What I Am Up To

Presentation of Personal Research Areas

Leo Wanner

November 20, 2003

Overview

1. Text Generation
    - Applied Text Generation
    - Sentence Planning in Text Generation (esp. the Role of the Information Structure in SP)
2. Machine Translation
    - Divergences in the Transfer Paradigm (in Cooperation with Prof. I. Mel’čuk, Université de Montréal)
3. Lexicology/Lexicography
    - (Multilingual) Collocational and Valency Lexica

4. The Use of Collocational Information in NLP

    - Text Generation, Machine Translation, Information Retrieval, and Information Extraction

5. Acquisition of Lexical Information from Corpora

    - Functional Roles of Modifiers
    - Collocations

6. Graph Grammars in Paraphrasing and Generation (Importing Findings from Theoretical Computer Science into NLP)

Text Generation

Text Generation: Mapping abstract data or knowledge representation structures onto surface realizations of natural language sentences.

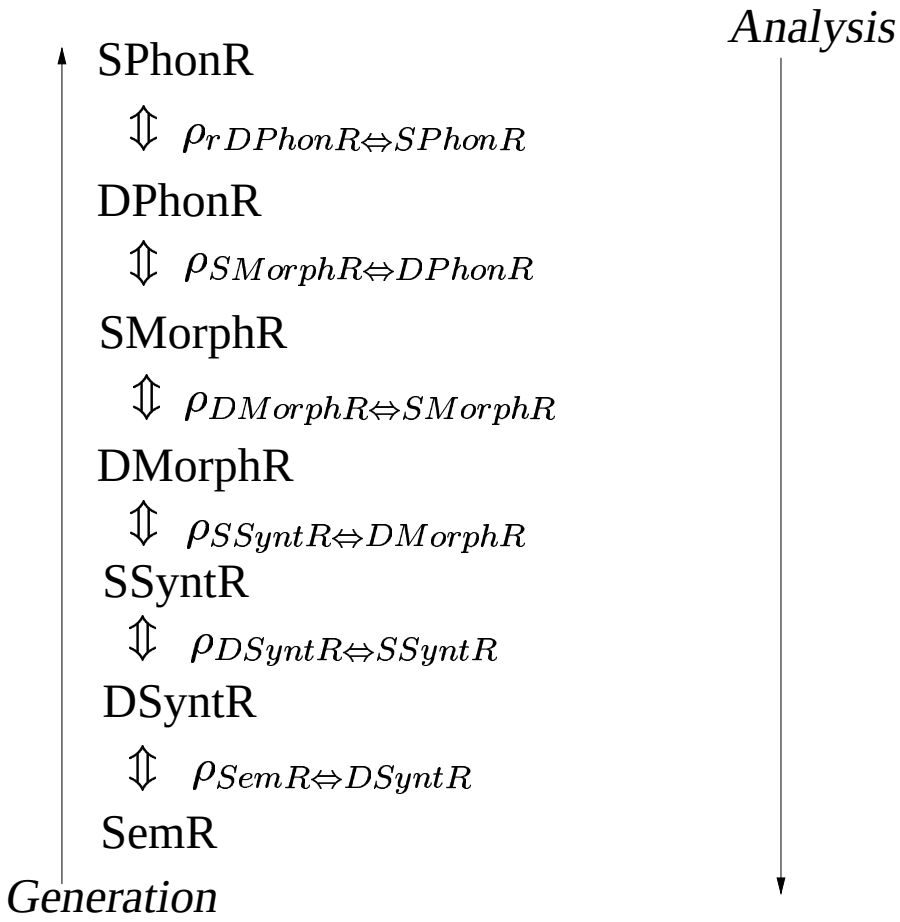
Text Generation Tasks:

- *Content Selection*: Selection of what is to be said depending on the *communicative goals* of the reader
- *Text Planning*: Determination of the *discourse structure* of the text to be generated.
- *Sentence Planning*: Cutting discourse units into sentences and determination of the form of these sentences.
- *Surface Realization*: Morphosyntactic realization of the sentences

Possible Starting Representation

```
<?xml version="1.0" ?>
<!DOCTYPE  msequences SYSTEM "Meas.dtd">
<msequences  date="14 08 2002" s="ozone" unit="mug/m3" l="Heilbronn"
   r="Mittlerer Neckarraum">
<msequence>
<rmin st="Esslingen" t="18" v="94"/>
<rmax st="Plochingen" t="18" v="217"/>
...
<measure t="07"  v="0"/>
<measure t="08"  v="2"/>
<measure t="09"  v="11"/>
<measure t="10"  v="54"/>
...
<measure t="17"  v="182"/>
<measure t="18"  v="198"/>
</msequence></msequences>
```

Underlying Linguistic Model (MTT)



Information Structure in Sentence Planning 1

The *Information Structure* is "superimposed" on the semantic structure of a sentence, dividing it along the following six dimensions:

1. Thematicity (*Rheme* vs. *Theme*),
2. Giveness (*Given* vs. *New*),
3. Focalization (*Focalized* vs. *Non-Focalized*),
4. Perspective (*Foregrounded* vs. *Backgrounded*),
5. Presupposedness (*Presupposed* vs. *Asserted*),
6. Unitariness (*Unitary* vs. *Articulated*).

Information Structure in Sentence Planning 2

Each parameter of each communicative dimension is reflected by specific syntactic and lexical means.

$\Rightarrow$

The *Information Structure* predetermines the syntactic structure and (to a certain extent) the lexical items of the sentences to be generated. Cf.:

1. *The ozone concentration*$_{Th.Given}$ *reached this morning at the Urquinaona monitoring site 193* $\mu gr/m^3$.

2. *The monitoring site at Urquinaona*$_{Th.Given}$ *reported this morning an ozone concentration of 193* $\mu gr/m^3$.

Information Structure in Sentence Planning 3

Research topics (in the *Meaning-Text Theory* Framework):

- How can the Information Structure be derived in order to be usable in Applied Text Generation?
  - Three major sources: (i) discourse structure of the text, (ii) communicative intentions in connection with the information on the domain, (iii) domain restrictions.

- How can the Information Structure be used in Applied Text Generation?
  - Extention of the Generation Grammars to handle (i) Information Structures, (ii) the constraints the Information Structure implies for the surface realization of the sentences.

Machine Translation: Mismatches between $\mathcal{L}_S$ and $\mathcal{L}_T$

Two equivalent structures S_S and S_T show a mismatch if they are not isomorphic. A mismatch can occur at the (i) semantic, (ii) lexical, (iii) syntactic, and/or (iv) morphological level.

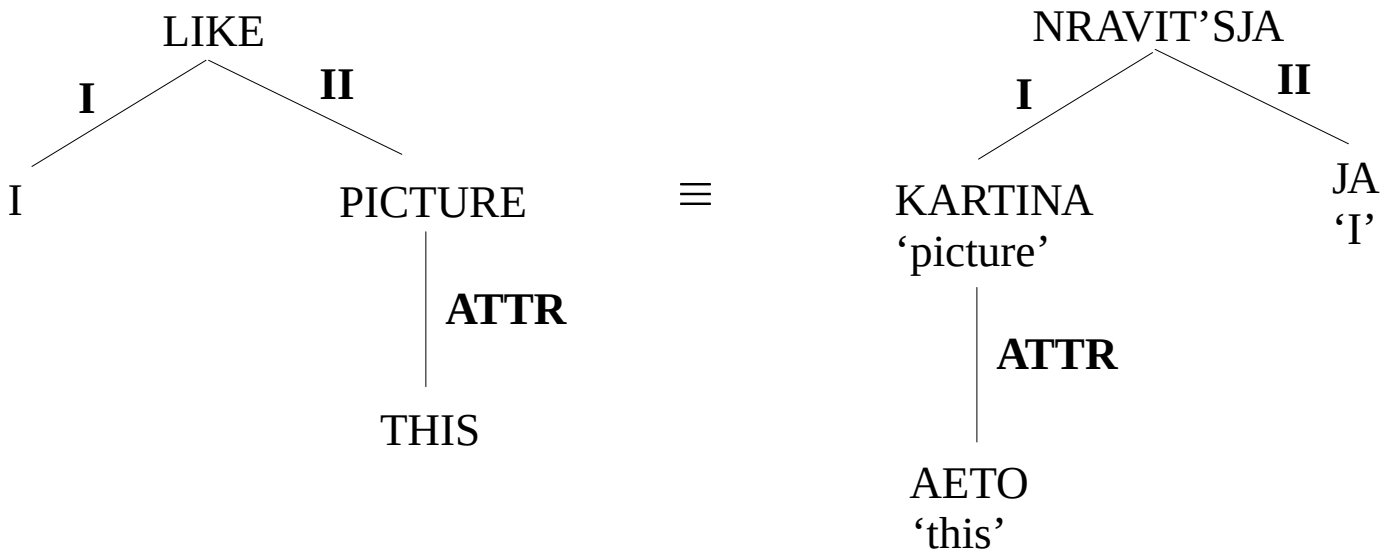
Examples of syntactic mismatches:

I like this picture. $\Leftrightarrow$ Rus. *Mne nravitsja èta kartina*
lit. 'To-me pleases this picture'.

I just learned that. $\Leftrightarrow$ Fr. *Je viens de l'apprendre*
lit. 'I come to learn that'.

Germ. *Ich schwimme gern* $\Leftrightarrow$ *I like swimming*.
lit. 'I swim with-pleasure'.

Machine Translation: Examples of Mismatches



Machine Translation

Research Topics:

- Develop and formally describe an exhaustive typology of lexical, syntactic, and morphological mismatches between $\mathcal{L}_S$ and $\mathcal{L}_T$

- Develop a model of syntactic transfer that is based on the paraphrasing paradigm.

  Underlying hypothesis:

  All types of syntactic mismatches that are encountered interlinguistically are also encountered intralinguistically

  $\Rightarrow$

  The transfer of a structure S_S in $\mathcal{L}_S$ into a structure S_T in $\mathcal{L}_T$ can be described in terms of a paraphasing grammar.

Lexicology/Lexicography:
Multilingual Collocational Lexica 1

Collocation: A *Collocation* is a combination of two lexical items, in which the semantics of one of the lexical items (the *base*) is autonomous from the combination it appears in, and where the other lexical item (the *collocate*) adds semantic features to the semantics of the base.

Examples of collocations:

bachelor	confirmed	hit	hard	accusation	prove
rain	heavy	breath	heavily	hint	take
demand	legitimate	cut	neatly	analysis	carry out
smoker	heavy	walk	steadily	attention	receive

Lexicology/Lexicography:
Multilingual Collocational Lexica 2

Important properties of a Collocation:

- it is a binary combination of lexical items
- it possesses a coherent syntactic structure, i.e., the base and the collocate always possess the same grammatical function with respect to each other
- it is a lexically restricted word combination

HOWEVER:

A partial semantic correlation between the semantic features of a base and the collocations this base occurs in exists! $\Rightarrow$ An efficient representation of collocational information is possible.

Examples

sentir	ADMIRACIÓN, ALEGRIA, APRENSIÓN, DECEPCIÓN, DESESPERACIÓN, ENDFADO, ENTUSIASMO, . . .
tener	ENTUSIASMO, ENFADO, ESPERANZA, IRA, MIEDO, ODIO, SORPESA, . . .
producir	ALEGRIA, APRENSIÓN, ENTUSIASMO, HORROR, PANICO, SORPRESA, . . .

Lexicology/Lexicography:
Multilingual Collocational Lexica 4

Research Topics:

- Lexicological Issues:

  Study of the correlation between the semantics of lexemes and the collocations these lexemes can occur in. In other words: Look for lexico-semantic features of lexemes that allow them to form non-free combinations with other lexemes.

- Artificial Intelligence Issues:

  Design inheritance techniques that allow for a generalization of collocational information in both monolingual and multilingual lexica.

Acquisition of Lexical Information from Corpora

Functional Roles of Modifiers:

These	*two*	*beautiful*	*electric*	*trains*	*with pantographs*
DEICTIC	NUMERATIVE	EPITHET	CLASSIFICATORY	OBJECT	QUALIFYING

Relevance, e.g., for Text Generation:

- The function of modifiers determines their order in an NP:

  **electric beautiful trains*; **two these trains*

- Two modifiers with different functions cannot be coordinated:

  **beautiful and electric trains*

Acquisition of Lexical Information from Corpora

Research Topics:

Automatic construction of lexica that contain, among other information, functional (and semantic) labels—with a focus on modifiers.

Approach:

- Bootstrapping from a small set of training examples
- Using linguistic clues (coordination, inflexion, etc.)

So far, a small lexicon for German has been compiled.

Automatic Recognition and Semantic Classifi cation of Collocations

Collocation: A *Collocation* is a combination of two lexical items, in which the semantics of one of the lexical items (the *base*) is autonomous from the combination it appears in, and where the other lexical item (the *collocate*) adds semantic features to the semantics of the base.

More formally:

A collocation $A \oplus B$ is a semantic phraseme such that the meaning composition 'X' expressed by $A \oplus B$ is constructed out of the signified 'A' and the meaning 'C', i.e. 'X' = 'A$\oplus$C', such that the lexical item B expresses 'C' contintent on A.

Other Definitions of Collocations

(Halliday, 1961:276)

> Collocation is the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur, at n removes (a distance of n lexical items) from an item X, the items A,B,C, ...

(Hausmann, 1976), (Smadja, 1993)

> Collocation is an arbitrary, recurrent and cohesive word combination which reveals an idiosyncratic syntactic structure and is specific to a given domain.

Typology of Collocations

Collocations can be classified according to the meaning component ‘C’; cf. examples of collocations with the same ‘C’:

A		B
ACCIDENT	$\xrightarrow{r}$	HAVE
EXAM	$\xrightarrow{r}$	GIVE
SPEECH	$\xrightarrow{r}$	DELIVER
STEP	$\xrightarrow{r}$	TAKE
SUICIDE	$\xrightarrow{r}$	COMMIT

$\Rightarrow$ *Lexical Functions* (Mel’čuk, 1996)

Syntagmatic Lexical Functions

A (*standard*) *syntagmatic Lexical Function* is a (directed) standard abstract relation r that holds between the base A and the collocate B of the collocation $A \oplus B$ and that denotes 'C' $\subset$ '$A \oplus C$' with '$A \oplus C$' being expressed by $A \oplus b$.

- Standard: r applies to a large number of collocations
- Abstract: r is sufficiently general and can therefore be exploited for purposes of classification

In total, about sixty different "simple standard" LFs are distinguished. They are assumed to provide an exhaustive semantico-syntactic typology of the stock of idiosyncratic binary word combinations in natural languages.

Examples of LFs 1

1.	‘perform’, ‘do’, ‘act’ ($Oper_1$)		3.	‘concern’, ‘apply to’ ($Func_2$)	
	BLOW	*deal*		ANALYSIS	*concern*
	OBSTACLE	*pose*		BLOW	*fall* [*upon*]
	RESISTANCE	*put up*		CHANGE	*affect*
	SUPPORT	*lend*		LECTURE	*be* [*on*]
2.	‘undergo’, ‘meet’ ($Oper_2$)		4.	‘act accordingly’ ($Real_1$)	
	BLOW	*receive*		ACCUSATION	*prove*
	OBSTACLE	*encounter*		PROMISE	*keep*
	RESISTANCE	*meet*, *run* [*into*]		SCHEDULE	*stick* [*to*]
	SUPPORT	*receive*		THREAT	*fulfil*

Examples of LFs 2

5.	‘happen’, ‘take place’ ($Func_0$)		7.	‘react accordingly’ ($Real_2$)	
	ACCIDENT	*happen*		DEMAND	*fulfil*, *meet*
	RAIN	*fall*		HINT	*take*
	RUMOUR	*circulate*		LAW	*abide* [*to*]
	SMELL	*linger*		CALL	*answer*
6.	‘originate from’ ($Func_1$)		8.	‘put an end to’ ($Liqu_1Func_0$)	
	ANALYSIS	*be due* [*to*]		SUPPORT	*withdraw*
	BLOW	*come* [*from*]		RESISTANCE	*put down*
	PROPOSAL	*stem* [*from*]		OBSTACLE	*remove*
	SUPPORT	*come* [*from*]		MEETING	*end*

An automatic classification of word combinations extracted from the corpus according to LFs would obviously be extremely helpful for both (computational) lexicography and NLP-applications.

The Basic Idea: Instance-Based Machine Learning

1. For each LF to be used in the classification procedure, collect a set of sample instances (i.e., compile “training sets”).

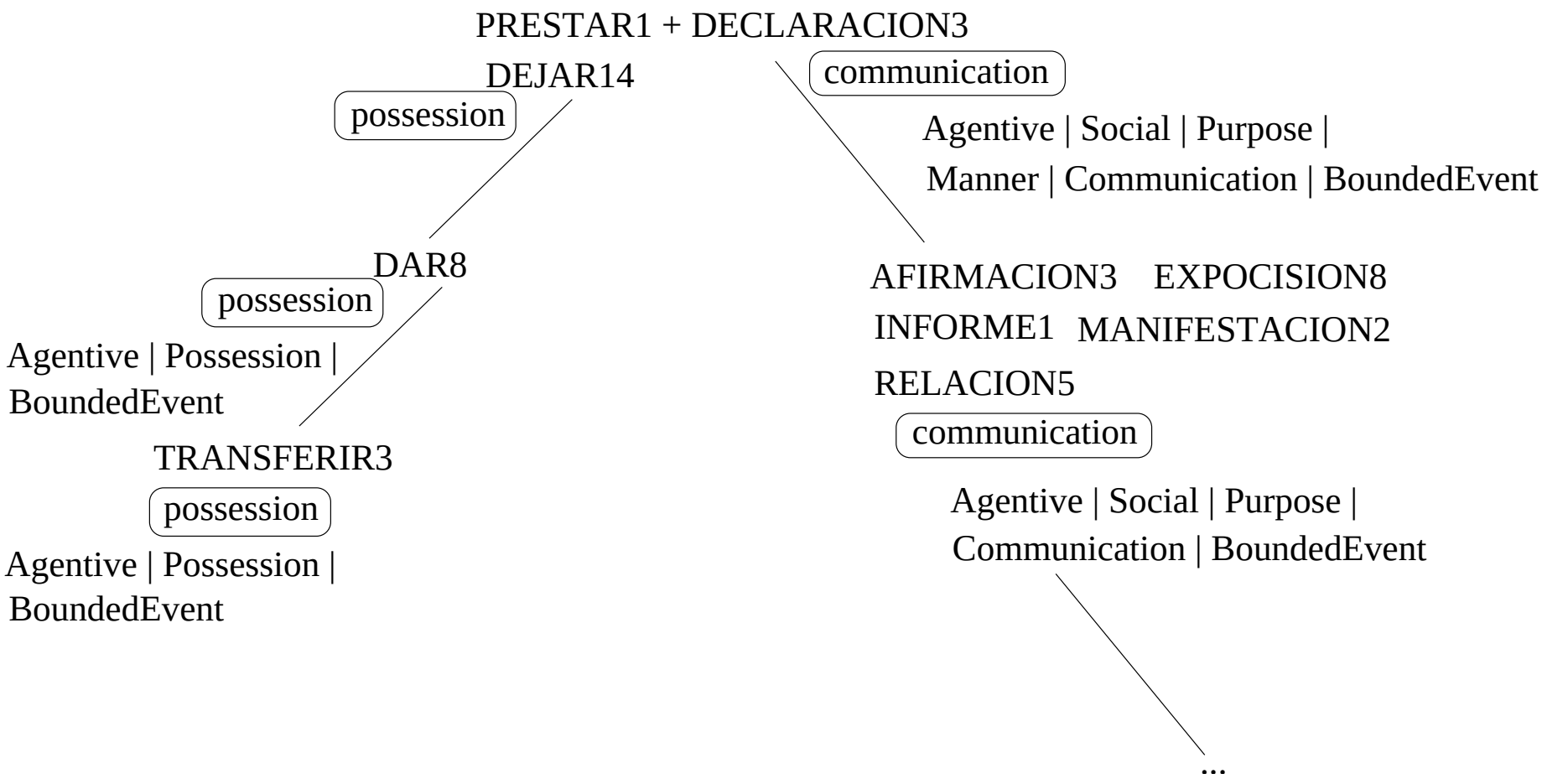
2. “Learn” what it means to be an instance of a specific LF by analysing the meaning of each instance in the corresponding training set.

3. Drawing on this analysis, calculate for each LF an artificial “prototypical” instance (called *centroid*).

4. Extract from the text corpus (by partial parsing) candidate binary word combinations that match the syntactic structure of at least one LF.

5. Compare the meaning of each candidate combination with the centroids of the LFs. If this meaning is sufficiently similar to a centroid, the candidate combination is considered as instance of the corresponding LF, otherwise not.

EuroWordNet as Source of Meaning Descriptions



Learning LF-Characteristic Relations

The EWN-hyperonymy hierarchies provide the componential descriptions of the meaning of the bases and collocates of the LF-instances in the training sets; cf. the FinFunc_0 instance *aprensión se disipa*:

```
((feeling APRENSION2 TEMOR1 RECELO4 PAVOR1
  feeling ESPANTO1 MIEDO1
  feeling Dynamic Mental Experience EMOCION1
  Dynamic Mental Experience SENTIMIENTO1
     Mental Property rasgo-psicologico1)
 (perception DISIPARSE2 DESVANECERSE4
  change Dynamic Location Existence DESAPARECER1 ESFUMARSE1))
```

The componential meaning descriptions allow us to calculate the centroids as figures depending on the relative pairwise co-occurrence frequency of the components in the base and in the collocate descriptions for a given training set.

Calculating the Centroids (Informal)

1. For each pair $< b_i, c_j >$, calculate how often b_i and c_j occur together and how often with other components (this gives the *relative pairwise co-occurrence frequency* of b_i and $c_j >$)

2. Calculate the importance (= the weight) of a pair $< b_i, c_j >$ relative to all other pairs in the descriptions of the instances in a training set

3. For each instance in a training set: sum up the weights of all pairs in its description

4. Calculate the centroid by summing up the total weights of all instances in a training set and dividing the sum by the number of instances in the training set.

Classifi cation Stage (Informal)

1. Extract candidate word combinations from the corpus

2. Acquire the componential descriptions of the elements in each candidate bigram

3. For each LF used in the classification procedure and each candidate bigram:
   (a) Sum up the weights of all pairs that occur in the candidate description to the score of the candidate (the weights have been calculated during the learning stage, see previous slide)
   (b) Compare the calculated score with the centroid of the LF in question; if the difference is smaller than an empirically determined Δ, than the candidate is an instance of this LF.

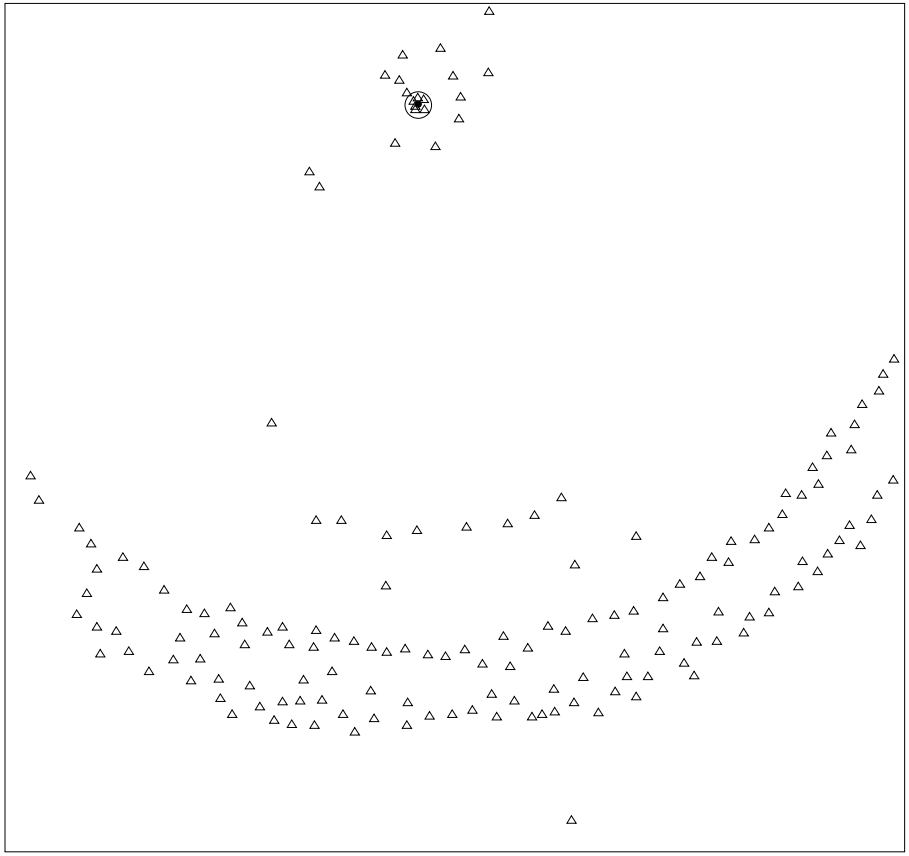
Deviance of the scores of some candidates from the centroids (C_2F_1 = $Caus_2Func_1$, O_1 = $Oper_1$, CO_1 = $ContOper_1$, FF_0 = $FinFunc_0$, IF_1 = $IncepFunc_1$)

	C_2F_1 (15.131)	O_1 (12.743)	CO_1 (3.481)	FF_0 (10.863)	IF_1 (1.337)
[la] admiración cesa	1.170	1.684	5.626	**0.029**	4.599
[la] admiración se desvanece	0.985	1.512	6.153	**0.139**	6.730
guardar admiración	1.297	1.289	**0.225**	1.344	1.306
sentir admiración	1.027	**0.004**	3.909	1.925	8.964
experimentar alegria	1.021	**0.013**	3.935	1.941	8.984
tener alegria	0.980	**0.108**	8.679	1.282	6.940
producir [una ADJ] decepción	**0.133**	2.175	5.218	1.780	9.987
[la] desesperación se apodera [de N]	1.089	1.448	1.519	1.180	**0.539**
provocar [una ADJ] desesperación	**0.470**	2.596	6.994	2.144	15.403
causar enfado	**0.910**	1.254	1.753	1.162	2.532
causar entusiasmo	**0.169**	2.208	5.188	1.676	9.503
conservar [la] esperanza	1.355	1.346	**0.322**	1.406	1.319
[la] esperanza se desvanece	0.968	1.556	6.753	**0.018**	7.407
dar horror	**0.074**	2.283	5.714	1.819	11.933

Oper$_1$-classification (with the initial centroid of 3.041775); D_{O_1} stands for 'deviance from the centroid of Oper$_1$'

candidate	D_{O_1}	candidate	D_{O_1}
poner reparos	**0.074**	presentar [una] demanda	**0.503**
soltar [una] indirecta	**0.100**	preparar [una] lista	0.507
sellar [un] compromiso	0.103	establecer [un] record	0.508
sacar apuntes	**0.113**	cometer [una] falta	**0.534**
celebrar elecciones	**0.116**	cumplir [un] deber	0.602
poner [una] queja	**0.210**	recibir [una] visita	0.633
reconocer acusación	0.305	hacer [un] examen	0.696
lanzar [un] grito	**0.378**	sacar [una] canción	**0.764**
acudir [a una] cita	0.399	imponer [el] orden	0.808
hacer [un] comentario	**0.405**	llevar [un] diario	0.837
pronunciar dictar [la] sentencia	**0.434**	organizar [una] manifestación	0.845
sacar [una] novela	0.447	provocar [una] discusión	0.882
responder [a una] objeción	0.449	dar alarma	**0.923**
promover [una] campaña	0.454	rendirse [al] chantaje	0.928
dar [una] explicación	**0.492**	dar [un] grito	**0.945**

Graphical illustration of the distance between the candidate scores and the centroid of $Oper_1$ in the field of emotion nouns:



Summary and Further Work

Summary:

- So far, two experiments with verb-noun LFs have been carried out: (1) classification of emotion noun-verb candidate bigrams; (2) classification of field-independent noun-verb candidate bigrams.
- In total, ten LFs have been used.

Current and Future Work:

- More extensive experiments
- Extension to all types of LFs (currently noun-adjective LFs)
- Experiments with different Machine Learning Techniques
- Application to the IULA corpus; automatic construction of a collocational dictionary in the IULA domain?

Why Is Automatic Recognition of LFs Important?

For instance:

- Automatic Compilation of Collocational Dictionaries
- Information Retrieval (Collaboration with Prof. M. Alonso Ramos, La Coruña):
  - more accurate indexing of documents,
  - semantic expansion of queries
  - higher quality retrieval
- Information Extraction
  - higher quality information extraction due to the possibility to restrict the search by specific collocation patterns (LFs)