SYSTRAN Intuitive Coding Technology

Jean Senellart

SYSTRAN S.A.
1, rue du Cimetière
95230 Soisy-sous-Montmorency
France
senellart@systran.fr

Jin Yang

SYSTRAN Software, Inc.
9333 Genesee Avenue
Plaza Level, Suite PL1
San Diego, CA 92121
USA
jyang@systransoft.com

Anabel Rebollo

SYSTRAN S.A.
1, rue du Cimetière
95230 Soisy-sous-Montmorency
France
rebollo@systran.fr

Abstract

Customizing a general-purpose MT system is an effective way to improve machine translation quality for specific usages. Building a user-specific dictionary is the first and most important step in the customization process. An intuitive dictionary-coding tool was developed and is now utilized to allow the user to build user dictionaries easily and intelligently. SYSTRAN's innovative and proprietary IntuitiveCoding® technology is the engine powering this tool. It is comprised of various components: massive linguistic resources, a morphological analyzer, a statistical guesser, finite-state automaton, and a context-free grammar. Methodologically, IntuitiveCoding® is also a cross-application approach for high quality dictionary building in terminology import and exchange. This paper describes the various components and the issues involved in its implementation. An evaluation frame and utilization of the technology are also presented. Future plans for further advancing this technology forward are projected.

1 Introduction

1.1 Background

Customizing a general-purpose Machine Translation (MT) system is an effective way to improve MT quality. MT customization projects have been implemented to varying degrees, and can be performed a) by the MT system developers; b) by the user, or c) by the collaboration of the two. For example, the SYSTRAN systems powering most of the Internet portals represent general-purpose MT systems based on the largest possible dictionaries. Their aim, though, is only to provide a general translation at the "gisting level" (Yang & Lange, 1998). In contrast, the MT systems used by the European Commission (EC) since 1976, have been deeply customized for the type of texts commonly used by the EC. Recent production-scale customization applications include the combination of controlled language and User Dictionary (UD) for a vehicle assembly process (Rychtyckyj, 2002), and the translation of online technical support documentation (Senellart, 2001). The MT system developers were actively involved in the above-mentioned customization projects. User participation in MT customization projects varies, and may include: the provision of domain and/or user specific glossaries, the review and ongoing assessment of translation quality, and suggestions for improvement. For production-scale customization projects, the size of user dictionaries is huge (50,000 to 200,000 entries). The domains are very specific, and the dictionary content is proprietary in nature. Considering the size of the user dictionaries, the need for present and future updates, and the proprietary nature of dictionary content, the MT users must, at a certain point, be able to independently maintain, expand and sustain continuous customization, with little or no involvement from the MT developers.

The ideal solution is to enable the user to perform customization tasks. There are, however, many challenges to this. First of all, users who are usually language specialists, do not necessarily have the computational linguistic expertise—let alone a deep understanding of MT in general, or the knowledge of a particular MT system. Instead, these language specialists are experts of certain domains, and in a specific language. Therefore, the immediate challenge is to rapidly and intelligently turn their specialized information into the knowledge representation of the MT system. This is the core of the Intuitive Coding (IC) process. In this process, the "coding" system is the interface between the target language and the language specialists. The interface needs to be flexible, interactive, robust, and most importantly, intuitive. This paper, describes the SYSTRAN IntuitiveCoding® technology (ICT), which powers the intuitive coding process of user dictionary entries into complex knowledge representation for customizing SYSTRAN general-purpose MT systems.

1.2 Intuitive Coding

The requirements for the Intuitive Coding process are as follows:

- The user: The user is a bilingual or multilingual language specialist of a particular domain. No other linguistic or MT expertise and experience is required.
- Intuitive dictionary representation: The representation of user dictionary entries should be simple and intuitive, such as are paper dictionaries.
- Automatic process: The information from user dictionary entries is automatically converted into the knowledge representation that the MT system requires. In other words, the process can transform the user dictionary entries into a functional MT dictionary without human intervention.
- Interactive process: The process can be interactive. The coding system outputs quality and "risk" analysis of the user entries. The user can provide any changes through a feedback cycle.
- Multi-level coding formalisms: Intuitive coding not only supports simple entries, but also includes advanced entries. The coding can be fully intuitive—let the system do the "magic". The user has more control via the advanced coding mechanism.
- Complete integration: User dictionary entries should be completely integrated into all-level processing of the MT system. The integration is sometimes language-specific—for example, Chinese entries may have impact on the Chinese word segmentation. Moreover, the interaction between user entries and the existing entries in the MT dictionary should bear defined parameters at the user-level.
- Easy-to-use graphic user interface: The coding interface should be implemented as an easy-to-use graphic user interface. In addition, a production-scale customization deals with large glossaries. Various issues, such as specific "sort" operations, duplicate checking, and real-time processing, need to be addressed.

With such a tool, the user can do the following:

- Give a technical equivalent for a general word;
- Define the specific meaning of a word with multiple possible translations;
- Add words that are not part of standard MT dictionaries;
- Add multi-word expressions that are not part of standard MT dictionaries; and
- Specify contextual rules.

Sample English-French user dictionary entries in the intuitive coding format are given as follows:

```
a download store=une boutique en ligne
a drive shaft=un arbre d'entraînement
a watering can=un arrosoir
"all rights reserved" (sentence)="tous droits réservés"
(sentence)
to save (context: money)=économiser
```

2 Technology Involved in Intuitive Coding

In essence, the IntuitiveCoding® technology enables the automatic and intelligent conversion of simple user dictionary entries into the knowledge representation of the required MT system. For the user, Intuitive Coding is the practice of adding intuitive grammatical clues to an entry, in order to provide more information on its nature. For the system, Intuitive Coding is the capacity of using implicit information to enrich user dictionary entries with general linguistics or specific MT information.

In practice, Intuitive Coding can be multi-level. The starting point is the representation and structures found in paper dictionaries. The upper limit (also known as advanced coding) allows high-level interaction with the MT dictionaries. More importantly, Intuitive Coding can cover the majority of dictionary entries needed for a production-scale MT customization.

2.1 Technical Components

Intuitive Coding Technology is based on:

- *Monolingual dictionaries*

The information in monolingual dictionaries is derived from the MT dictionaries. Each entry (single word, or multi-word expression) consists of morphological, syntactical and semantic features as coded and used by the MT system. In other words, the linguistic information used in the source language analysis and the target language generation are the basis for the monolingual dictionaries, but not the information for source-target transfer.

- *A statistical guesser*

The statistical guesser is used to compute a list of potential categories, their morphological codes, possible syntactic features and probability weights for any word (found or not found in the monolingual dictionaries).

- *A statistical context-free description of compound structures*

The statistical context-free description of compound structures is used to analyze compound entries. The linguistic description is a context free grammar, with associated linguistic probability. For example, to analyze the French “moyenne tension electrique” as an [(ADJ (NOUN) ADJ)] noun phrase, the system uses the following rules:

noun adjective $\rightarrow_{0.99}$ noun
noun noun $\rightarrow_{0.3}$ noun
adjective(+left) noun $\rightarrow_{0.99}$ noun
adjective(+right) noun $\rightarrow_{0.6}$ noun

- *A set of intuitive clues*

The intuitive clues are the rules describing how to interpret intuitive information presented in user dictionary entries. For example, the rules extract information from the particle, determiner, natural agreement or “elision” manifestation in entries, as illustrated in the following examples.

Rule	Entry	Information presented in the intuitive clue
to + verb (infinitive)	to rule	“to” can be the signature of English verbs
s’ + verb se + verb (starting with vowel)	s'attendre	The French verb is reflexive. And there is elision between the pronoun and the verb
noun (ms) + prep + noun (fs) + adj (ms)	pilote de course fameux	Adjective does not agree with feminine noun “course”, excluding the NOUN PREP (NOUN ADJ) structure.

- *Confidence*

The statistical nature of Intuitive Coding is represented by the “confidence” factor. The confidence for each entry is computed using the statistical probability associated to the rules, the result of dictionary lookup, and the probability produced by the guesser. The result of the intuitive coding process is a list of possible “coding” options for a user entry. The multiple results of each monolingual dictionary entry are used, along with the alignment rules, to reach the final confidence score for a bilingual or multilingual entry.

- *Alignment rules*

The coding result of a multilingual entry for each language is a list of possible structures with probability weights. Alignment rules select the combination of monolingual entries with the higher weight. The final weight of the multilingual entry is computed based on the combined weight of each monolingual entry and alignment rule.

en:to refill= fr:boucher	“*boucher*”: *to fill* (verb), *butcher* (noun) The verb entry of “*boucher*” is selected per alignment rule with the explicit English verb structure “*to refill*”.
fr:avocat= en:lawyer	“avocat”: *lawyer* (noun, human), *avocado* (noun, fruit) The human meaning of “*avocat*” is selected per alignment rule <N:HUMAN>=<N:HUMAN> with English “*lawyer*”
fr:petite ferme= en:small farm	“petite”: *small* (adj) “ferme”: *farm* (noun) The structure ADJ-NOUN is selected for the French, instead of the more common NOUN-ADJ structure.

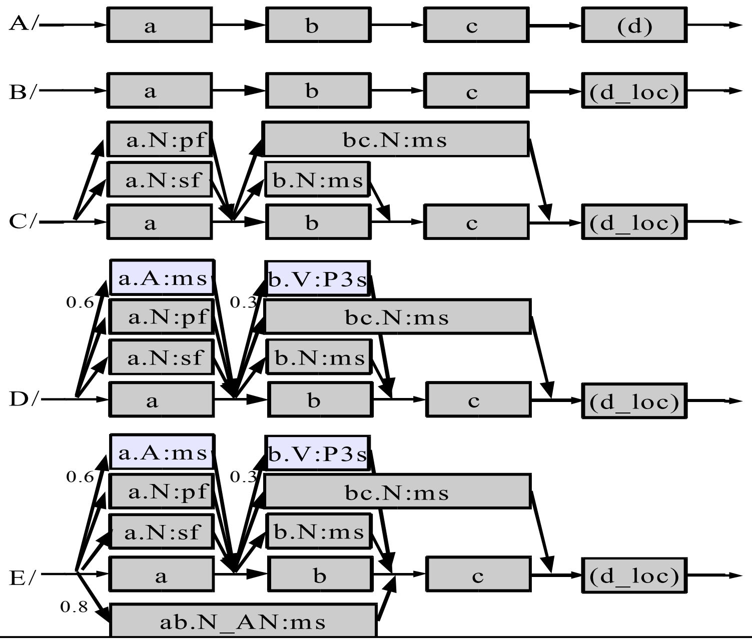


Figure 1 - Steps of monolingual coding. **A/** represents original entry where *(d)* is a clue. In B/ clue is normalized. In C/ dictionary applies on entry. In D/ dictionary is completed by guesser. E/ show transformation of automaton structure by application of morphological rule.

Typically, the alignment rules are used to select the part of speech of a homographic entry, the structure of an ambiguous compound structure, or even syntactic and/or semantic features of the entry. Examples are shown above.

- *Advanced coding*

Advanced coding allows a higher level of customization. It is the practice of adding sophisticated linguistic information (semantic, syntactic, morphological, and contextual) on the nature of an entry. Each language has its own set of linguistic clues (e.g. http://www.systransoft.com/Support/Dicts/Tables/latest/en.html). See Table 1 (sample English linguistic clues.)

Reference	Variants	Category	Example					
				noun	proper noun	acronym	verb	adj
A		intuitive	a lower hinge	X				
To		intuitive	to right-click				x	
Singular	sg	morphology	news (singular)	X				
Plural	pl	morphology	Trousers (plural)	X				
to somebody	to sbdy	syntax	to talk (to somebody)				x	
Somebody	sbdy	syntax	to call (somebody)				x	
Something	sth	syntax	to design (something)				x	
last name		semantic	Bush (last name)		x			
Human	hu	semantic	absorber (human)	X	x			
company name		semantic	Apple (company name)		x			
Object		contex	to configure (object : bridging)				x	

Table 1. Sample Linguistic Clues (English)

- *Finite-state automaton*

All dictionary lookup and rule matching are performed using a low level FSA library, which provides exact matching lookup and localized matching operations. The localized matching includes lookup mechanisms for generic language-specific variants, for example, traditional and simplified Chinese characters for Chinese, acceptation of unvowellized form in Arabic etc.

2.2 Steps

The steps of the Intuitive Coding process of monolingual entries are illustrated in Figure 1.
The figure plans the coding of an example entry "abc (d)", and the graphics shows the enrichment of the structure during the process:

1. The entry is tokenized. The structure is basically a list of tokens.
2. The linguistic clues are normalized, including localization and abbreviated clues.
3. Dictionary lookup is performed. The structure is enriched with different dictionary entries. In the example, "a" can be either noun (+plural, +feminine) or noun (+singular +feminine). Lexical compounds are equally matched (bc). Each path of the structure represents a different analysis.
4. The guessing process is applied (the light gray box). This module introduces confidence on transition (i.e., "a" can be an adjective with a confidence 0.6).
5. The composition rules are applied and syntagms are built. Here "ab" is analyzed as a noun with adjective-noun structure, and confidence is 0.8.

The monolingual coding for a bilingual or multilingual entry is performed in parallel for each language. Multilingual alignment is then applied based on the monolingual coding results. An example of multilingual alignment is shown below.

Language $_1$	Language $_2$	Language$_3$	Result
$(a_{Noun}\ (bc)_{Noun})_{Noun}$ /0.63	S_1 /p_1	T_1 /q_1	$(a_{Adj}\ (bc)_{Noun})_{Noun}$ =S_2 =T_1
$(a_{Adj}\ (bc)_{Noun})_{Noun}$ /0.52		T_2 /q_2	
	S_3 /p_3	T_3 /q_3	

Table 2 Multilingual Alignment

Each column represents the list of results for each language. The "Result" column is the selection of the most likely combination of monolingual results, according to alignment rules.

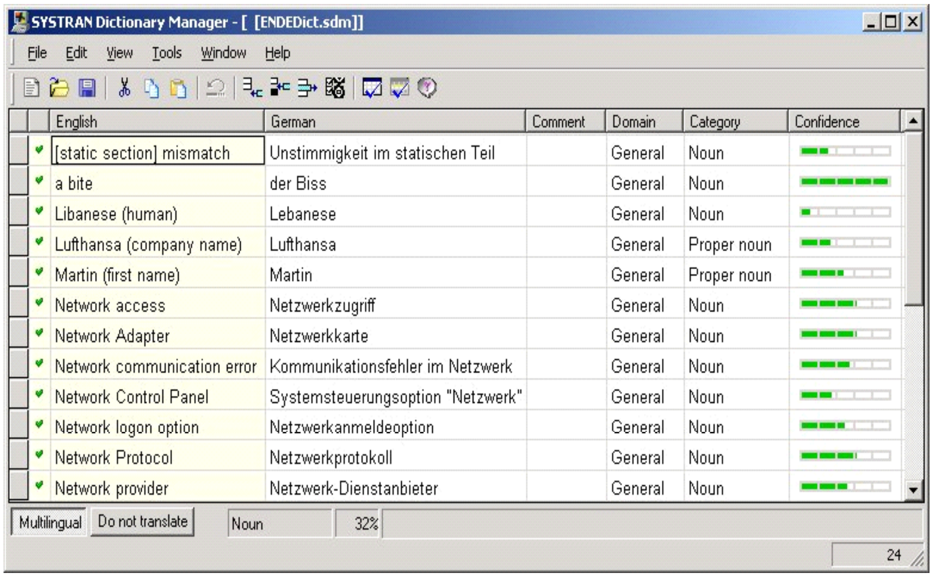
2.3 Issues

When an entry in the user dictionary is intended to replace the translation from the MT system, conflicts of linguistic constrains may arise. See the following English-Chinese cases:

User Entry	Entries coded in the MT system	Example Sentences or Notes
house= 房子	White House (白宮)	*The official White House site is whitehouse.gov.*
	house (verb)	*The MAP-H21 can house a 2.5-inch hard disk drive.*
to address= 寻址	address (涉及) … issue	*10 issues you must address.*
	address (涉及) … concern	*Studies Address Concern of Mercury-Tainted Fish*
relation=联系	in relation to (关于)	*An overview of molecular forces in relation to protein structure.*
please=请	please is coded as adv and verb	For a specific domain (e.g. user manuals), "please" may never been used as a verb. In this case, the user may not want any interaction with the MT dictionary to avoid incorrect part-speech determination.

User-level clues are provided to allow for different levels of interaction. However, the interaction between a user dictionary and the general MT dictionary is a complex issue. The anticipation of the kind of interaction that a given entry would have with the MT system is impossible. As illustrated in the examples, the interaction may have several forms and each may depend on the way the MT system handles the conflicting expression: simple transfer entries, idiomatic expressions, complex syntactic rules, and even preprocessing normalization (Gerber & Yang 1998). It is neither a goal nor possibility that the user understands or is even aware of the various mechanisms.

The core of IntuitiveCoding® technology, therefore, is to integrate the diverse coding tools into a functional methodology allowing the user to identify unexpected results, and to modify the entries based on impact evaluation.


SYSTRAN Dictionary Manager - [[ENDEDict.sdm]]
File Edit View Tools Window Help

	English	German	Comment	Domain	Category	Confidence
✓	[static section] mismatch	Unstimmigkeit im statischen Teil		General	Noun	🟩🟩
✓	a bite	der Biss		General	Noun	🟩🟩🟩🟩🟩
✓	Libanese (human)	Lebanese		General	Noun	🟩
✓	Lufthansa (company name)	Lufthansa		General	Proper noun	🟩🟩
✓	Martin (first name)	Martin		General	Proper noun	🟩🟩🟩
✓	Network access	Netzwerkzugriff		General	Noun	🟩🟩🟩
✓	Network Adapter	Netzwerkkarte		General	Noun	🟩🟩🟩
✓	Network communication error	Kommunikationsfehler im Netzwerk		General	Noun	🟩🟩🟩
✓	Network Control Panel	Systemsteuerungsoption "Netzwerk"		General	Noun	🟩🟩
✓	Network logon option	Netzwerkanmeldeoption		General	Noun	🟩🟩🟩
✓	Network Protocol	Netzwerkprotokoll		General	Noun	🟩🟩🟩
✓	Network provider	Netzwerk-Dienstanbieter		General	Noun	🟩🟩🟩

2.4 Intuitive Coding Methodology

As described above, Intuitive Coding is a "technology" practiced as a coding methodology, comprised of interactive steps.

1. The user dictionary is automatically converted to the knowledge representation of the MT system, making the runtime user dictionary available from the start.
2. Entries are identified by the Coding Engine as questionable, and are submitted for user review.
3. Translation differences are extracted and reviewed.
4. Depending on possible side effects on sentences, the user can enrich the entries to smooth the integration.

This complete loop is performed using the SYSTRAN Review Manager—preparing the corpus, performing quality analysis, and providing management of review cycle tasks. Figure 2 illustrates the Intuitive Coding process.

3 Evaluation & Applications

3.1 Languages

The IntuitiveCoding® technology is already implemented in the following languages: French, Spanish, Italian, Portuguese, English, German, Dutch, Chinese, Japanese, Greek, and Russian. It is under development for Arabic, Farsi, Korean, Danish, Swedish, and Finnish.

3.2 Evaluation Criteria

The following criteria are used to evaluate the Intuitive Coding system:

a. Speed of the system: Starting from first raw N thousand entry glossary and time required for building the first runtime dictionary
b. Initial Recall: the ratio of entries in the first runtime dictionary that require enrichment
c. Post-processing Precision: the ratio of entries identified as “problematic” during the corpus quality assessment review
d. Final Accuracy: the ratio of “non-codable” entries

The above evaluation is both language and domain-dependent, but also contingent upon the

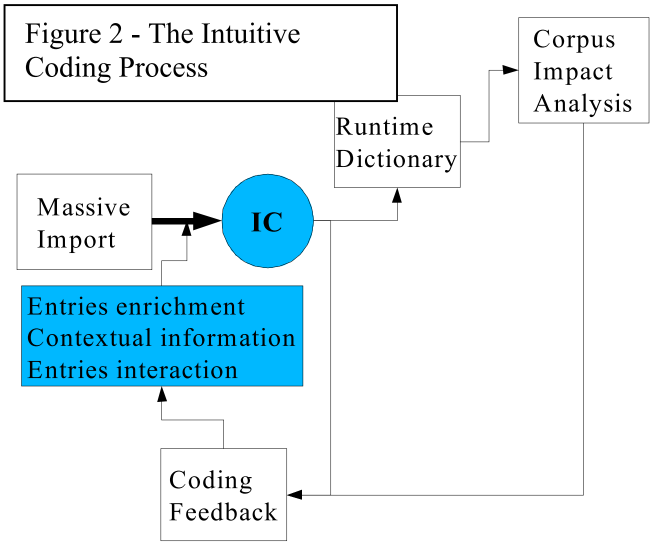


Figure 2 - The Intuitive Coding Process

quality of the initial user glossary. Furthermore, for certain specialized applications, such as the chemical glossary, the first step is to customize the coding rules (e.g., customized context-free compound rules) since the grammar of the expression is domain-specific, and the confidence ratio needs to be re-evaluated for this purpose.

Typical values used in customer dictionary coding are:

a. 10 entries per second on a 600MHz computer
b. 7-12% of entries receive feedback from the IC engine and should be reviewed and correcte
c. Number of problematic entries: 5%
d. Less than 1% comprise the non-codable entries, due to grammatical features not covered by a generic set of advanced coding clues.

3.3 Applications and associated tools

The IntuitiveCoding® technology was developed in the SYSTRAN Dictionary Manager, which provides users with state-of-the-art terminology management capabilities. The tool guides the user through the process of adding their own terms and expressions to the user dictionary, allows the user to import and export Text and Excel file-formatted glossaries, and create user domains for greater term specification.

The coding process is also integrated in a comprehensive review process based on the "SYSTRAN Review Manager". This tool allows users to define review tasks for a large corpus, compute quantitative evaluation against criteria defined by the administrator, and provides a side-by-side concordance tool for the extraction and evaluation of general translation side-effects.

4 Conclusion

The concept of Intuitive Coding is not dependent upon a given MT system. The main goals are to exploit implicit linguistic information present in a raw dictionary, formalize missing information via feedback to the user, dispatch the coding results to defined internal modules, and provide real-time interaction based on MT output. IntuitiveCoding® technology powers a user dictionary management tool that enables straightforward coding and integration of production-scale user dictionaries. This empowers the user to perform MT customization, without the intense involvement of the MT system developers. The next steps include a) terminology extraction; and b) terminology management.

After integration of the initial user glossary and the addition of missing terminology, the next step in the customization process is to code extracted multi-word terms. SYSTRAN's Terminology Extraction tool, currently under development, is also based on the IntuitiveCoding® technology. The differences between its "coding user-dictionary" and "extracting terminology lie in: a) The entries in the user dictionary are expected to be in lemmatized forms, where terminology extraction deals with inflected forms in the corpus. b) The user dictionary entries are expected to be valid grammatical units, where terminology extraction does not know the expression's priority boundary.

The combination of the IntuitiveCoding® engine and the methodical approach of feedback and corpus validation coding, facilitates the intelligent import of the user dictionary by a non-MT expert. This technology is used in commercial large-scale MT customization projects.

Acknowledgments

We would like to thank Nick Mote for editing the paper, and other anonymous reviewers.

References

Gerber, Laurie & Yang, Jin: SYSTRAN MT Dictionary Development. Machine Translation: Past, Present and Future. In: Proceedings of Machine Translation Summit VI. October 29 – November 1997. San Diego, CA, USA (1997) 211-218.

Senellart, Jean, Plitt, Mirko, Bailly, Christophe: Resource Alignment and Implicit Transfer. In: Proceedings of the Machine Translation Summit VIII. Santiago de Compostela, Galicia, Spain. (2001)

Senellart, Jean, Dienes, Peter, Varadi, Tamas: New Generation of Systran Translation System. In: Proceedings of the Machine Translation Summit VIII. Santiago de Compostela, Galicia, Spain. (2001)

Rychtyckyj, Nestor: An Assessment of Machine Translation for Vehicle Assembly Process Planning at Ford Motor Company. Machine Translation: From Research to Real Users. In: Proceedings of 5h Conference of the Association for Machine Translation in the Americas, October 2002, Tiburon, CA, USA (2002). 207-215

Yang, Jin & Lange, Elke. SYSTRAN on AltaVista: A User Study on Real-Time Machine Translation on the Internet. Machine Translation and the Information Soup. In: Proceedings of Third Conference of the Association for Machine Translation in the Americas, AMTA ’98. Langhorne, PA, USA. (1998). 275-285.