CHIST-ERA Conference 2011

A Micro-Systemic Approach for Dependable Natural Language Processing

Sylviane CARDEY & Peter GREENFIELD

Centre Tesnière, Université de Franche-Comté, France

http://tesniere.univ-fcomte.fr
sylviane.cardey@univ-fcomte.fr

i) Today's Starting Point

Natural Language Processing Today

For the heterogeneous data for **From Data to New Knowledge (D2K)** in the form of **natural language**, this latter is often the **weakest link** in complex systems connecting natural and artificial elements (e.g. the 1977 Tenerife airport disaster - 583 fatalities).

Compounded with this, with the exception of controlled languages, **natural language processing is notorious in defying even elementary engineering practices** where quality relies on norms and without which reliable interoperability is impossible.

NLP Reliability Today

When we look at NLP applications what strikes us first is their **unreliability**:

- Machine translation exists**, reliable** machine translation does not
- Information searching with **noise** and with **weak signals** ignored

Why is this so?

Why is NLP So Unreliable Today?

Many would say:

- The basic precepts of engineering practice are ignored (normalisation, case based testing, traceability,…)
- Evaluation/tuning counts more than fundamental research
- Corpus linguistics approaches are too favoured even though these are limited as being performance based (rather than competence) & sample based (rarely exhaustive)

But why are *these* so? Why these impasses?

What Really is the Problem with NLP Today?

- We contend that the regrettable state of NLP today is at least in part because one has forgotten (or one cannot admit) that natural language *per-se* is natural; it is **not** an artifact.
- We contend too that web semantics (e.g. RDF and SPARQL), taxonomies and such like which are suitable for artefacts are **not** part of the practice of NLP (which is not to say that they cannot be interfaced with NLP).

Complexity & Society

- Natural language is **very complex:**

  One is confronted with the well known (to linguists) language inherent phenomena such as openness (neologisms...), ambiguity, homophony, homonymy, synonymy, anaphora, 'levels' (phonology, lexis, syntax, semantics...) etc.

- Natural language is a **social phenomenon**

  One has to contend with (normalise or exploit) 'real and authentic' language as practised by real human beings (slang, 'errors', dialects…).

it deosn't mttaer in waht oredr

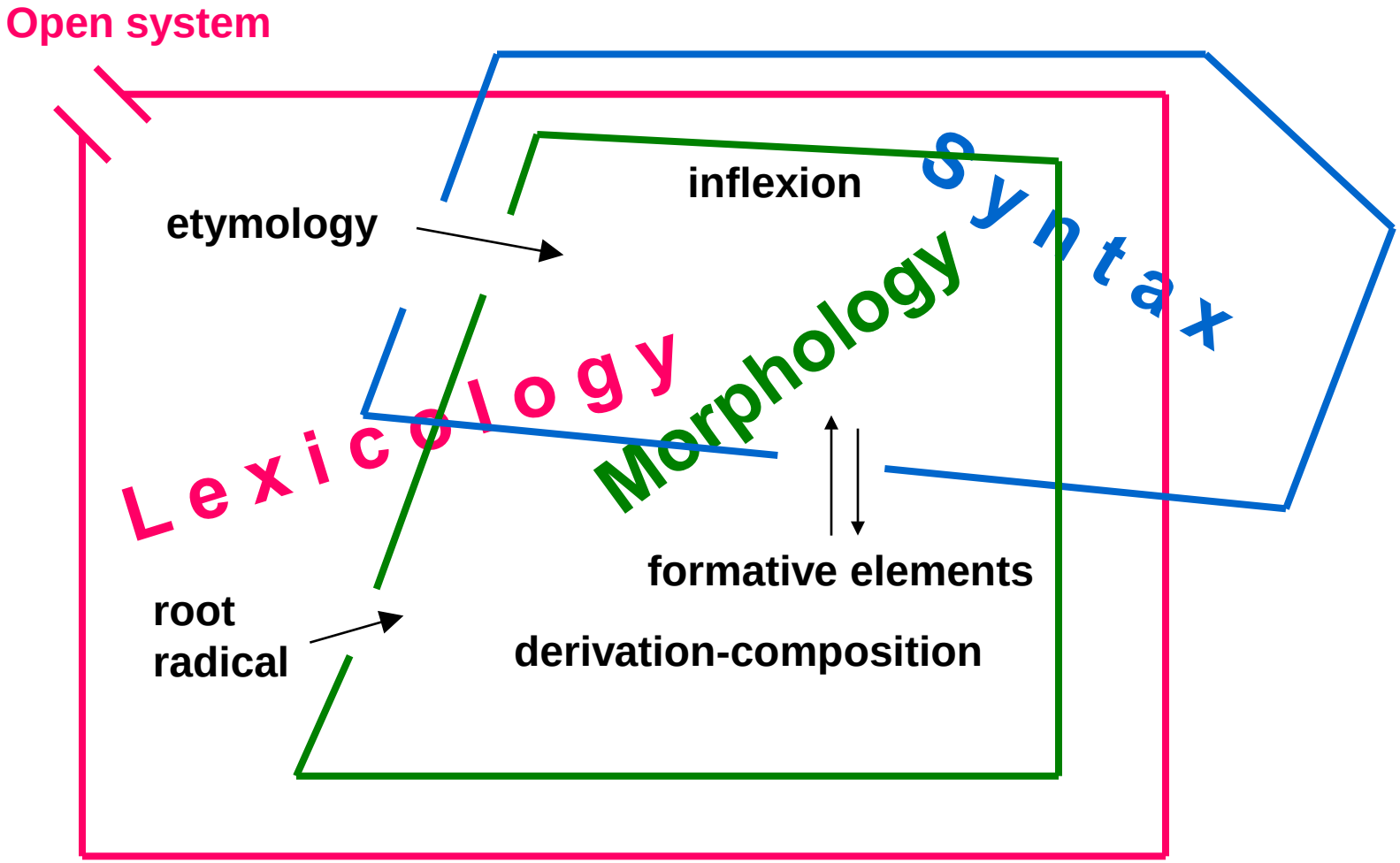
it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a total mses and you can sitll raed it wouthit a porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the ...

Information

- How can we filter and interpret information and how can we construct it and translate it?

- A message which is malformed or incorrectly interpreted or not interpreted can provoke serious catastrophes.

French morphological system



How do you spell …?

- model + ing?
- model + er?
- distil + ing
- frolic + ing?

Polycategories

In the French sentence:

'**la méchante rigole car le petit est malade**'

(the nasty woman laughs because the little boy is ill)

out of context, all the lexical units are ambiguous...

Lexical Unit	Categories
la	{Art., Nom, Pro. pers.}
méchante	{Nom, Adj.}
rigole	{Nom, Verbe coni.}
car	{Nom, Conj.}
le	{Art., Pro. pers.}
petit	{Nom, Adj.}
est	{Nom, Verbe conj.}
malade	{Nom, Adi.}

Logic, Mathematics and Poetry

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe
…

"Jabberwocky" , Lewis Carroll. Through the Looking-Glass, and What Alice Found There (1872).

Keywords

Important words and non-important ones.
The problem is what is an **important** word?
The main question is **what is a word**?

He is a has been, he has been working on the same methodology for too long.

The product ought to be perfect.

The consumer is really saying:

The product ought to be perfect **but it is not**.

? *For some months this product is no longer as it was before.*

? *The product would be very good without garlic.*

Confusions

- ***perdre*** de l'altitude */* ***prendre*** de l'altitude
- *dess**ou**s / dess**u**s* (specially for Anglophones)
- ***al****titude / **at**titude / **la**titude*
- ***up****locked / **un**locked*

NLP & Dependability Compliance for Life/Safety-Critical Applications

- Dependability compliance is not possible with statistical, keyword and other non systems based approaches.
- However, though dependability compliance is potentially possible with systems ('rule') based approaches, **in reality this is not the case**. One cannot analyse a language(s) in its entirety due to its complexity.

Machine Learning & Other 'Short-Cuts'

Machine learning approaches/existing resources/'standards' concerning **natural** language as a substitute for manual linguistic analysis ("language is too complex/too much work/too… so this is why I use a 'shortcut'") are not a panacea.

- "But the corpus did not include this case…"
- "But the tags in the 'standard' tag set make no sense for this application…"
- "But the (pre-ordained) annotation effort was at least as much in person-hours as a linguist's in-depth analysis… And the results are not-reusable…"
- "But we do not have an exhaustive case-based benchmark…"
- "But the dictionary (in extension) cannot handle neologisms…"
- …

So, what can we do?

Can we put Natural Language Processing (NLP) on a firm footing, and admit NLP to the world of dependability engineering for life/safety-critical applications?

If we can meet this challenge, then NLP for less demanding applications, today's and the future's, can surely benefit.

ii) Future Trends

NLP must provide reliable applications

We contend that the future trends in information technology, such as conformity to **mandatory regulations** concerning dependability, will impose that NLP **must provide reliable applications**. Thus we will have to:

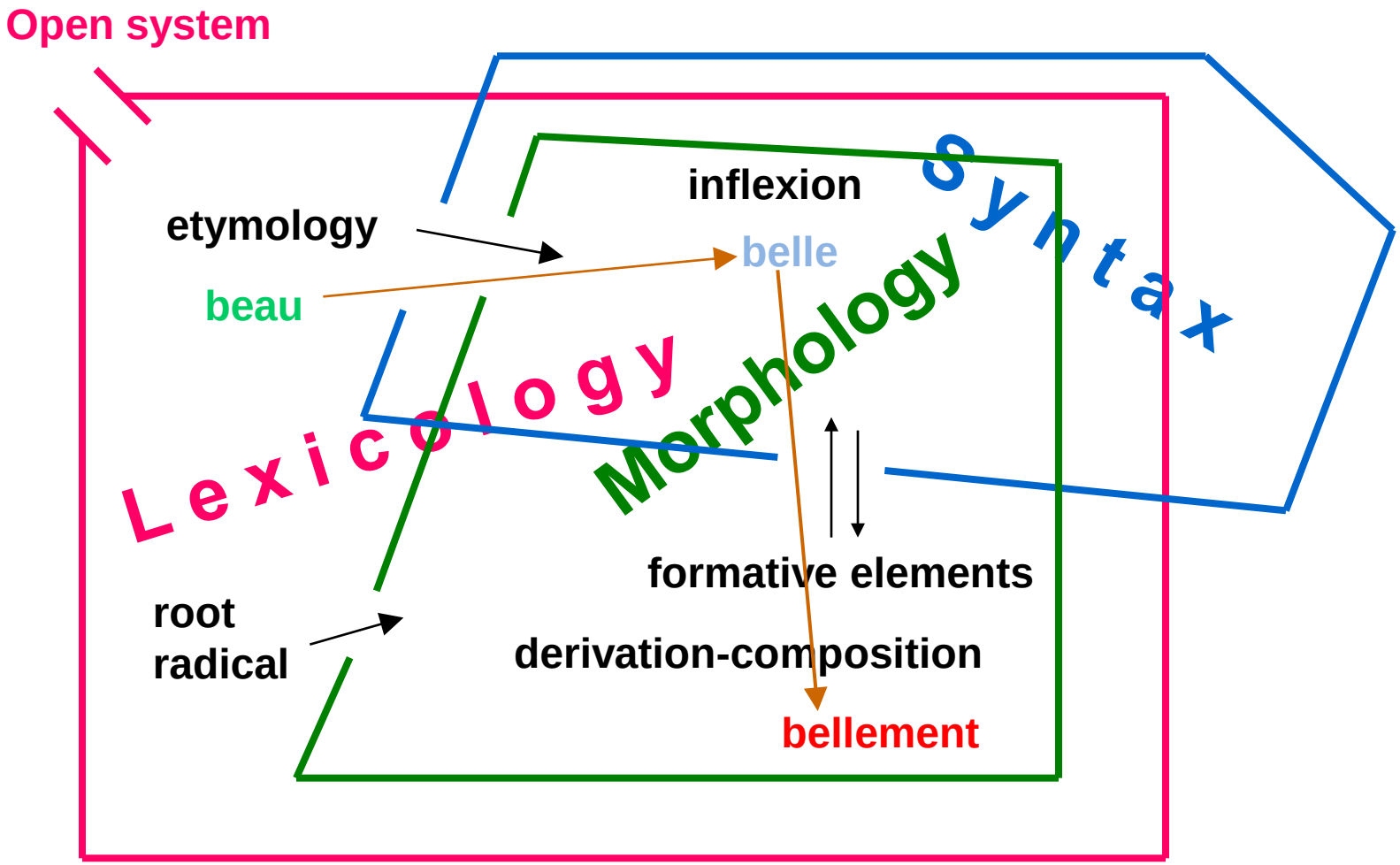
1. admit that natural language is very complex and that natural language is a social phenomenon.
2. devise the appropriate conforming analysis techniques leading to reliable NLP applications.

Centre Tesnière has and is working in this direction.

A Micro-Systemic Approach for Dependable NLP

Centre Tesnière's micro-systemic linguistic analysis approach proposes that to be processed safely languages have to be decomposed into systems which can be analysed by a human being and by machine because they are small enough but also complete so as to be able to work together as a unified system. As well as this, the systems so delimited can interact with other such systems, and this interaction is a property of language. Nothing is independent: lexis, morphology, syntax are linked.

French morphological system



Our Model

We have developed, using discrete constructive mathematics, a stable (zero obsolescence) abstract core model-theoretic model.

During the analysis for some application the ensuing processes of instantiating this model prones:

- exhaustive analyses
- fine analyses
- compositional analyses
- normalisation to promote intra/interoperability
- the linguist generalises (competence)

Micro-Systemic Linguistic Analysis

The Mathematical Model

⇒ Analysis Methodology **⇒ Algorithm Generation**

For an observed linguistic phenomenon, 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 ⇒ **Partition $\boldsymbol{P}_{nc}$**	of the canonical forms in relation with the variant forms in isolation, the context being limited to just the canonical and variant forms themselves
'in-context' (ic) categorisation ⇒ **Partition $\boldsymbol{P}_{ic}$**	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 internally related linguistic systems are involved

Given that we have partitions ($\boldsymbol{P}_{nc}$ and $\boldsymbol{P}_{ic}$), from the fundamental theorem on equivalence relations, it follows that there exist two corresponding equivalence relations $\boldsymbol{E}_{nc}$ and $\boldsymbol{E}_{ic}$, both over the binary ordered relation between the canonical forms and the variant forms $\boldsymbol{CV}$.

The Model

We model the system $\boldsymbol{S}$ which we call 'super-system' over the linguistic phenomenon by means of the binary ordered relation between the equivalence relations $\boldsymbol{E}_{ic}$ and $\boldsymbol{E}_{nc}$ (in this order) each of these over $\boldsymbol{CV}$.

Thus we have $\boldsymbol{E}_{ic}\ \boldsymbol{S}\ \boldsymbol{E}_{nc}$.

From this relation $\boldsymbol{S}$, which corresponds to a micro-system's mathematic structure and which itself is necessarily and usefully a **total surjection**, we can formulate functions for specific purposes and also generate algorithms.

Example of a Linguistic Micro-System: The Super System $S_{Doubling\_or\_not}$

Doubling or not of the final consonant in English words before –ed, -ing, -er, -est, -en

– eg: (froli**c** → froli**ck**ing)

•The relevant words in their base form (canonical),

– eg: disti**l**, mode**l**, froli**c**

•The relevant words in their derived/inflected form (the variants),

– eg: disti**ll**ing, mode**l**ing, mode**ll**ing, froli**ck**ing

model+ing is spelt modeling or modelling?

The Super System $S_{Doubling\_or\_not}$

Doubling or not of the final consonant in English words before –ed, -ing, -er, -est, -en

Conditions

Id	Condition text
cv	word with final consonant in English taking a -ed, -ing, -er, -est, -en
a	word of a syllable of the form C-V-C
b	terminated by C--V--C or by C-V(pronounced)-V(pronounced)-C
c	last syllable accented
d	terminated by -l or -m
e	used in England
f	"(un)parallel"
g	"handicap, humbug"
h	"worship, kidnap"
i	terminated par -ic
j	"wool"

Operators

Id	Operator text
N	No doubling of the consonant
D	Doubling of the consonant
K	The words terminating in –ic take –ck

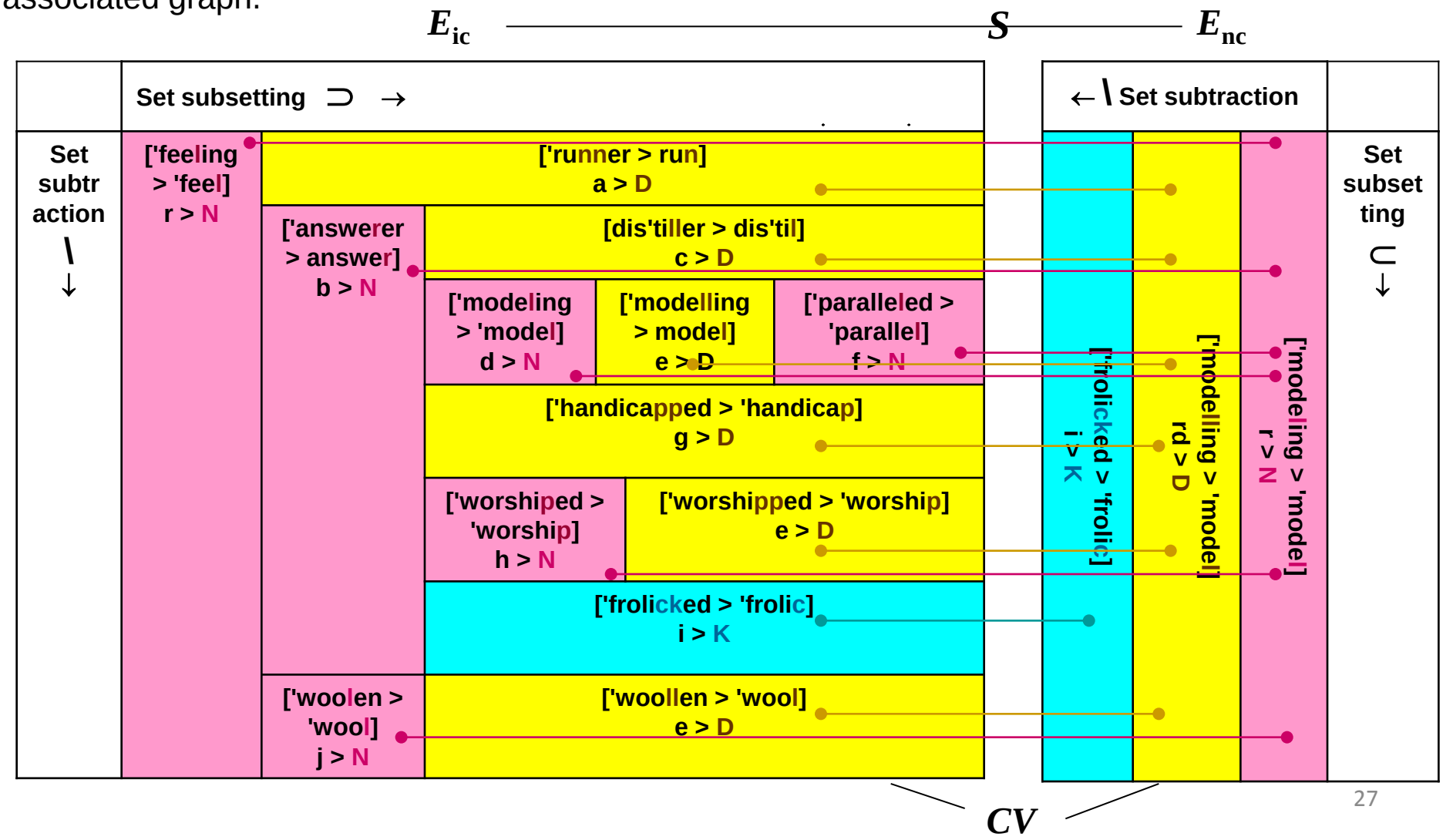
Algorithm with justifications

Level	Condition>Operator	(Canonical, Variant, Corpus)
0	cv > N	('feel, 'feeling, CD)
1	a > D	('run, 'runner, CD)
1	b > N	('answer, 'answerer, CD)
2	c > D	(dis'til, dis'tiller, CD)
2	d > N	('model, 'modeling, WD)
3	e > D	('model, 'modelling, CD)
4	f > N	((un)'parallel, (un)'paralleled, CD)
2	g > D	('handicap, 'handicapped , CD)
2	h > N	('worship, 'worshiped , WD)
3	e > D	('worship, 'worshipped , CD)
2	i > K	('frolick, 'frolicked , CD)
1	j > N	('wool, 'woolen , WD)
2	e > D	('wool, 'woollen , CD)

Legend: true, false, *undefined* (i.e. not visited), CD=CEOED 1971, WD=WNCD 1981

Representation of the Model of Super System $S_{Doubling\_or\_not}$

$S_{Doubling\_or\_not}$ is formulated as the binary ordered relation (usefully a **total surjection)** S between the equivalence relations E_{ic} and E_{nc}, each over CV, shown here with the materialisation of its associated graph.



Studygram Interactive Interpreter
Problem Source Code in Spread-Sheet Form

File Edit View Insert Format Tools Data Window Help Adobe PDF

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Title	Redoublement de la consonne finale en anglais devant les finales -ed, -ing, -er, -est, -en												
2	Version	20110216												
3	Conditions													
4	cv	Mot terminé par une consonne en anglais auquel on peut ajouter 1 des finales -ed, -ing, -er, -est, -en												
5	a	un mot d'une syllabe formée de C-V-C												
6	b	un mot terminé par C-V-C ou par C-V(prononcée)-V(prononcée)-C												
7	c	avec la dernière syllabe accentuée												
8	d	terminé par -l ou -m												
9	e	en Angleterre												
10	f	(un)parallel												
11	g	List	list_g											
12	h	List	list_h											
13	i	terminé par -ic												
14	j	wool												
15	Lists													
16	list_g	handicap	humbug											
17	list_h	worship	kidnap											
18	Operators													
19	D	Redoublement de la consonne												
20	N	Pas de Redoublement de la consonne												
21	K	Les mots termines par -ic prennent -ck												
22	Organigramme													
23	0	0	cv					'feel	N	'feeling	CEOED 1971			
24	1	1		a				'run	D	'runner	CEOED 1971			
25	2	1		b				'answer	N	'answerer	CEOED 1971			
26	3	2			c			dis'til	D	dis'tiller	CEOED 1971			
27	4	2			d			'model	N	'modeling	WNCD 1981			
28	5	3				e		'model	D	'modelling	CEOED 1971			
29	6	4					f	(un)'parallel	N	(un)'paralleled	CEOED 1971			
30	7	2			g			'handicap	D	'handicapped	CEOED 1971			
31	8	2			h			\worship	N	'worshiped	WNCD 1981			
32	9	3				e		\worship	D	'worshipped	CEOED 1971			
33	10	2			i			frolic	K	'frolicked	CEOED 1971			
34	11	1		j				\wool	N	'woolen	WNCD 1981			
35	12	2			e			\wool	D	'woollen	CEOED 1971			
36	End													
37														
38														
39														
40														
41														

démarrer Microsoft Excel - micr... FR 99% 17:42

Studygram
Consultation with trace

model+ing is spelt modeling or modelling?

```
Studygram
Centre Tesnière, Université de Franche-Comté, France
Choose application:
1         studygram_table               Simple interactive interpreter
2         studygram_organigramme        Interactive interpreter with full trace
3         model                         Generate model-theoretic model
4         studygram_test                Auto-test
q         quit                          Quit Studygram
1/2/3/4/q : 2
Microsystem: 'Redoublement de la consonne finale en anglais devant les finales -ed, -ing, -er, -
est, -en'
0:        cv: Mot terminé par une consonne en anglais auquel on peut ajouter 1 des finales -ed, -
ing, -er, -est, -en ? y/n : y
1:                a: un mot d'une syllabe formée de C-V-C ? y/n : n
2:                b: un mot terminé par C-V-C ou par C-V(prononcée)-V(prononcée)-C ? y/n : y
3:                          c: avec la dernière syllabe accentuée ? y/n : n
4:                          d: terminé par -l ou -m ? y/n : y
5:                                    e: en Angleterre ? y/n : y
6:                                              f: (un)parallel ? y/n : n
5                                     Last true condition: e Canonical: '''model'
5                                     Operator_Id: D : 'Redoublement de la consonne' Variant:
'''modelling' Justification: 'CEOED 1971'
Trace = 5 - [cv,\+ a,b,\+ c,d,e,\+ f,covj('''model',[ovj('D','''modelling','CEOED 1971')])]
```

Studygram
Generate Model-Theoretic model

```
Another consultation? y/n : n
...
3         model                                    Generate model-theoretic model
...

Model:
0 - [cv,\+ a,\+ b,\+ j,covj('''feel',[ovj('N','''feeling','CEOED 1971')])]
1 - [cv,a,covj('''run',[ovj('D','''runner','CEOED 1971')])]
2 - [cv,\+ a,b,\+ c,\+ d,\+ g,\+ h,\+ i,covj('''answer',[ovj('N','''answerer','CEOED 1971')])]
3 - [cv,\+ a,b,c,covj('dis''til',[ovj('D','dis''tiller','CEOED 1971')])]
4 - [cv,\+ a,b,\+ c,d,\+ e,covj('''model',[ovj('N','''modeling','WNCD 1981')])]
5 - [cv,\+ a,b,\+ c,d,e,\+ f,covj('''model',[ovj('D','''modelling','CEOED 1971')])]
6 - [cv,\+ a,b,\+ c,d,e,f,covj('(un)''parallel',[ovj('N','(un)''paralleled','CEOED 1971')])]
7 - [cv,\+ a,b,\+ c,\+ d,g,covj('''handicap',[ovj('D','''handicapped','CEOED 1971')])]
8 - [cv,\+ a,b,\+ c,\+ d,\+ g,h,\+ e,covj('''worship',[ovj('N','''worshiped','WNCD 1981')])]
9 - [cv,\+ a,b,\+ c,\+ d,\+ g,h,e,covj('''worship',[ovj('D','''worshipped','CEOED 1971')])]
10 - [cv,\+ a,b,\+ c,\+ d,\+ g,\+ h,i,covj(frolic,[ovj('K','''frolicked','CEOED 1971')])]
11 - [cv,\+ a,\+ b,j,\+ e,covj('''wool',[ovj('N','''woolen','WNCD 1981')])]
12 - [cv,\+ a,\+ b,j,e,covj('''wool',[ovj('D','''woollen','CEOED 1971')])]
#Model = 13
```

The Labelgram Disambiguating Tagger

'la méchante rigole car le petit est malade'

Lexical Unit	Tagger ref.	Categories	Disambig n°	Disambig ref.	Category
la	2.12	{Art., Nom, Pro. pers.}	5	45	Art.
méchante	41.2	{Nom, Adj.}	28	339	Nom
rigole	360.4	{Nom, Verbe coni.}	8	79	Verbe conj.
car	47.5	**{Nom, Conj.}**	10	144	**Conj.**
le	pre_dict	{Art., Pro. pers.}	5	45	Art.
petit	279.1	{Nom, Adj.}	28	339	Nom
est	pre_dict	{Nom, Verbe conj.}	8	74	Verbe conj.
malade	13.1	{Nom, Adi.}	28	346a	Adj.

No other system is known that can disambiguate sequences of polycategorial ambiguities.

Here we have the forward functional composition of 3 micro-systems:

- Ambiguous terms of French ;
- POS ambiguities of French ;
- Possible French syntactic structures

The full (execution) trace indicates that the first word to be disambiguated is **car**

{Nom, Conj.} → Conj.

Syntactically ambiguous sentences can be tagged too e.g.

'la petite brise la glace.'

Labelgram & Neologisms

'Twas brillig, and the slithy toves did gyre and gimble in the wabe;

Lexical unit	Out-of-context	In-context
	Categories	**Category**
it	{PROpers}	PROpers
was	{V}	V
brillig	{ADJ}	ADJ
,	{PUNCT}	PUNCT
and	{CONJ}	CONJ
the	{ADV, DET}	DET
slithy	{ADJ}	ADJ

toves	{Nplu, V3sing}	Nplu
did	{Aux}	Aux
gyre	{V}	V
and	{CONJ}	CONJ
gimble	{V}	V
in	{ADV, ADJ, PREP}	PREP
the	{ADV, DET}	DET
wabe	{N}	N
;	{PUNCT}	PUNCT

Not only have all the ***neologisms*** been tagged, but Labelgram has also disambiguated ***toves***.

Here, micro-systemic linguistic analysis has resulted in **intensional** linguistic data.

Sense Mining (information searching)
(In collaboration with Nestlé)

The product ought to be perfect **but it is not**

Centre Tesnière's micro-systemic methodology uses lexis, morphology, syntax and semantics represented by rules and sets in interrelated micro-systems.

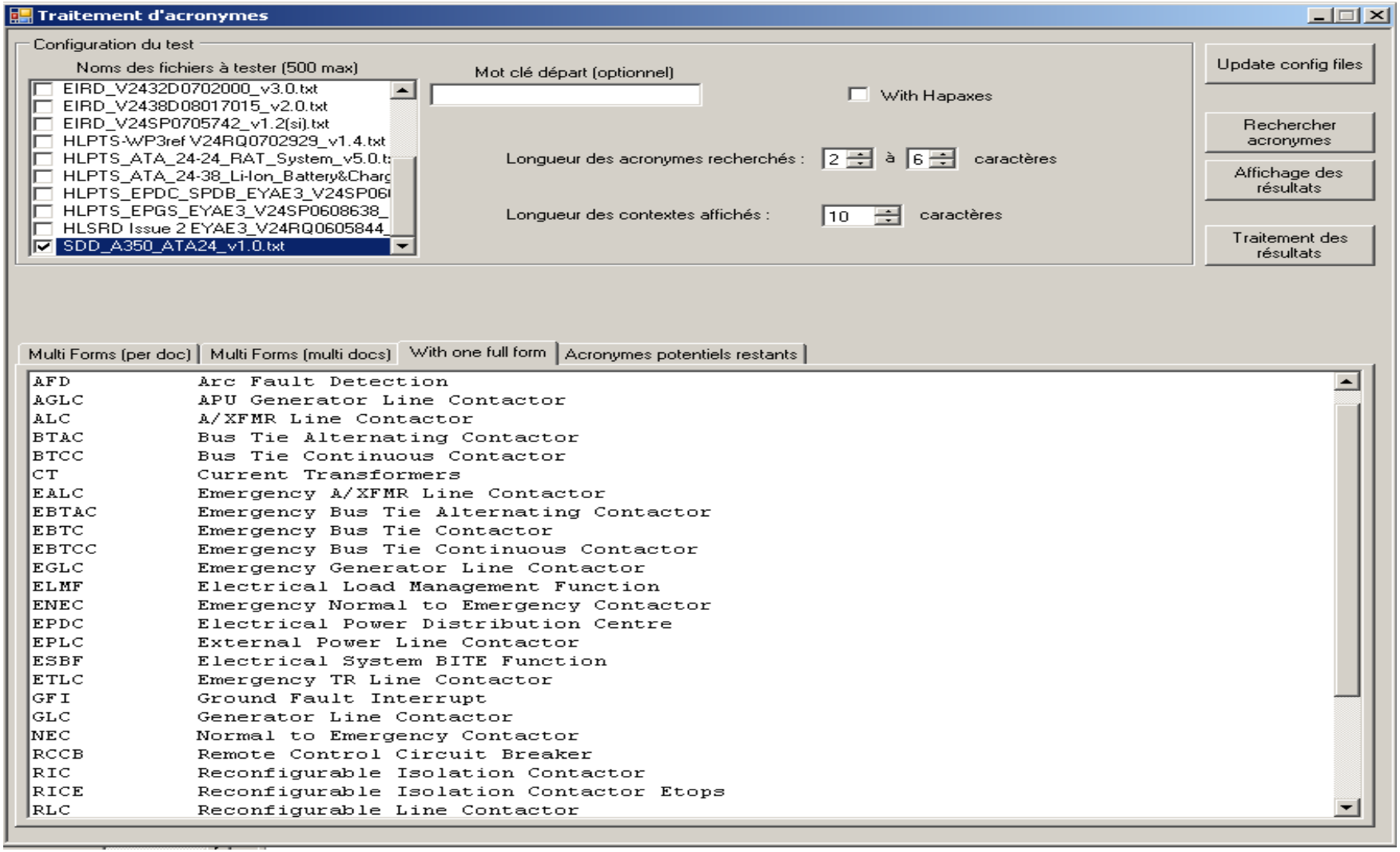
At the end we have a grammar of synonymous formulæ (rules) which allows finding one or many senses in a given text knowing that different texts can have the same meaning.

Tests have been done on 150,000 verbatims. The rate of success with raw text (email, verbatims, letters) without any preparation is 84%, and **after normalisation** it is 99%.

Automatic Acronym Extraction
Centre Tesnière & Airbus France

Automatic acronym extraction from Airbus system specification documents has been developed in view of normalising, this by using micro-systemic linguistic analysis with the linguistic data in intension (patent application in progress).

Automatic Acronym Extraction
Centre Tesnière & Airbus France



MultiCoDiCT: Formal specification in Z (J.R. Abriel)

State space extended to the domain of on-line dictionaries

On_Line_Dictionary

General_Dictionary	Include schema *General_Dictionary*
entry_collocations : seq *WORD_FORM* ⟶» *COLLOCATION*	Total surjection between the canonical entry word forms and their collocations - state invariant
entry_word_forms, orphan_word_forms : seq P *WORD_FORM*	

# *entry_word_forms* = # *entry_collocations* = # *orphan_word_forms* = *DEGREE*	Degree of the dictionary
∀ *l* : 1 .. *DEGREE* • *entry_word_forms*(*l*) = dom *entry_collocations*(*l*) ∧ (< ran *entry_word_forms*(*l*), ran *orphan_word_forms*(*l*) > partition { *can_wf, col_wf* : *WORD_FORM*, *collocation* : *collections*(*l*) | *col_wf* ⤇ *can_wf* ∈ ran *collocation* • *can_wf* })	Partitioning of the word forms in the collocations into the canonical entry word forms and the canonical orphan word forms - state invariant

Evaluation…

The following procedure which was initially developed and itself evaluated in an agro-food industry (Nestlé) application which has the benefit of incorporating case-based testing and the production of a complementary exhaustive benchmark.

Initially, we construct the raw input working corpus, partitioning this into the initial data set enabling ‘boot-strapping’ of the analysis, and sample test data sets.

These latter are subsequently used in the incremental regression testing, instrumentation and manual (i.e. the linguist) qualification sub-tasks (involving for example stability/asymptotic criteria satisfaction).

This overall approach ensures incremental verification of the validity of the system in respect of the system’s definition.

Manual qualification procedure

Recognition		Nature of context		Demonstrates	Action
Correctly recognised		Corpus attested context		Specific analysis	None – success
		Linguist's competence context	Attested category in other context(s) (e.g. in same corpus)	Generality	Add attestation to context & to automatic case-based benchmark datum
			Category not attested in e.g. same corpus, but attested in other corpus/a	Cross corpus generality	Add attestation to context & to automatic case-based benchmark datum
Error	Not recognised	Lack of cover: category and/or context missing		Location of error	Insert category and/or context, do regression test
	Incorrectly recognised	Context error		Location of error	Correction of context, do regression test

iii) Priorities for the Call

From Data to New Knowledge (D2K)

Reliable NLP

Our experience is that reliable NLP does not imply some alchemist's philosopher's stone.

For the analyses involving many very different applications undertaken in Centre Tesnière as witness the funded projects, patents and PhD theses, these all have a common basis: micro-systemic linguistic analysis.

A micro-systemic linguistic analysis properly carried out is necessarily reliable.

Disparities, Norms & Divergences intra-language & inter-language and their modelling: divergences/normalisation

Domain/Application	Modelling: S ou S^{-1}	IntRA/IntER language	Disparities/Norms/Divergences
Acronyms - détection	Variant → Canonical	IntRAlanguage	Divergences/Norms
Classificatim : Seme tables Sense-mining Seme translation	 Variant → Canonical Variant → Canonical Canonical → Variant	 IntERlanguage IntRAlanguage IntERlanguage	 Divergences Disparities/Norms – Semes Divergences
Dialectes, Jargons	Variant → Canonical	IntRAlanguage	Disparities – Provenance/Interpretation
Interférences oral-écrit-oral	Variant → Canonical	IntERlanguage	Disparities/Norms
Labelgram : POS tagging	Variant → Canonical	IntRAlanguage	Divergences
Languages, Quality and Gouvernance	Variant → Canonical	IntRAlanguage	Disparities/Norms – Measurement
CL ; TACT : Controlled Language ; Machine Translation	 Variant → Canonical ; Canonical → Variant	 IntRAlanguage IntERlanguage	 Disparities/Norms → Canonical Pivot ; Canonical Pivot → Divergences
MultiCoDiCT : Synonymy Polysemy Regionalisms	 Variant → Canonical Canonical → Variant Canonical → Variant	 IntRAlanguage IntERlanguage IntRAlanguage	 Norms Norms Divergences
Text Normalisation	Variant → Canonical	IntRAlanguage	Disparities/Norms
Paraphrase	S ou S^{-1}	IntRA/IntER	Disparities/Norms/Divergences
Sense-mining	Variant → Canonical	IntRAlanguage	Disparities/Norms – Info + Evaluation
Studygram : Linguistic problem Dialogue language	 Canonical → Variant Canonical → Variant	 IntRAlanguage IntERlanguage	 Norms – Normative grammar Divergences

Centre Tesnière Projects, Patents & Collaborators: Micro-Systemic Linguistics

- 2002: Airbus Operations SAS : **Patent** Domain of **controlled languages** (Brevet n° 02 07774 "Procédé et dispositif pour élaborer une forme abrégée d'un terme quelconque qui est utilisé dans un message d'alarme destiné à être affiché sur un écran du poste de pilotage d'un aéronef").
- 2003: **Paten**t Domain of **machine translation** (Numero 03 05149 "Procédé et dispositif de traduction automatique dans une langue cible d'au moins un énoncé comprenant une suite de mots, formé dans une langue source").
- 2005 – 2006: **Project** Classificatim. Domain of **sense mining** conducted in collaboration with an agro-food industry enterprise, Nestlé.
- 1/2007 – 12/2009: **Project** LiSe. ANR-06-SECU-007. Domain of **sense mining, machine translation** and **controlled languages**. Linguistique, normes, traitement automatique des langues et Sécurité : du « data et sense-mining » aux langues contrôlées, http://projet-lise.univ-fcomte.fr/. Coordinator Centre Tesnière. Airbus Operations SAS was a partner
- 1/2008 – 8/2009**: Project** MESSAGE. JLS/2007/CIPS/02. Domain of **controlled languages** Alert Messages and Protocols JLS/2007/CIPS/02., http://message-project.univ-fcomte.fr Coordinator Centre Tesnière.
- 2011 – 18 months. **Project** Sensunique. ANR Programme Emergence, Edition 2010. Domain of **controlled languages** Optimization of software for high quality technical writing: a pilot application in the field of health. Coordinator Centre Tesnière.
- Current: Airbus Operations SAS **Patent Application** Domain of **sense mining,** automatic acronym extraction.

[PhD Theses (micro-systemic linguistic analysis):Centre Tesnière] 1/3

1997 : **Bilal AL SHAFI**, mention **très honorable à l'unanimité du jury** « Traitement informatisé des signes diacritiques pour une utilisation automatique et didactique »

1997 : **Zahra EL HAROUCHY**, mention **très honorable et félicitations à l'unanimité du jury** « Dictionnaire et grammaire informatisés pour la levée des ambiguïtés »

1998 : **Hui-Lan CHAO**, mention **très honorable et félicitations à l'unanimité du jury** « Compréhension automatique de phrases interrogatives françaises et chinoises, application dans le cadre de bases de données »

2000 : **Mi-Seon HONG**, mention **très honorable et félicitations à l'unanimité du jury** « Modèle théorique et représentation formelle de la sémantique des langues éloignées : application au couple coréen-français en traduction automatique »

2000 : **Frédérique DEPAIN-DELMOTTE**, mention **très honorable et félicitations à l'unanimité du jury** « Proposition d'un modèle linguistique pour la résolution d'anaphores en vue du traitement automatique des langues »

2001 : **Walid EL ABED**, mention **très honorable et félicitations à l'unanimité du jury** « Méta modèle sémantique et noyau informatique pour l'interrogation multilingue des bases de données en langue naturelle (théorie et application) »

2001 : **Eliza GAVIEIRO (CIFRE Airbus)**, mention **très honorable et félicitations à l'unanimité du jury** « Vers un modèle d'élaboration de la terminologie d'une langue contrôlée ; application aux textes d'alarmes en aéronautique pour les futurs postes de pilotage »

2002 : **Dalila LIMAME**, mention **très honorable et félicitations à l'unanimité du jury** « Vers un Système de Traduction des Expressions Polysémiques ; le sytème S.T.E.P. »

2002 : **Izabella THOMAS**, mention **très honorable et félicitations à l'unanimité du jury** « Vers un modèle d'interprétation du groupe Adjectif Nom/Nom Adjectif en vue de la traduction automatique (application du français vers le polonais) »

2003 : **Haytham ALSHARAF, mention très honorable et félicitations à l'unanimité du jury** « Vers un système de traduction automatique du langage juridique du français vers l'arabe »

[PhD Theses (micro-systemic linguistic analysis):Centre Tesnière] 2/3

2003 : **Yihui SHEN**, mention **très honorable et félicitations à l'unanimité du jury** « Formalisation des phrases injonctives : application à la traduction automatique chinois-français »

2004 : **Gaëlle BIROCHEAU**, mention **très honorable à l'unanimité du jury** « Etiquetage morphologique et contribution à la désambiguïsation automatique des ambiguïtés morphologiques sur un lexique anglais »

2004 : **Igor SKOURATOV**, mention **très honorable à l'unanimité du jury** « caractéristiques typologiques des néologismes dans le français contemporain : aspects linguistiques et sociolinguistiques »

2004 : **Séverine VIENNEY**, mention **très honorable et félicitations à l'unanimité du jury** « Analyse et syntaxe pour la correction automatique : application à l'accord du participe passé et du verbe en général »

2004 : **Héléna MORGADINHO**, mention **très honorable et félicitations à l'unanimité du jury** «Analyse pour un système d'étiquetage morphologique et de désambiguïsation morphosyntaxique : Labelgram espagnol »

2004 : **Hsiang-I LIN**, mention **très honorable et félicitations à l'unanimité du jury** « Vers une traduction automatique des expressions figées françaises en chinois : la traduction canonique »

2005 : **Mounira BIOUD**, mention **très honorable et félicitations à l'unanimité du jury** « Une normalisation de l'emploi de la majuscule et sa représentation formelle pour un système de vérification automatique des majuscules dans un texte »

2006 : **Kyoko KURODA**, mention **très honorable et félicitations à l'unanimité du jury** « Traduction Automatique : Divergences de Traduction entre le japonais et le français »

2006 : **Xiaohong WU**, mention **très honorable et félicitations à l'unanimité du jury** « Conception d'une langue contrôlée pour un système de traduction automatique de protocoles médicaux ; application aux domaines de l'echinococcose et au clonage moléculaire »

2006 : **Hadnane ECHCHOURAFI**, mention **très honorable à l'unanimité du jury** « Vers une reconnaissance des composés pour une désambiguïsation automatique (composés à trois, quatre, cinq et six éléments »

2007 : **Gabriel SEKUNDA**, mention **très honorable à l'unanimité du jury** « Vers une classification des emplois des structures de la langue française contenant un infinitif en vue de leur traduction en langue polonaise »

2007 : **Aleksandra DZIADKIEWICZ**, mention **très honorable et félicitations à l'unanimité du jury** « Vers une reconnaissance et une traduction automatique de phraséologismes pragmatiques (application du français vers le polonais »

[PhD Theses (micro-systemic linguistic analysis):Centre Tesnière] 3/3

2007 : **Sombat KHRUATHONG**, mention **très honorable et félicitations à l'unanimité du jury** « Vers une analyse micro-systémique en vue d'une traduction automatique thaï-français : application aux verbes sériels »

2007 : **Abdelouafi GHENIMI**, mention **très honorable et félicitations à l'unanimité du jury** « Conception d'un modèle de traduction automatique arabe-français appliqué au domaine des mathématiques »

2007 : **Eun Soon YOU**, mention **très honorable et félicitations à l'unanimité du jury** « Le traitement des unités lexicales polysémiques (l'adjectif et le verbe)
vers un système de traduction automatique »

2007 : **Rosita CHAN**, mention **très honorable et félicitations à l'unanimité du jury** « Vers un modèle de dictionnaire pour le traitement de divergences de traduction français-espagnol-français. Applications au domaine du tourisme »

2008 : **Thierry LECOLINET,** mention **très honorable** «Termes de la mythologie : évolution de sens ou de forme en diachronie »

2008 **: Joseph BAUDOUIN,** mention **très honorable «** Les ambiguïtés de la langue arabe »

2009 : **Hj Md Said Mohd SAUPI**, mention **très honorable**, « Le malais: études en diachronie et représentation formelle »

2010 : **Ziad MIKATI**, mention **très honorable et félicitations à l'unanimité du jury,** « Du Data mining au Sense mining : Modèle pour une analyse de la langue arabe et ses représentations formelles en vue d'une application à des domaines demandant une haute sécurité »

2010 : **Mohammed AL-ZAHRANI**, mention **très honorable et félicitations à l'unanimité du jury, « Théorie systémique et microsystémique :** vers une modélisation de règles en vue d'applications au français et à l'arabe »

2010 : **Julie RENAHY**, mention **très honorable et félicitations à l'unanimité du jury,** «Conception d'une langue contrôlée généralisante (Application aux domaines de la santé et sécurité civile) »

Some Revealing Experiences

- Standards for linguistic elements serve of little use:
  - The French POS tags for translating French to Arabic are not the same as for French to Chinese
- Existing resources serve of little use – adjusting/reworking/interfacing takes longer than creating what is really wanted from scratch – and what you get is reliable too

iii) Priorities for the Call
NLP specific priorities 1/3

- NLP must enter the domain of real engineering practice enabling dependability compliant applications for life/safety-critical applications
- NLP needs to be supported by discrete & constructive mathematical model-theoretic approaches
- Such approaches must provide inherent analysis, synthesis & evaluation capabilities

iii) Priorities for the Call
NLP specific priorities 2/3

- Abstract mathematical model which is accessible to the regulatory authorities, the theoretical linguist and the software engineer
- Formal methods
- Analysis process proning:
  - exhaustiveness,
  - fine analyses and
  - normalisation
- Incremental verification of the validity of the system
- Objective results evaluation with blind evaluation

iii) Priorities for the Call
NLP specific priorities 3/3

- Inherent tracing (static & dynamic)
- Correctibility (source of errors)
- Case based testing enabling exhaustive objective benchmarks
- Declarative 'linguistic' programming
- Representations' well-formedness verification and algorithm generation/optimisation (abstract interpretation)
- Light-weight: only analyse what the application requires
- Intensional rather than extensional linguistic data

Selected References 1/2

- CARDEY S., GREENFIELD P., (2003), « Disambiguating and Tagging Using Systemic Grammar », in: proc. 8th ISSC, Santiago de Cuba, Actas I :559-564.
- CARDEY S., GREENFIELD P., (2005), « A Core Model of Systemic Linguistic Analysis », in: proc. of RANLP-2005, Borovets, Bulgaria, 21-23 September 2005 :134-138.
- CARDEY, S., (2006), How to avoid interferences with other languages when constructing a spoken controlled language, in: «La comunicazione parlata/Spoken Communication», ISBN 88 7092 238 3, 2006, Naples, Italy.
- CARDEY S., GREENFIELD P., (2006), « Systemic Linguistics with Applications », in: Linguistics in the Twenty First Century, Eds. E.M.Bermúdes and L.R. Miyares, Cambridge Scholars Press, United Kingdom. ISBN 1904303862, 2006 :261-271.
- CARDEY, S., GREENFIELD P., BIOUD, M., DZIADKIEWICZ H., KURODA K., MARCELINO I., MELIAN C., MORGADINHO H., ROBARDET G., VIENNEY S. (2006), « The Classificatim Sense-Mining System », in: Advances in Natural Language Processing, Springer-Verlag – LNAI 4139, ISBN 3-540-37334-9 :674-684.
- CARDEY S., GREENFIELD P., (2008), « Micro-systemic linguistic analysis and software engineering: a synthesis », in: revue RML6, Actes du Colloque International en Traductologie et TAL, 7 et 8 juin 2008, Oran.
- CARDEY, S., GREENFIELD, P., ANANTALAPOCHAI, R., BEDDAR, M., DEVITRE, D., JIN, G., (2008), « Modelling of Multiple Target Machine Translation of Controlled Languages Based on Language Norms and Divergences », in: proc. ISUC2008, Osaka, Japan, December 1516, 2008, IEEE Computer Society, ISBN 978-0-7695-3433-6, pp 322-329.

Selected References 2/2

- CARDEY S., (2009), « Controlled Languages for More Reliable Human Communication in Safety Critical Domains », in: proc. 11th ISSC, Santiago de Cuba, Cuba, 19-23 January 2009, ISBN:978-959-7174-14-1 :330-335.
- CARDEY S., DEVITRE D., GREENFIELD P., SPAGGIARI L., (2009), « Recognising Acronyms in the Context of Safety Critical Technical Documentation », in: proc. ISMTCL, BULAG, PUFC, ISSN 0758 6787, ISBN 978-2-84867-261-8 :56-61.
- CARDEY S., BOGACKI K., BLANCO X., MITKOV R., (2010) « Resources for Controlled Languages for Alert Messages and Protocols in the European Perspective », in: proc. LREC 2010, Malta, 17-23 May 2010.
- CARDEY S., (2011), « Machine Translation of Controlled Languages for More Reliable Human Communication in Safety Critical Applications », in: proc. 12th ISSC, Santiago de Cuba, Cuba, January 17-21, 2011, Vol. II, ISBN: 978-959-7174-19-6 :953-958.
- GREENFIELD P. (1998) « Invariants in multilingual terminological dictionaries », in: P.-A. BUVET (ed.), Figement et traitement automatique des langues naturelles, Bulletin de linguistique appliquée et générale, 1998, n° 23, 111-121.
- KAMPEERA W., CARDEY S., (2011), « Paraphrases in Natural Language Processing », in: proc. 12th ISSC, Santiago de Cuba, Cuba, January 17-21, 2011, Vol. II, ISBN: 978-959-7174-19-6 :963-9677.
- SPAGGIARI, L., CARDEY, S., (2010), « A System to Control Language for Oral Communication », in: Advances in Natural Language Processing, coll. Lecture Notes in Artificial Intelligence, Springer-Verlag, LNAI 6233, ISSN 0302-9743, ISBN-10 3-642-14769-0, ISBN-13 978-3-642-14769-2 :393-400.

Conclusion

- At Centre Tesnière we are of the opinion that natural language processing must become **reliable** – **and that this is possible.**
- We are of the view that we need one theory but the applications being different, tags etc. certainly vary.
- We do not need extensive corpora. What we do need however are **representative** corpora.
- We systematically try and solve the most difficult problem first.
- Centre Tesnière's applications with Nestlé (sense mining) and Airbus France (controlled languages) are very different, but both are based on the same theory. Both applications are light-weight and have been built from scratch, so we have no 3rd party dependencies.

"Languages respect some norms; if not, it would be impossible to learn a language",
(CARDEY *et al*., 2006)

CHIST-ERA Conference 2011

A Micro-Systemic Approach for Dependable Natural Language Processing

Sylviane CARDEY & Peter GREENFIELD

Centre Tesnière, Université de Franche-Comté, France

http://tesniere.univ-fcomte.fr
sylviane.cardey@univ-fcomte.fr