here, into a more general machine translation framework. Identifying the chemical names in a complex document and translating them inde-
minedly is related to the text mining field of entity extraction. One method of chemical name entity extraction that works moderately well is simply to pass phrases to chemical name parsing software, and if it is able to return a result, the phrase is assumed to have represented a chemical name. While this is a reasonable first approximation, it is limited by the current machine-interpretation rates of chemi-
catalogues) is unlikely to be significantly worse than from foreign language documents (such as patents or compound catalogues) is unlikely to be significantly worse than from foreign language documents. This means that using software approaches similar to those described, the recall and indexing of terms from foreign language documents (such as patents or compound catalogues) is unlikely to be significantly worse than from native English language documents. The availability of special-purpose translation software for converting chemical names is also likely to assist and greatly simplify the task of preparing these technical documents in languages other than English.

ACKNOWLEDGMENT
The author wishes to thank OpenEye Scientific Software for the freedom to investigate the issues of chemical name translation. Andrew Grant of AstraZeneca Pharmaceuticals was instrumental in originally motivating the problem and assisting with Swedish, Spanish, and Italian support. Ikutoshi Matsuura (retired from Chugai), Takahiro Ohe of OpenEye Japan, and Yutaka Yamada of Cybernet Systems assisted with Japanese support. Maria Gallardo de Zaragoza assisted with Spanish support. Stanislaw Włodek assisted with Polish support. Kristzina Boda assisted with Hungarian support. Tudor Oprea and Sorel Muresan assisted with Romanian support. Peter Erli assisted with Slovak support. Many thanks also to Paul Hawkins at OpenEye, for proofreading and correcting early drafts of this manuscript and to the anonymous reviewers for their insightful comments. I would also like to acknowledge the impressive Internet resources made available by the International Union of Pure and Applied Chemists (IUPAC), by Tomoyuki Sadahiro tsadahiro@nifty.com on Japanese chemical nomenclature, and by the Wikipedia Foundation, for Wikipedia www.wikipedia.org an unparalleled resource for translators.

REFERENCES AND NOTES
(1) Commission of the European Communities. Rechtswrifschriften für Gefährliche Stoffe: Einstufung und Kennzeichnung in der Europäischen Gemeinschaft. 67/548/EWG, 1987.
(2) Kirby, G. H.; Lord, M. R.; Rayner, J. D. Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 6. (Semi-)
(3) Garfield, E. Chemico-linguistics: Computer Translation of Chemical Nomenclature. Nature 1961, 192–194.
(4) Vander Stouw, G. G.; Elliott, P. M.; Isenberg, A. C. Automated Conversion of Chemical Substance Names to Atom-Bond Connection Tables. J. Chem. Doc. 1974, 14 (4), 185–193.
(5) Cooke-Fox, D. I.; Kirby, G. H.; Rayner, J. D. Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 1. Introduction and Background to a Grammar-based Approach. J. Chem. Inf. Comput. Sci. 1989, 29 (2), 101–105.
(6) Google. http://www.google.com/ (accessed December 10, 2008).
(7) The Klingon Language Institute. http://www.kli.org/ (accessed December 10, 2008).
