

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

### The SyGULAC Theory<sup>1</sup>

#### 1. Introduction

In this paper we show why and how we use the SyGULAC theory (Systemic Grammar Using a Linguistically motivated Algebra and Calculus), which has its roots and its calculability in a systemic approach and indeed a micro-systemic approach, and in discrete mathematics, this with the view not only for the analysis of languages, but also for their generation. From this basis, we present two examples of methodologies which have been developed and refined, and which have been applied to three examples: controlled languages, machine translation and data-sense mining, all three in security critical domains which means that the results must be reliable.

#### 2. The theoretical point of view

##### 2.1 Syntactic analysis

We discuss here only the two analysis methods (and their developments) which are in most current use and then compare them with our own.

Knowing that the first step consists in performing a discourse category analysis, one can then add to this analysis, 1) an immediate constituent analysis (Chomsky) which has the intention of rendering visible the functions relating words with each other, and 2) another analysis can be equally used and which is based on valency theory (Tesnière), a verb's valency being the number of complements necessary for constructing a simple and complete utterance, these complements being able to be verbal or sentence complements. One can thus classify a language's verbs in terms of their valencies.

At first sight these two theories (Chomsky and Tesnière) seem to be different. However, what is interesting is the easy transfer from one to the other as shown in Figure 1.

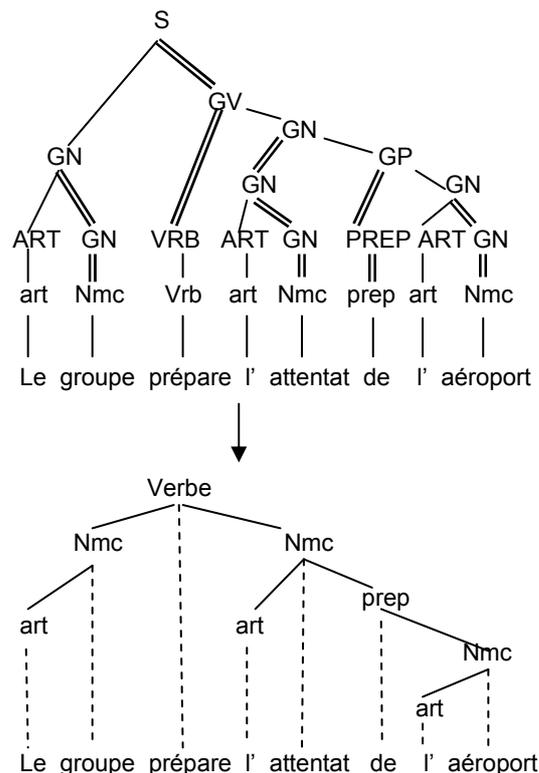


FIGURE 1. Chomsky type schema with double arcs, which if suppressed, leads to Tesnière

Generally speaking, there would seem to be two ways of representing an utterance's cohesion when trying to represent the thought, or the communicative act. The former consists in describing the sentence as a nested structure of its constituents (in generative terms, the sentence being generated from a unique symbol by means of

<sup>1</sup> Keynote speech in this 13<sup>th</sup> International Symposium on Social Communication.

rules that are independent of the words and morphemes making up the sentence), and the latter consists in showing that there is a sort of attraction that the lexical elements exercise on each other (this idea following from Tesnière's stemma). Neither of the two ways involve semantics, or at most very little.

## 2.2 *Semantic analysis*

Semantics, which is concerned with sense in general, is present at different levels. For example, Tesnière said that one finds semantics at the level of links. In fact semantics has only really been studied very recently and especially concerning lexis (homonymy, polysemy). However, syntax, inflexions and word order are as much elements which are involved in the sense as are other phenomena such as anaphora, dislocation, and still others too.

## 2.3 *The SyGULAC theory*

Contrary to what has just been said, the SyGULAC theory (Systemic Grammar Using a Linguistically motivated Algebra and Calculus), based on systemic linguistic analysis [1,2], both developed in Centre Tesnière, does not have as goal describing the whole of a language by means of some global representation of the different 'layers': lexis, syntax, morphology, semantics separately. Rather, SyGULAC advocates firstly delimiting the problem or the analyses' needs concerning some specific application. According to the needs, a specific system is constructed that resolves and represents the problem. This system can be manipulated and represented, which is contrary to what can be done for a language in its totality which can neither be delimited nor manipulated. Thus only the necessary elements, be these lexical, morphological, syntactic are represented in the single system, this latter being able to be related to other such systems. We do not even mention semantics because, finally, this is the only 'layer' that interests us; whatever the operations one does, in reality it is to be able to access the sense. Thus we do not need a complete description of the language, or languages, concerning their lexis or their morphology as habitually one tries to do. This allows us to resolve problems thanks to analyses which are much lighter in quantity and in time spent.

## 3 **Methodologies**

In the Introduction we said that two methodologies have been developed and refined for language analysis and generation. The same formalism can be used for one or the other.

### 3.1 *Analysis*

Analysis is seen here from the point of view of extracting sense whether for machine translation or for data-sense mining. The methodology, whether it be applied to one or the other, is the same. The description model uses the same rules format. In fact, the methodology in general can be used for diverse applications.

We see below (Figure 2) two rules, one for machine translation and the other for data-sense mining; the formal representation is the same for both.

Machine Translation:

opt(neg1) + lexis('يحب') + opt(neg2) + nver + arg1(acc) + opt(opt(precomp1),comp1(n)) + opt(opt(precomp2),comp2(n)) + pt

Data-sense Mining:

l(d) + '['( ) + 从 + l(chiffres) + 到 / 至 + l(chiffres) + l(temps) + ']'( ) + l(f)

FIGURE 2. Example of rules for machine translation and data-sense mining

### 3.2 *Generation*

For us generation concerns for example utterances output by machine translation systems. We briefly note here the sometimes poor results provided by those on-line machine translations systems which function by means of transfer for the best known and which pass via English when one wants to go for example from French to Arabic, Chinese or Russian. Translation memory systems exist too, but their quality is only adequate for repetitive source utterances due to the method which consists in finding translations of sentence segments which have already been translated [3]. However, in the context of security, amongst others, it is impossible to take the risk of not finding the searched for segment even if one has recourse to a machine translation system of variable quality in the event of failure.

## 4 **Our methodology for machine translation**

Our methodology does not require pre-edition in the conventional sense of the term. We use a controlled language [4, 5] which at the outset serves for giving a good interpretation of the information in suppressing ambiguities and all that can harm the information's clarity. However, in practice we soon found that even after controlling the source language, the results were far from what we had hoped. We therefore also controlled the target language. This means that a very fine level comparative analysis of target languages with French has been undertaken. We have thus been able to extract similar mega- and micro-structures not only for French but also for target languages and between target languages too. We have built our translation system based on these resemblances [6], each divergence being subsequently handled at the transfer level specific to each language [7, 8].

The following example shows how and why this control is necessary. Take the following non-controlled sentence:

*Refroidir immédiatement la brûlure, en l'arrosant avec de l'eau froide durant 5 minutes.*

After controlling we obtain:

*Verser de l'eau froide sur la brûlure immédiatement durant 5 minutes.*

The reasons for the control are as follows:

- The sentence contains two distinct pieces of information; one is injunctive *arroser* and one is explicative *refroidir*. It seems more logical to start by stating the action to be effected and only afterwards provide the motive(s) for this action. Furthermore, to ensure understanding and also ease of translation of the sentence, the controlled language imposes one and only one verb per sentence. This prohibition applies also to using the gerundive *arrosant*.
- Controlling is necessary for the three target languages, Arabic, Chinese and English, because of the verb *arroser*. These three languages use this verb only when it is followed by an argument of vegetable type.
- So as to avoid any error in the identification of a pronoun's antecedent, pronouns are prohibited in the controlled language.

Controlling solves numerous linguistic problems. Nevertheless, some of them remain when the controlled sentences are translated by commercially available machine translation systems.

We give below examples of translations made by machine translation systems other than our own.

We start with the controlled sentence:

*Verser de l'eau froide sur la brûlure immédiatement durant 5 minutes.*

For the translation to Chinese with 'reverso' we obtain:

在 5 分钟期间在烧伤上立刻倒冷水 (préposition在 5min pendant brûlure maintenant verser froide eau)

The problem concerning Chinese is at the structural level and also the lexical inexactitude of *pendant*. “在 ...期间” can only be used for a duration which is much longer than “minute”, such as “année” (year) for example. As far as the error arising from *brûlure*, in this context one would rather use *blessure* (wound) in Chinese.

For the translation to Arabic with 'reverso':

الدفع (صب) بعض الماء البارد على أن تحترق بعد 5 دقائق (poussée (verser) quelque eau froide pourvu que tu brûles après 5 minutes)

Finally, to English, still with 'reverso':

Cool at once the burn, by spraying him(it) with some cold water for 5 minutes.

There are two problems with the English: the position of the complement *at once* which is understandable but not standard, and the problem with the pronoun.

The results produced by our own machine translation system are as follows:

- Chinese LiSe:  
立刻在伤口上浇冷水5分钟,
- Arabic LiSe:  
يجب صب الماء البارد على الحرق فوراً لمدة 5 دقائق .
- English LiSe:  
Pour cold water on the burn immediately for 5 minutes.

## 5. Our methodology for data-sense mining

As opposed to data-mining which consists in searching for words in a sentence, indeed in a text, in order to extract if possible important points in the text, we work at the level of sense in general, that is to say, all the elements (morphemes (inflexions both syntactic and derivational (lexical), simple and compound lexis, etc.) and their organisation and distribution in the sentence or in the lexis for the morphemes. In this application domain, there are currently two major methodologies, one based on keywords and the other on statistics, which itself uses keywords, so both use only a part of the lexis and not the lexis in its totality. Thus lists of 'non-important' words are created (also called 'empty' words), so that only those words (terms) deemed 'important' or 'keywords' are recognised. The problem is, “What is an important word?”. The principle question is “What is a word?”. Take the example *Ce produit était parfait*, (This product was perfect) which has as understatement *il ne l'est plus* (it is not any more). So, if we use just *parfait* and *produit*, we get a bad interpretation. It must be added that these two methodologies require the systems to be trained and/or pre-edition to be effected, and urgency renders these impossible in the event of a crisis due to the lack of time.

Our methodology [9], data and sense-mining, not only uses the lexis in its totality, principally its morphology but also and above all syntax and of course semantics and their intersections morpho-syntax, lexico-syntactico-semantics, etc., represented by rules and sets of structures, and systems functioning in interrelation. The methodology interprets texts which contain no so-called keyword and it analyses the whole of the text.

We give below an example of the method for sense (sème) extraction using a rule structure. When the sème is found in the text, it is compared with current knowledge that one has of the same sème.

- Chinese example:  
Relation\_Seme\_Name: 年代,année  
Text Before: ""  
Matched Text: " 从80到90年代 , "

Text After:

"该导弹是前苏联地该导弹是前苏联地对地导弹。劳动是朝鲜研制的中程弹道导弹“飞毛腿C”  
 射程 ) 500射程约 , 的改良型 ( 公里1300。公里"

Compared information: [从,80,到,90,年代]

Structure used: 1. I(d) + 'I( ) + 从 + I(chiffres) + 到 / 至 + I(chiffres) + I(temps) + 'I( ) + I(f)

With the same structure, one can also find the following Chinese sentences:

- 从60到90年代 ,
- 从60至80年代 ,
- 从六十至八十年 ,
- etc.

– Arabic example:

The first information found:

المجموعة حضرت الاعتداء في لبنان → the group prepared the attack in the Lebanon

The structure that has allowed us to find this information:

opt(Part. interrog) + opt(Part. nég) + opt(Part.prob) + Primat + opt(Part.prob ) + Prédicat +  
 opt(Pronom.connect) + opt(Cod1>>Cod2>>Cod3) + opt(Prép) + opt(Coi) + opt(opt(Prép) + Ccirt) +  
 opt(opt(Prép) + Ccircl) + opt(opt(Prép) + Cman)

With the same structure, we can find other types of information:

Information 2 found:

العصابة سرقت المخزن بالأمس → the bandits stole from the shop yesterday

Information 3 found:

المجرم قتل الشرطي في العاصمة → the criminal killed the police officer in the capital

**6. Conclusion**

To conclude, we can say that it would have been impossible to have obtained such reliable results and across several languages without the support of the SyGULAC theory and the various methodologies which share and respect the same formal model.

Finally we show in Figure 3 neology processing that no other current system can achieve, our example being the first two lines from Lewis Carroll's *Jabberwocky* [10] ("*Twas*" has been normalised to "it was").

*'It was brillig, and the slithy toves did gyre and gimble in the wabe;'*

Lexical unit	Out-of-context Categories	In-context Category
it	{PROpers}	PROpers
was	{V}	V
<b>brillig</b>	{ADJ}	ADJ
,	{PUNCT}	PUNCT
and	{CONJ}	CONJ
the	{ADV, DET}	DET
<b>slithy</b>	{ADJ}	ADJ
<b>toves</b>	{Nplu, V3sing}	Nplu
did	{Aux}	Aux
<b>gyre</b>	{V}	V
and	{CONJ}	CONJ
<b>gimble</b>	{V}	V
in	{ADV, ADJ, PREP}	PREP
the	{ADV, DET}	DET
<b>wabe</b>	{N}	N
;	{PUNCT}	PUNCT

Figure 3. Results of the disambiguated tagging of the first two lines of *Jabberwocky*

All the **unknown words** have been automatically labelled by our system.

**References**

[1] S. Cardey, P. Greenfield, *Systemic Linguistics with Applications*, in *Linguistics in the Twenty First Century*, Cambridge Scholars Press, United Kingdom, 2006, ISBN 1904303862, pp. 261-271.  
 [2] S. Cardey, P. Greenfield, *A Core Model of Systemic Linguistic Analysis*, Proceedings of RANLP-2005, Borovets, Bulgaria, 21-23 September 2005, pp. 134-138.

- [3] 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.
- [4] E. Gavieiro-Villatte, L. Spaggiari. *Demonstration of the open ended overview of controlled language*, Proceedings LREC, Athens, July 2000, pp. 1133-1134.
- [5] S. Cardey, *Controlled Languages for More Reliable Human Communication in Safety Critical Domains*, Proceedings of the 11th International Symposium on Social Communication, Centre for Applied Linguistics, Santiago de Cuba, Cuba, 19-23 January 2009, ISBN:978-959-7174-14-119-23, pp. 330-335.
- [6] S. Cardey et al., *Le projet LiSe « Linguistique, normes, traitement automatique des langues et sécurité : du data et sense mining aux langues contrôlées*, in actes du WISG 2010, Workshop Interdisciplinaire sur la Sécurité Globale, Université de Technologie de Troyes, 26 & 27 Janvier 2010, 10 pages, CDROM
- [7] 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.
- [8] M. Beddar, *French to Arabic Machine Translation: Isomorphic Syntax, Use of Terminal Sequences*, 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.
- [9] S. Cardey et al., *The Classification Sense-Mining System*, in Advances in Natural Language Processing, Springer-Verlag – LNAI 4139, ISBN 3-540-37334-9, pp. 674-68, 2006.
- [10] Carroll, L., *Through the Looking Glass*, in: Alice's Adventures in Wonderland and Through the Looking Glass (1872), Puffin Books, Penguin Books Ltd (1974).