SYLVIANE CARDEY
Centre Tesnière, Université de Franche-Comté
Besançon, France
sylviane.cardey@univ-fcomte.fr

Machine Translation of Controlled Languages for More Reliable Human Communication in Safety Critical Applications[1]

1. Introduction

The research results presented in this paper concern linguistics and controlled language machine translation, an application within the LiSe (Linguistique et Sécurité) project[2] [1], this in the context of security and crises in general, and in particular where communication ought to be rapid and correct [2].

2. Controlled Language Machine Translation

In an emergency or crisis, not only have alert messages to be controlled at source [3, 4], but we do not necessarily know the messages that must be translated beforehand. As well as this, the translated messages too have to be controlled for the target language audiences. However any machine translation must be accomplished with neither manual pre- nor post-edition as both are unacceptable in an emergency due to time constraints. In the LiSe project, controlled French has been the source controlled language (SCL) and for the target controlled languages (TCLs) these are controlled French (identity – exploited for testing), Arabic, Chinese, English (abbreviation An) and Thai; see Figs. 1 and 2.

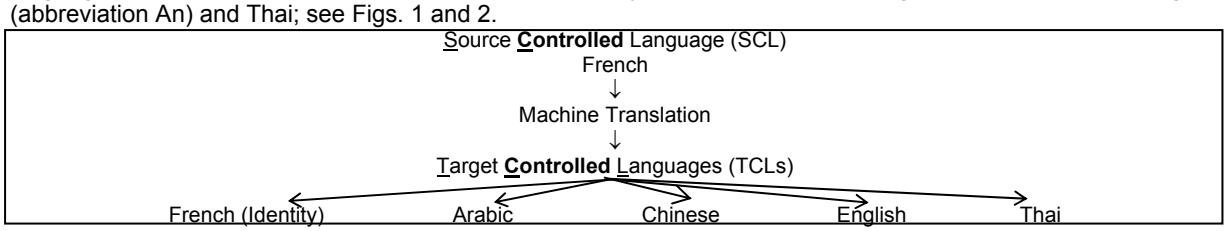


Figure 1. Machine translation of controlled languages

Corpus source: http://www.interieur.gouv.fr/misill/sections/a_votre_service/votre_securite/conseils-incendie/incendie-chez-vous/view					
UncontrolledText	**SCL**	**ControlledText**	**SCL = TCL?**	**TCL**	**Translation**
Ne branchez pas trop d'appareils sur la même prise	French	Ne pas brancher trop d'appareils sur la même prise	SCL = TCL	French	Ne pas brancher trop d'appareils sur la même prise
			SCL ≠ TCL	Arabic	لا توصلوا الكثير من الأجهزة بنفس منشب التيار
				Chinese	不要把太多的电器插在同一个插销
				English	Do not connect too many electrical appliances to the same plug
				Thai	อย่าเสียบปลั๊กเครื่องใช้ไฟฟ้าจำนวนมากเกินไปบนที่เสียบอันเดียวกัน

Figure 2. Example illustrating Translation

3 Machine Translation Architecture

In our architecture, what is new is that both the source language SL (French) and the target language(s) TL(s) are controlled, not only to conform to normal controlled language constraints, but also mutually for translation knowing that each and all influence the others [5]. Pre-edition is avoided as well, as control of the source is provided during message entry by means of the 'SL to CPSL' User Interface [1]. The TLs being controlled, this obviates post-edition. All this has resulted in a novel hybrid (pivot + transfer) rule-based machine translation architecture in which the pivot language PL is French controlled also for translation (**C**ontrolled **P**ivot **S**ource **L**anguage CPSL), and where the Transfer System is directed by the various source-target language divergences. In the case of French, we have French → French: identity (Ø divergences). To situate our architecture and its scaleability, Fig. 3 illustrates how our own "Controlled Pivot Source Language + Transfer" architecture is related to conventional pivot and transfer architectures.

[1] Keynote speech in this Symposium.
[2] The LiSe project was funded by the French Agence Nationale de la Recherche (ANR) programme: Concepts Systèmes et Outils pour la Sécurité Globale, contract n° ANR-06-SECU-007-02.

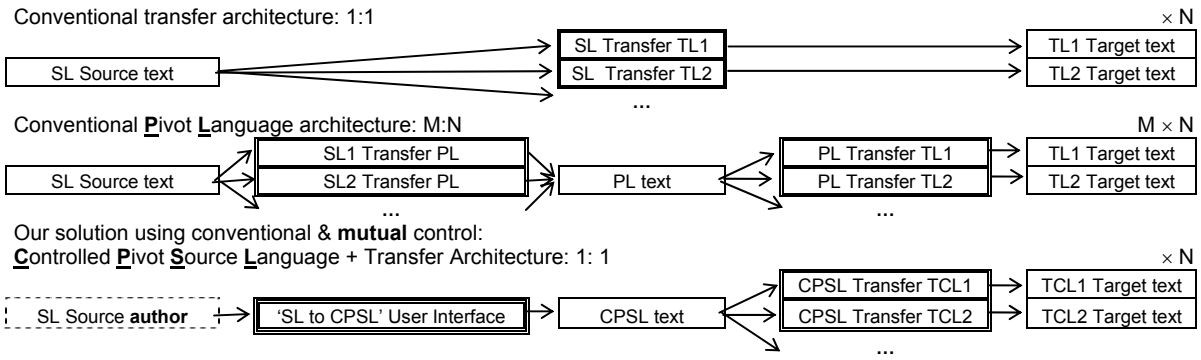


Figure 3. MT architectures, scaleability and applicability

4. Author, Pivot Language & Transfer

The author composes messages syntagmatically in the controlled French pivot language using the 'SL to CPSL' User Interface. For example "*Ne pas brancher trop d'appareils électriques sur la même prise.* " is expressed in the CPSL by the structure:

```
RegroupageEnSyntagms_LS =      [
        ['Ne',                                  neg1],
        ['pas',                                 neg2],
        [brancher,                              vinf],
        ['trop d'appareils électriques',        arg1],
        [sur,                                   prep_v],
        ['la même prise',                       arg2],
        [.,                                     pt]]
```

This structure is passed to the Transfer System for translation into each target language and the mutual divergences with the controlled French direct the translation process. Transfer divergences can occur at variously the sentence, syntagmatic and lexical levels. For example for the first two (these include the lexical level):

– Sentence Level

  French:

  opt(neg1) + opt(neg2) + vinf + arg1 + prep_v + arg2 + pt

  Ne pas brancher trop d'appareils électriques sur la même prise.

  Chinese:

  opt(arg0) + opt(neg1) + indicateur(['把','让']) + arg1 + v + arg2

  不要把太多的电器插在同一个插销

– Syntagmatic Level

  `ad + adj + n = adj + cl + n`

  `la même prise` = `同一个插销`

  - The French `ad` has no corresponding lexis in Chinese.
  - The Chinese `cl` has no corresponding lexis in French.
  - The French `adj` corresponds to the Chinese adj (linked via the dictionnaireLexical).
  - The French n corresponds to the Chinese n (linked via the dictionnaireLexical).

Transfer is effectuated for each source-target language couple as follows:

i. Source language sentence level selection is determined by the 'SL to CPSL' User Interface which subsequently enables each unique corresponding target language sentence level structure to be selected (= in the case of French).

ii. Lexical analysis of each syntagm in the controlled French source sentence (the suite of syntagms delivered by the 'SL to CPSL' User Interface) is performed. This results not only in the construction of the lexical transfer table for each SCL/TCL couple but also the selection of the relevant syntagm transfer structure couple.

iii. Transfer is effectuated by the target language as requestor and **not** the source language as provider. This ensures that no linguistic rule placement coding is necessary where:
   - A source language (French) source lexis has no corresponding lexis in the target language
   - A target language lexis has no corresponding lexis in the French source.

Finally target surface level post transfer operations are effected: e.g. syntagm lexies (government) agreement (e.g. Arabic adjectives with nouns) and clitics.

5. Micro-Systemic Linguistic Analysis of Machine Translation

We now turn to the micro-systemic linguistic analysis that we have carried out [6, 7]. For each (∀) linguistic phenomenon observed when translating from the Source Controlled Language to each Target Controlled Language we observe (∃):

– a single 'canonical' case where there are no language divergences (identical source and target controlled language phenomena)

– ‘variant’ cases encompassing the divergences between each target controlled language and the source controlled language

The whole language identity case (variant = canonical) applies to all linguistic phenomena observed when the Source Controlled Language ≡ Target Controlled Language. For an observed linguistic phenomenon when translating, in classifying the variant cases, the linguist establishes two categorisations in the form of two partitions and then puts these into relation, one with the other. The categorisations are:

– 'non-contextual' (nc) categorisation of the canonical forms **C** in relation with the variant forms **V** in isolation, the context being limited to just the canonical and variant forms themselves. This results in the Partition $\mathbf{P_{nc}}$.
– 'in-context' (ic) categorisation of the canonical forms in relation with the variant forms in terms of the linguistic contexts of the variant forms. The systemic analysis reveals precisely what other related linguistic systems are involved. This results in the Partition $\mathbf{P_{ic}}$.

Given that we have partitions ($\mathbf{P}_{nc}$ and $\mathbf{P}_{ic}$), from the fundamental theorem on equivalence relations, it follows that there exist two corresponding equivalence relations $\mathbf{E}_{nc}$ and $\mathbf{E}_{ic}$, both over the binary ordered relation between the canonical forms and the variant forms **CV**. We model the system over the linguistic phenomenon by means of the binary ordered relation **Ss** between the equivalence relations $\mathbf{E}_{nc}$ and $\mathbf{E}_{ic}$, each over **CV**. From this relation **Ss** we can generate algorithms.

6. Micro-systemic Translation Architecture

Applying micro-systemic linguistic analysis, we have classified and organised the equivalences and divergences in the form of a compositional micro-system structure. In this structure the resulting micro-systems are expressed in a declarative manner by means of typed container data structures in the form of a database together with their contents so as to be incorporated in the machine translation process; see Fig. 4 [8].

Super µSystem	Explanation
Ss_TACT	Root µsystem – **T**raduction **A**utomatique **C**entre **T**esnière
Ss_frC	Controlled French
Ss_frC_frC	Controlled French → Controlled French (**Identity – no divergences**, constructed from **Ss**_frC)
Ss_frC_arC	Controlled French → Controlled Arabic
…	…

Figure 4. Machine Translation Super Micro-Systems

In Figs. 5 and 6 we show the typed container data structures corresponding to **Ss**_frC_frC and **Ss**_frC_arC; note the presence of example & corpus attributes in the data base. For **Ss**_frC_arC we mark with a grey background the divergences with **Ss**_frC_frC (attribute and/or content); for example, unlike French, Arabic has a case system.

Ss_frC_frC:

table
frC
frC_groupesVerbaux_frC
frC_args_frC
frC_groupes
frC_dictionnaireLexical_frC
frC_catégories_frC
frC_catégories

frC_groupesVerbaux_frC					
	verbe_frC	**verbe_frC**	**groupe_frC**	**exemple**	**corpus**
frC_args_frC					
	arg_frC		**arg_frC**	**exemple**	**corpus**
frC_groupes					
	groupe_frC		**structure_frC**	**exemple**	**corpus**
frC_dictionnaireLexical_frC					
	lexique_frC	**catégorie_frC_frC**	**lexique_frC**	**catégorie_frC**	**exemple** / **corpus**
frC_catégories_frC					
	catégorie_frC_frC		**exemple**	**corpus**	
frC_catégories					
	categoryie_frC		**exemple**	**corpus**	

Figure 5. Typed container data structure corresponding to Ss_frC_frC

Ss_frC_arC:

Table
arC
arC_groupesVerbaux_frC
…

arC_groupesVerbaux_frC					
	verbe_arC	**verbe_arC**	**groupe_arC**	**exemple**	**corpus**
arC_args_frC					
	arg_frC		**arg_arC**	**exemple**	**corpus**
arC_groupes					
	groupe_arC		**structure_arC**	**exemple**	**corpus**
arC_dictionnaireLexical _frC					
	lexique_frC	**catégorie_frC_arC**	**lexique_arC**	**noLexique_arC**	**accLexique_arC** / **…** / **tdefLexique_arC** / **catégorie_arC…**
arC_catégories_frC					
	catégorie_frC_arC		**exemple**	**corpus**	
arC_catégories					
	categoryie_arC		**exemple**	**corpus**	

Figure 6. Typed container data structure corresponding to Ss_frC_arC

7. Micro-systemic Translation Programming

For ergonomic linguistic programming reasons, the concrete (content containing) form of the typed container data structures is realised by means of spread-sheets upon which is mapped a typed & interpretable data structure

representing the Machine Translation micro-system. For example at the spread-sheet level: worksheet & column names' languages (other than for **Ss**_TACT) are formulated as follows (in BNF):

LS = Langue Source (SCL), LC = Langue Cible (TCL)
Full worksheet name :: LS | LC | LC_worksheetname | LC_worksheetname_LS
Full column name :: columnname | columnname_LS | columnname_LC | columnname_LS_LC

(Worksheet names & column names cannot contain '_'). At the cell level we have:

cell_types([atom, plus_to_list, term, atom_list]).

Fig. 7 summarises the target controlled language divergences with controlled French.

Sentential syntagmatic order	Transfer lexis POS ambiguity	Target clitic:
Syntagm lexical order		prefixing
Category (Part of Speech)	Lexis:	postfixing
Super Categories (POS, Case…)	**Source** / **Target**	infixing
Syntagm lexies agreement	yes / yes	
Case (declension)	yes / no	
Classifier	no / yes	
Indicator		

Figure 7. Target Language Divergence with Controlled French

In Fig. 8 we give examples of coding for some of these divergences.

Target Language Divergence with Controlled French		**Controlled Language**					**Example coding**
		frC	anC	arC	chC	thC	
Category (Part of Speech)		Ø	yes	yes	yes	yes	frC: | même | **adjs** |, arC: | même | **det** |
Super Categories (POS, Case…)		Ø	yes	yes	yes	yes	arC: **n** = [nms, nmp, nfs, nfp, nmph, nmpnh, nfph, nfpnh] arC: **cases** = [no, acc, t, nodef, accdef]
Transfer lexis POS ambiguity		Ø	yes	yes	yes	yes	thC: frC: adj**1** + nmp + adj**2** thC: n + adj**2** + adj**1**
Lexis							
Source	**Target**						
no	yes	Ø	yes	yes	yes	yes	arC: arC: artu_+ n_A(acc)+lexis_('ال')+ adj_A(acc)

Remark: the Lexis example also illustrates Case (declension) and Target clitics (artu_ and lexis_) – both prefixing clitics.

Figure 8. Examples of coding for some of the controlled language divergences with controlled French

Our implementation involves a single target language independent kernel in which the Target Language divergences with Controlled French are abstracted as propositions. For example: 'Case (declension)':

Database (spread-sheet cell content):

arC_catégories: cases = [no, acc, t, nodef, accdef]
anC_catégories: cases = []

Kernel code:

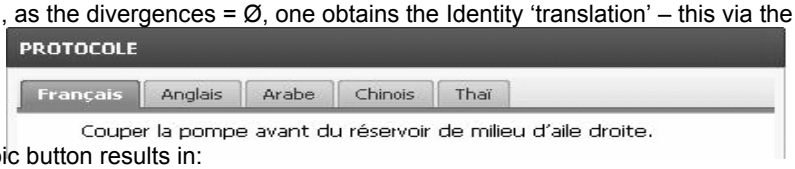
Cases_TL = [_|_] ⇒ Target Language with case system (e.g. LC = arC Arabic)
Cases_TL = [] ⇒ Target Language without case system (e.g. LC = anC English)

8. Problems put into evidence and solved by our Controlled Language Machine Translation system

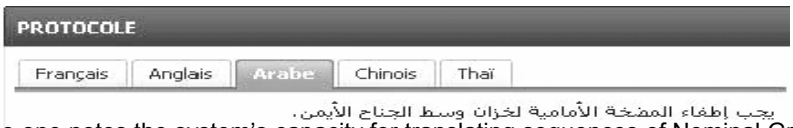
We take the case of the machine translation of French → Arabic [9] with an example from the domain of aeronautics:

*Couper la pompe avant **du** réservoir **de** milieu **d**'aile droite.*

This sentence, in French, is entered into the 'SL to CPSL' User Interface. From the target language buttons, if French is selected, as the divergences = Ø, one obtains the Identity 'translation' – this via the Transfer System:



Selecting the Arabic button results in:



- In this example one notes the system's capacity for translating sequences of Nominal Groups nested with the preposition '**de**'. This type of sequence of NGs requires an in-depth analysis in respect of translation to Arabic for several reasons.
- There are two types of determination in Arabic; one being with the definite article and the other being a particular determination involving annexation.
- As well as this, Arabic has a morpho-syntactic characteristic linked to its inflexional morphology and to its case marking which concerns nouns and adjectives (Case (declension)).
- Syntactic linking has also to be taken into account because it is necessary to make the agreement between nouns and adjectives (Syntagm lexies agreement).
- To these problems must be added the cliticisation of certain parts of speech in Arabic (Target clitic).

9. Tracing

The controlled language machine translation system includes logged dynamic tracing:

i. Input JSON (interface language (www.json.org) with the 'SL to CPSL' User Interface) text logged in pretty print format for subsequent automated benchmark/regression testing

ii. For each target language:
   a. Prepare for transfer :
      A. RegroupageEnSyntagms_LS = value
      B. LC_GroupeVerbal_LS = value
      C. Regroupage_En_Arguments_LS = value
      D. Unites_Source = value
      E. Do transfer (driven by Target Language)
   b. Regroupage_En_Arguments_LC = value
   c. Post transfer:
      A. government_agreement (if applicable)
      B. Unites_Cible = value

iii. Traductions = value

iv. JSON to return to the 'SL to CPSL' User Interface - Translations

10. Implementation Process

To situate and relate our project in terms of the real world of application engineering we terminate with the implementation process devised in conjunction with the different project actors and which is summarized in Fig. 9.

Project Steps ⇒			
µ-systemic linguistic analysis of: Controlling & Translation	**Testing**	**µ-systemic modeling**	**Application**
Corpora ↓ Source (→controlled) language ↕ Translation Analyses ↕ Target (→controlled) languages ↑ Corpora	Domain experts ↕ (→User interface) ↕ Source controlled language ↕ Translation Analyses (→ norms & divergences + benchmarks) ↕ Target controlled languages ↕ Domain experts	Norms & divergences ↓ Interrelated µ-systems of canonical & variant forms ↓ Container structures & formalization languages ↓ MT system: Database schema + Language independent compiler & transfer kernel + Linguistic data	Authors ↕ User Interface ↕ Source controlled language ⇓ MT system ⇓ Target controlled languages ⇓ End-users
⇐ Regression Testing			

Figure 9. Controlled language Machine Translation implementation process

11. Conclusion

We have explained our novel hybrid rule-based machine translation architecture which involves as pivot language the source language controlled also for translation, and in which the transfer system encompasses source-target language divergences. We have related this to the real world of application engineering by terminating with the implementation process. We stress that the micro-systemic architectures for the controlled language machine translation have been devised by linguists, and that micro-systemic linguistic analyses are exhaustive and provide traceability, essential qualities in safety critical applications. In like manner the linguistic programming depends on concepts, formalisations and specialised coding languages conceived by linguists which together result in table driven declarative programming. Language is innately human and could well play the key role in the management of crises.

Acknowledgements

We wish to thank Peter Greenfield and all those who have actively participated in the project: our LiSe project partners Airbus (Laurent Spaggiari), SERF (Dominique Vuitton), Ciprian Melian of Melian Multimédia and the researchers of Centre Tesnière: Raksi Anantalapochai, Mohand Beddar, Dilber DeVitre, Gan Jin, Kyoko Kuroda, Julie Renahy and Xiaohong Wu.

References

[1] S. Cardey et al., *Le projet LiSe « Linguistique, normes, traitement automatique des langues et sécurité : du data et sense mining aux langues controlées*, in actes du WISG 2010, Workshop Interdisciplinaire sur la Sécurité Globale, Université de Technologie de Troyes, 26 & 27 Janvier 2010, 10 pages, CDROM

[2] J. Hutchins. *Has machine translation improved? some historical comparisons* MT Summit IX: Proceedings of the Ninth Machine Translation Summit, New Orleans, USA, September 23-27, 2003. [East Stroudsburg, PA: AMTA], pp. 181-188.

[3] E. Gavieiro-Villatte, L. Spaggiari. *Demonstration of the open ended overview of controlled language*, Proceedings LREC, Athens, July 2000, pp. 1133-1134.

[4] S. Cardey, *Controlled Languages for More Reliable Human Communication in Safety Critical Domains*, Proceedings of the 11th International Symposium on Social Communication, Santiago de Cuba, Cuba, 19-23 January 2009, ISBN: 978-959-7174-14-119-23, pp. 330-335.
[5] S. Cardey et al.**,** *Modèle pour une Traduction Automatique fidèle, le système TACTmultilingue, Projet LiSE (Linguistique et Sécurité)*, in actes du WISG'09, Workshop Interdisciplinaire sur la Sécurité Globale, Université de Technologie de Troyes, 28 & 29 Janvier 2009, 10 pages, CDROM
[6] S. Cardey, P. Greenfield, *Systemic Linguistics with Applications*, in Linguistics in the Twenty First Century, Cambridge Scholars Press, United Kingdom, ISBN 1904303862, pp. 261-271, 2006.
[7] S. Cardey, P. Greenfield, *A Core Model of Systemic Linguistic Analysis,* Proceedings of RANLP-2005, Borovets, Bulgaria, 21-23 September 2005, pp. 134-138.
[8] S. Cardey, P. Greenfield, R. Anantalapochai, M. Beddar, D. DeVitre, G. Jin, *Modelling of Multiple Target Machine Translation of Controlled Languages Based on Language Norms and Divergences*, Proceedings of ISUC2008, Osaka, Japan, December 15-16, 2008, Proceedings published by the IEEE Computer Society, ISBN 978-0-7695-3433-6, pp 322-329.
[9] M. Beddar, *French to Arabic Machine Translation: Isomorphic Syntax, Use of Terminal S*equences, Proceedings of ISMTCL, Besançon, July 1-3, 2009, International Review Bulag, PUFC, ISSN 0758 6787, ISBN 978-2-84867-261-8, 2009, pp. 38-42.