SYSTRAN MT DICTIONARY DEVELOPMENT

Laurie Gerber, Jin Yang

SYSTRAN Software, Inc.
7855 Fay Avenue, Suite 300
La Jolla, CA 92037
(619) 459-6700
lgerber@systransoft.com, jyang@systransoft.com

Abstract

SYSTRAN has demonstrated success in the MT field with its long history spanning nearly 30 years. As a general-purpose fully automatic MT system, SYSTRAN employs a transfer approach. Among its several components, large, carefully encoded, high-quality dictionaries are critical to SYSTRAN's translation capability. A total of over 2.4 million words and expressions are now encoded in the dictionaries for twelve source language systems (30 language pairs - one per year!). SYSTRAN'S dictionaries, along with its parsers, transfer modules, and generators, have been tested on huge amounts of text, and contain large terminology databases covering various domains and detailed linguistic rules. Using these resources, SYSTRAN MT systems have successfully served practical translation needs for nearly 30 years, and built a reputation in the MT world for their large, mature dictionaries.

This paper describes various aspects of SYSTRAN MT dictionary development as an important part of the development and refinement of SYSTRAN MT systems. There are 4 major sections: 1) Role and Importance of Dictionaries in the SYSTRAN Paradigm describes the importance of coverage and depth in the dictionaries; 2) Dictionary Structure discusses the specifics of dictionary structure and types of information represented; 3) Dictionary Creation and Update describes the strategy and mechanics of the dictionary development; 4) Past. Present and Future Development provides some perspective on where SYSTRAN has come from and where it is going.

1. The Role and Importance of Dictionaries in the SYSTRAN Paradigm

The problem of automatic linguistic analysis is one of accumulating and utilizing linguistic knowledge. Much of the needed knowledge is static and can be built into the system and its dictionaries, e.g. standard syntax, and conventional word use, including the variations typical of different text types and domains. MT systems encode knowledge in a variety of ways. Linguistic rules, probabilities, and databases of examples are typical encoding methods. SYSTRAN, which is a rule-based transfer system, encodes linguistic knowledge primarily in two ways: 1) In the dictionaries which contain “bottom-up” parse rules along with extensive syntactic and semantic information about each word; 2) In linguistic programs which contain the general “top-down” rules in the parser.

Dictionary coverage, or size, is critical to high-quality translation. Words not found in a system's dictionary are a challenge to parse correctly. For SYSTRAN, broad coverage has always been a high priority because of the extensive use of SYSTRAN for information gathering in a wide variety of text types and domains (Gachot 1996.)

Depth, or richness, of encoding is a defining feature of SYSTRAN's dictionaries. A priority is placed on broad vocabulary coverage, and lexical items are heavily encoded with syntactic and semantic features. This has meant that: 1) During development, dictionary building is a relatively slow, hands-on process. 2) In practice, the resulting systems have high accuracy because the static idiosyncrasies of each word are captured in the dictionaries. An evaluation of major commercial MT systems by CompuServe in 1993 stated, "Dictionary complexity, on the other hand, often provides a rough gauge of the output quality and improvability of a system. MT systems with highly detailed information about terms often provide better translations simply because they have access to more information" (Flanagan, 1993).

The realization of the coverage and depth required is made possible by a well-defined development procedure and a number of tools. Although dictionary building is labor-intensive, it is still done efficiently. Educated bilinguals without extensive training can do most of the lexicography. The assignment of even hard-to-understand codes can be determined using examples and templates provided in SYSTRAN training and reference materials. The emphasis on the dictionary encoding of words has meant that a large number of lexicographers and a small number of highly trained linguists can do development of new MT systems.

The history of dictionary building at SYSTRAN has been one of constant striving for efficiency, and evolutionary refinement of procedures. Methods for obtaining terms, and even the internal record structure of dictionary entries, have been changed gradually as improvements became possible. However, we have continued to be satisfied with the type and depth of information stored in the dictionary entries. The time-tested validity this informational content continues to serve SYSTRAN MT systems well, while innovations in methodology are incorporated.

2. Dictionary Structure

For large-scale production MT systems, knowledge acquisition is not the only challenge. As more and more knowledge is built into the system, managing the resulting complexity becomes a vital issue (Scott 1994.) Within SYSTRAN, one strategy is to keep word-specific rules out of the linguistic modules, so that they contain only the generalized rules. Word-specific rules reside, as much as possible, in the stem and expression dictionary, and to some extent in the transfer programs called "lexical routines" which disambiguate multiple meaning words, and perform other processing for high-frequency words, or word classes which require special transfer processing. In addition to the advantage of keeping the main linguistic modules relatively uncluttered, the accumulation of ad hoc rules in the expression dictionary often provides excellent data on which generalizations can later be based.

Two types of dictionaries are developed for each source language: stem dictionaries, and expression dictionaries. The main function of the stem dictionary is to provide lexical information for linguistic analysis. The expression dictionary supplements the general rules in linguistic analysis and transfer with word-specific rules. Most of SYSTRAN's dictionaries are multitarget, with source information for a single source word or expression, and target information for as many languages as the source language links to. A typical entry consists of a source language portion, a target language portion, and one optional field for transfer information on the translation of prepositions.

While the stem dictionaries are based on single words in the source language, entries in the expression dictionaries (called "LS" dictionaries in-house) come in a variety of forms, as described in section 2.3. Their function ranges from creating "pseudo-stems" for multi-word adverbials, to conditional rules for the analysis or translation of single words, or providing meanings for collocations such as technical noun phrases.

2.1 Stem Dictionaries

The base forms of single words in the source language are entered in the "stem" dictionary. Each source language word is accompanied by extensive grammatical information about its morphology, syntactic behavior, semantic content, and homographs[1]. Properties assigned to target language meanings include part of speech, morphology, syntactic behavior, and preposition translation. Multiple meanings may be assigned in the stem dictionary for different domains (via "topical glossary"), transitive and intransitive meanings on verbs, animate and inanimate uses of nouns. Unlimited additional meanings may be assigned conditionally via the linguistic programs or the expression dictionaries.

2.1.1 Linguistic Knowledge - Source Language Information

SYSTRAN's stem dictionaries contain a variety of lexical data. This information, represented by various codes, is all assigned to the source language portion of a stem dictionary entry, to be used in the analysis phase of translation. This essential core of data includes:

- Part of speech code: Typically 80-100 parts of speech are defined per language.
- Gender and Number
- Homograph pattern: Between 40 and 100 patterns per language. Homographs are resolved by linguistic programs, with routines for each of the defined patterns.
- Inflection or pattern code: Identifies an entry with a table of valid endings. This is used for matching words during dictionary lookup, and parsing tense, person, number, and any other information available from inflection.
- Semo-syntactic codes: Many of these describe a word's syntactic behavior, others include semantic content as well. 150 codes are defined for use in all languages:

  ANSUB = Prefers animate subject (coded on verbs)
  ADVNN = Adverb may modify noun (coded on adverbs)
  GI = Word can govern infinitive (coded on various)
- Semantic tagging: SYSTRAN has defined 500 semantic categories which are organized into six hierarchical trees. Stem dictionary entries may be assigned as many categories as necessary to define the semantic content of a word. The semantic codes are organized hierarchically to allow lower nodes to inherit the properties of all superior nodes. For example, the semantic category ORALEX (oral expression) is a lower node on the PROCES (process - used mainly for verbs and verbal nouns) tree. A word encoded with ORALEX, also will automatically inherit the properties of the superior nodes. The semantic code field on such a word would look like this:

  ```
  SEM- ORALEX (SONIFY, TRANSM, GIVOUT, GIVE, PRPHY, ACT, PRGEN, PROCES)
  ```

  Only the code ORALEX was explicitly assigned. The inherited codes are in parentheses, and are listed from the specific to the general:

SONIFY	to make sound
TRANSM	to transmit something
GIVOUT	to give out or emit something
GIVE	to give
PRPHY	physical process
ACT	action
PRGEN	general process
PROCES	process (head of taxon)

[1] In SYSTRAN terminology, "homograph" refers only to words with identical written forms and multiple parts of speech, e.g. "judge" (noun or verb). A polysemous single written form is referred to as a "multiple meaning word".

- Domain usage: Vocabulary generally limited to a specific domain may be tagged with one of 77 domain codes.
- Document type: Vocabulary generally limited to a specific document type, e.g., patents, minutes, etc., may be tagged as such.

2.1.2 Linguistic Knowledge - Target Language Information

As mentioned above, SYSTRAN's dictionaries are typically multitarget. They contain source language entries, accompanied by the above attribute assignments, and target language entries for every target language accessible from that source language. The core of target language information includes:

- Preferences for preposition translation: Strictly speaking, this may be considered a "transfer" field. A translation may be specified for any preposition which governs, or is governed by the source language word.

- Target language meaning, together with morphology and information on syntactic behavior:
    - Part of speech
    - Meaning identifiers (domain-specific meaning, animate or inanimate noun, transitive or intransitive verb, reflexive or non-reflexive verb)
    - Inflection pattern
    - Article usage information
    - Gender and number
    - Syntactic information - comparable to the semo-syntactic codes assigned on the source level, but without the semantic component. 38 codes defined.

Note again that for a single target language, multiple meanings may be assigned as follows:

- As many parts of speech as necessary.
- All parts of speech may be assigned specialized meanings for any or all of 20 different domains[2].
- Nouns which, in the source language are animate/inanimate ambiguous may be assigned animate and inanimate meanings.
- Verbs which in the source language have transitive and intransitive uses, may be assigned transitive and intransitive target language meanings.
- Polysemous words may have multiple meanings which can be selected by transfer phase programs called "lexical routines".

2.2 Expression Dictionaries

Several categories of entries may be encoded in the expression dictionaries. They are listed here in order of increasing complexity. Each of the types is followed by a typical example in English (rules which would be written in SYSTRAN notation are paraphrased) to illustrate the type of expression:

[2] At SYSTRAN, these are referred to as "topical glossaries". We do not maintain separate domain-specific source dictionaries, all words are part of the same stem dictionary, the domain differences are handled via source language codes described below, and the availability of a variety of domain specific meanings. These meanings are selected by the user, who may indicate up to 4 topical glossaries, in order of preference, when running a translation.

- Idiom Replace: This type of entry allows frozen idiomatic expressions and multi-word adverbials to be . fused into a single “pseudo stem” which can then be entered in the stem dictionary with a single part of speech. In the resulting pseudo stem, the words are joined by periods to yield a single string.
  be that as it may -> be.that.as.it.may (adverb)
- Collocation: Used to assign meaning to strings of words which may be unconditionally replaced with a target string. Most useful for conventionalized technical noun phrases. Constituent words may still be inflected.
  butterfly nut
  hermetic seal
- Conditional Expression: Used when meanings or other target language information needs to be added only under certain conditions. Conditions for meaning assignment may be quite elaborate, and may specify any of the syntactic or semantic relationships or semantic properties defined. These expressions are invoked at the transfer phase of the translation process, and may also assign target language information or perform other transfer phase parse modifications.
  If word is *"eat"* and direct object has semantic category METAL, translate as *"corrode"*
- Parsing Expression: Used to apply word-specific rules to the parse. Especially useful for early resolution of polysemous words, or those with multiple syntactic usage patterns. Any source language dictionary or parse information may be added or modified, and semantic attributes added or deleted.
  If word is *"turn"* check right for *"off"*. Force *"off"* to function as an adverb, resolve *"turn"* as a verb.
- Homograph Resolution - These expressions disambiguate and assign the correct part of speech for a single word. Used on a limited basis for simple, unique homograph problems, they are considered to be a specialized sub-type of parsing expression. See example for parsing expression.

3. Dictionary Creation and Update

Various techniques are employed throughout the dictionary development process to build and refine these databases of the translation systems.

3.1 Development Principles

Production oriented: MT development at SYSTRAN has always focused on rapid development of production systems, which motivates efficiency and practicality. At the same time, as a rule-based system, building lexicons and grammars is the most-time consuming concern (Nirenburg et. al. 1996). And since the parsers depend so heavily on rich dictionary encoding, quality cannot be sacrificed. The result of these two demands is that dictionary development is automated as much as possible, while adhering to quality assurance by machine-aided manual validation.

Linguistic rules empirically motivated: Dictionary development as well as linguistic rule development at SYSTRAN is typically done using a combination of two types of resources: 1) extensive live text examples; 2) published linguistic and grammatical reference works, with the emphasis on live text examples (Yang and Gerber 1996).

Technical domain focus: Technical text tends to follow the usage conventions of the domain. This focus means that relatively predictable use of words dominates.

3.2 Dictionary Building Tools

Automate where possible – "fast coding" methods: Various tools have been developed to automate the dictionary building process as much as possible, using whatever electronic resources are available. These methods are sometimes as simple as inputting word lists (where electronic resources are unavailable) together with desired meanings, running them through the system and modifying meanings or adding entries. When more electronic resources are available, words may be coded directly from electronic dictionary resources. When text corpora are available, frequency lists are generated to facilitate addressing high-frequency items first. Concordance listings may be generated to facilitate coding of difficult-to-code words. These tools, which facilitate the process, are referred to as "fast coding" tools or interfaces, and include a variety of utilities to help lexicographers identify words and expressions which need to be entered and pre-validate their work.

Information retrieval tool for developing conditional and parsing expressions: When preparing to add a conditional or parsing expression rule to the expression dictionary, lexicographers are expected to gather representative data from on-line corpora to ascertain whether the rule will work in all and only the cases intended. For this purpose, they use a parsing diagnostic text retrieval tool now referred to as InfoRaptor (Gachot, Lange and Yang 1996)

Customer Specific Dictionary (CSD): Since the commercial release of SYSTRAN Professional for Windows, a dictionary building tool has been provided as part of the product. This allows users to build their own termbanks to supplement, or in some cases override, the terms supplied in the main dictionaries. Users often operate in very limited domains for which only one possible use of a word is necessary. The CSD utility allows encoding of the primary parts of speech, and generates morphological endings automatically.

3.3 Dictionary Verification Process

Managing rapid expansion: Work done by lexicographers is incorporated regularly (usually monthly) into the central dictionary for each source language. The incorporation process is called "dictionary update" and is partially described in this section. These regular updates, followed by regression testing, ensure smooth coordination between each lexicographer's work and linguistic rule development.

Manual validation after automatic acquisition: All of the automated dictionary development tools are interactive, presenting the results for human review. Because the priority is on quality of coding, quotas and speed are not emphasized to lexicographers. Instead, the work of individual lexicographers is subject to a final review or spot-check to facilitate ongoing training and skill development.

Dictionary "Edit" and dictionary "Audit": The dictionary update process consists of several steps, beginning with validation. Validation detects improperly formed lexical entries on various levels. The most basic check, "edit", is for the correctness of the dictionary format. The second check, "audit", is for the logic and consistency of codes of each lexical entry. For example, noun syntactic codes should not be coded on verb entries, proper nouns (in English) should have capitalized meanings, etc.

Incorporation of new entries: This process does further validation which will prevent the incorporation of exact duplicate entries, or entries with incorrect linking as cross-references to existing entries.

3.4 Dictionary Refinement

In addition to the entry-by-entry validation, quality refinement is also done on the system as whole. In order to increase the translation quality of a system, ongoing review and testing help to ensure consistency and thoroughness in coding, and to protect the system from accidental degradation.

Dictionary review: This generally consists of a thorough review of groups of entries. For example, all of the entries with a particular part of speech or syntactic code may be extracted and reviewed as a group to make sure that the coding has been consistently and appropriately applied. This method helps to catch the occasional error, as well as making sure that our dictionaries adhere to the rules and conventions for building them.

Regular regression testing: Although conscientious efforts are made to check and review dictionary entries, the impact of new terms or modifications on translation is impossible to predict. Various types of risk require testing the impact of dictionary modifications on large volumes of text. Three of the most common are:

- Translation meaning selection: Lexicographers select the translation they anticipate will be the most generic meaning that still encompasses the full meaning of the source word. Typically the target language meaning is selected from a bilingual dictionary. However, the target language word that provides the best translation in the context of a bilingual printed dictionary, may not always work well in the context of the words it will appear with in the translation output.
- Part of speech assignment: It is our experience that printed dictionaries do not represent the same distribution of usage found in live text. Sometimes additional part of speech entries are necessary after viewing the translation of words in context.
- Unintended side effects: Improvements in one area occasionally interfere with necessary functionality in another.

Dictionary updates are followed by regression tests run on large (1,000-6,000 sentences), well-varied texts. The risk of degradation in performance is taken very seriously and managed aggressively with regular testing of output that is compared against previous versions.

4. Current Status and Future Development - Punched Cards to Automatic Extraction

Dictionary development at SYSTRAN has evolved along with available technology and resources for nearly 30 years. The early process was laboriously manual: lexicographers checked entries to be added against a paper printout of the existing dictionary. The new entry was hand-written on a letter or legal-sized form - one word per page. Each stack of 50 new words was submitted for data entry on punched cards. The punched cards were checked, and then updated on a mainframe. In those days, a lexicographer who could average 50 entries per day (~1,000 per month) was a marvel!

These days, it is not unheard of for a lexicographer to generate 5,000 entries per month. The increase in speed has been made possible by the increased power and flexibility of the development environment (PCs and UNIX) and the availability of electronic dictionaries and corpora have greatly facilitated the creation of tools to automate the process.

Because of the need to keep all of SYSTRAN production systems production-ready, change to the methodology tends to come in evolutionary steps rather than risky overhauls. We are excited about the potential of the newer methods and ideas, particularly corpus manipulation, discourse structure, and statistical inference to provide new momentum and leverage to complement our existing strength.

SYSTRAN continues to explore further avenues to automate high-quality dictionary building. Two tools under current development for implementation in-house are:

- An automatic phrase finder to facilitate the location of multi-word terms in live text.
- An automatic conditional expression builder. The prototype of this tool utilizes the SYSTRAN parser to extract pairs of words/phrases in specified syntactic relationships which often require special translations. For example: verb+direct object; verb+prepositional phrase; subject+verb.

As we pursue greater efficiency and automation in dictionary building, we will preserve the rich, detailed encoding that has provided such a solid foundation for the parsers, and is an integral part of the SYSTRAN architecture.

References

Flanagan, M. 1993. Evaluating MT for Message Translation. In *Proceedings of the 34th Annual Conference of the American Translators Association.*

Gachot, D. 1996. Assimilation or Dissemination? That is the Question. In *Proceedings of the Second Conferee of the Association for Machine Translation in the Americas.* October 2-5, 1996. Montreal, Quebec, Canada.

Gachot, D., Lange, E. and Yang, J. 1996. The SYSTRAN NLP Browser: An Application of MT Technology in Multilingual Information Retrieval. In *Proceedings of Workshop on Cross-Linguistic Information Retrieval,* August 22, 1996. Zurich, Switzerland.

Nirenburg, S., Beale, S., Helmreich, S., Mahesh, K., Viegas E., and Zajac, R. 1996. Two Principles and Six Techniques for Rapid MT Development. In *Expanding MT Horizons: Proceedings of the Second Conference of the Association for MT in the Americas.* October 2-5, 1996. Montreal, Quebec, Canada.

Scott, B. 1994. The Logos View. In *Proceedings of the first Conference of the Association for Machine Translation in the Americas.* October 5-8, 1994. Columbia, Maryland, USA.

Yang, J. and Gerber L. 1996. SYSTRAN Chinese-English MT System. In *Proceedings of the International Conference on Chinese Computing '96.* June 4-7, 1996. Singapore.