New tools for the encoding of lexical data extracted from corpus

Núria Bel[1], Sergio Espeja[2], Montserrat Marimon[3]

IULA – Universitat Pompeu Fabra
La Rambla 30-32
08002 – Barcelona
{nuria.bel,sergio.espeja,montserrat.marimon} at upf.edu

Abstract

This paper describes the methodology and tools that are the basis of our platform AAILE.[4] AAILE has been built for supplying those working in the construction of lexicons for syntactic parsing with more efficient ways of visualizing and analyzing data extracted from corpus. The platform offers support using techniques such as similarity measures, clustering and pattern classification.

1. Introduction

The encoding of lexical units for a computational grammar is a complex task. It is difficult for humans because it demands to be both exhaustive and consistent. With large corpora available for lexicographers to look into real data, one could think that the exhaustiveness problem, at least, was solved. Different tools and platforms, mainly based on concordances, have been offered to lexicon developers to inspect, sum up and select data. But, no human can easily cope with the direct examination of, for instance, the 2,988 occurrences that an adjective like *clinical* has in the 3.7 million word corpus on medicine.

We present here a set of new tools, the AAILE platform, devised specifically for lexicographers working on the encoding of computational lexica. These tools allow, on the one hand, the easy inspection and visualization of the syntactic information contained in concordances, and, on the other hand, the computation of specific comparison and prediction functions. These functions are a further attempt to contribute to the solution of the problem of lexical coverage of deep analysis computational grammars (Marimon & Bel 2004).

The basis for AAILE tools is the mapping of the textual contexts into vector spaces such that syntactic information can be handled quantitatively. This projection is the necessary step for computing similarity measures, applying clustering techniques, and predicting the assignment of syntactic features by pattern classification techniques.

Furthermore, a vector representation also becomes a complete and compact representation for human inspection of linguistic data. To have the means to visualize actual syntactic contexts in a more compact way than the commonly used made of word and tag strings facilitates a quick verification of the lexicographer expectations.

The paper is organized as follows. Section 2 presents the process for mapping textual concordances into 'bag of features' vectors. Section 3 presents how AAILE displays data. First, we supply a short overview of the compact representation, where the feature vectors are just listed offering a summary of the data. Later, the graphical view is introduced, together with an explanation of how a similarity measure based on the *cosine distance* between two vectors is used for plotting vectors in a 2-D space. Section 4 presents how the mapping of textual information into such a mathematical object allows for further manipulation and inference of new linguistic information. This information, clustering and confidence measures, is supplied to lexicographers to assist them in the process of encoding lexical entries. The following section briefly accounts for the intended use of AAILE platform. And, finally Section 6 supplies with the most technical details of the platform.

2. Vectors representing word occurrences

Vector spaces for representing textual data were first used in Information Retrieval. Vector representation is very convenient as it allows the use of mathematical techniques, in our case measuring similarity between different instances, clustering of occurrences and Bayes based pattern classification, as we will see below.

In order to represent word occurrences as vectors, we use Regular Expressions (RE's) that search for local syntactic information –sequences of tags– in a part of speech tagged corpus. In designing the particular RE's, we follow linguistic criteria to identify those linguistic cues that are considered to play a role on the characterization of the syntactic properties of the grammatical category –part of speech– under consideration.

Different RE's are used to check whether a number *n* of particular cues are found in each occurrence of a word in a corpus. The positive or negative results of each checking are stored as binary values of *n* dimension vectors. Thus, we create a 'bag of features' representation (Rosenfeld, 1997) for each occurrence.

Nevertheless, in order to have a complete picture of a word's behavior, one has to see the characteristics of all its occurrences in a representative corpus (Bel, 2004) to judge whether certain context is more or less frequent,

[1] *Ramón y Cajal* Program of the Spanish Ministerio de Educación y Ciencia.

[2] Supported by the Spanish Ministerio de Educación y Ciencia and *Fondo Social Europeo*, PTA-CTE/1370/2003

[3] *Juan de la Cierva* Program of the Spanish Ministerio de Educación y Ciencia.

[4] AAILE stands for *Automatic Acquisition of Lexical Information*, the project funded by the Spanish Ministerio de Educación y Ciencia (HUM2004-05111-C02-01/FILO) that has partially supported the development of the work presented here.

whether there are particular cues that are constant in more or less occurrences, etc.

AAILE's aim is to help the lexicographer to generalize and abstract from the characteristics of particular occurrences into the characteristics of the whole set of occurrences. A word's *signature* σ is the set of vectors resulting of the transformation from all its occurrences in a corpus and it represents the complete syntactic behavior of that particular word, more technically of that lemma-category pair.

Table 1 shows the *signature* of the adjective *adjacent*, the vectors resulting of applying 14 RE's to every occurrence of this adjective in a corpus of 5 million words (*Corpus Tècnic de l'IULA,* Cabré et al. 2000). Each RE checks whether a particular context is displayed in the occurrence. All occurrences are inspected by the whole collection of RE's.

#occ.	lc1	lc2	lc3	lc4	lc5	lc6	lc7	lc8	lc9	lc10	lc11	lc12	lc13	lc14
124	1	0	0	0	0	0	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0	0	0	1	0
23	0	1	0	0	0	0	0	0	0	0	0	0	1	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	1	0	0	0	0	0	0	0	0	0	1	0
1	0	0	1	0	0	0	0	0	0	1	0	0	0	0
1	1	0	1	0	0	0	0	0	0	0	0	0	0	0

Table 1. Signature for the English adjective *adjacent*. The RE's have check for the following linguistic cues: lc_1, whether the item is located before a noun; lc_2, whether the item is located after a noun; lc_3, if the item is located after a copulative verb; lc_4 after a gradual adverb, and from lc_5 to lc_{14}, whether the item is before a particular preposition.

For the example shown in Table 1, the RE's have check for the following linguistic cues: lc_1 whether the item is located before a noun; lc_2 whether the item is located after a noun; lc_3 if the item is located after a copulative verb; lc_4 after a gradual adverb and lc_5- lc_{14} whether the item is located before a particular preposition.

Additionally, the first column in Table 1 refers to the number of times that the resulting vectors turn to be the same. We will call *profile* to a particular vector configuration when it is the very same results of the fourteen RE's checking for different occurrences. Hence, the first column refers to the absolute frequency of a *profile* in a *signature*.

In the example of Table 1, the most frequent *profile* (124) corresponds to the occurrences where the only observed cue is its being in pre-nominal position, marked in column lc_1. The positive values in columns lc_2 and lc_3 show that in other *profiles* there were occurrences displaying post-nominal and predicative positions, respectively. And lc_{13} shows the number of times that this adjective was found immediately before the preposition *to*. The invariance of one of the linguistic cues across different profiles will be crucially used when computing the assignment of syntactic features to *signatures*. In short, we take for granted that it is very likely that the preposition *to*, which appears in three of the seven profiles, has to be considered a bound preposition of that adjective.

Table 2 and Table 3 show, for the same linguistic cues than in Table 1, the signatures of two other adjectives, *countless* and *adhesive*, that display different results. For instance, *adhesive* has been found in a predicative position, but not *countless*.

#occ.	lc1	lc2	lc3	lc4	lc5	lc6	lc7	lc8	lc9	lc10	lc11	lc12	lc13	lc14
16	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2. Signature for English adjective *countless*. Columns refer to the same information than in Table 1.

#occ.	lc1	lc2	lc3	lc4	lc5	lc6	lc7	lc8	lc9	lc10	lc11	lc12	lc13	lc14
16	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	1	0	0	0	0	0	0	0	0	0	0	0

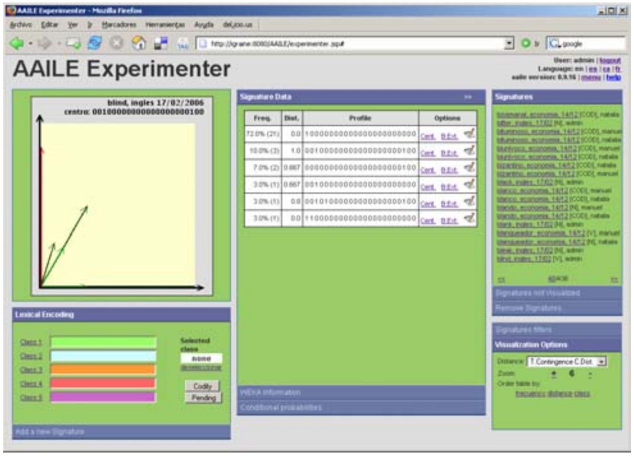
Table 3. Signature for English adjective *adhesive*. Columns refer to the same information than in Table 1.

A *signature*, that is, the set of vectors that represents all occurrences of a particular word in a corpus, is the input for the AAILE platform.

AAILE data acquisition module permits the introduction of signatures either by using web services that directly consult the corpus, or by uploading plain files. When uploaded into the platform, a *signature* also contains other information: the lemma and part of speech the language and the identification of the corpus where concordances were extracted.

3. Vector visualization in AAILE

Once the textual data are translated into a vector space, the information can be displayed, analyzed and interpreted in the most suitable ways for supporting lexicographer's work. Figure 1 shows AAILE's main view where most of the information is displayed.



AAILE Experimenter

blind, ingles 17/02/2006
centro: 0010000000000000000100

Freq.	Dist.	Profile	Options
72.0% (21)	0.0	1000000000000000000000	Cent. B.Ext.
10.0% (3)	1.0	0010000000000000000100	Cent. B.Ext.
7.0% (2)	0.667	0000000000000000000100	Cent. B.Ext.
3.0% (1)	0.667	0010000000000000000000	Cent. B.Ext.
3.0% (1)	0.0	0010100000000000000100	Cent. B.Ext.
3.0% (1)	0.0	1100000000000000000000	Cent. B.Ext.

Distance: T.Contingence C.Dist.

Figure 1: Main window of AAILE's encoding page. At the left-upper corner, the display shows the different vectors with a representation of its similarity (as calculated by *cosine distance*). The centre of the window shows the *signature*, the set of vectors that represents the occurrences of a particular word, in this case the adjective *blind*. The right column contains visualization choices (zoom, ordering criteria, etc.) and other administrative information.

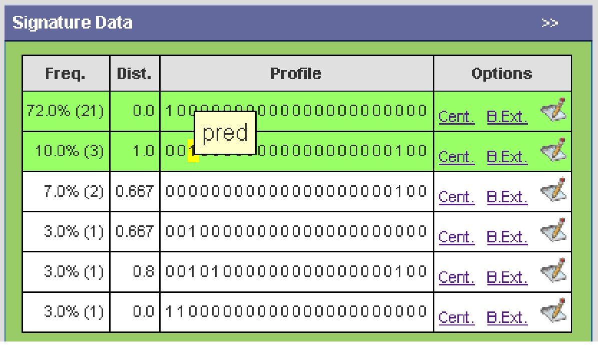
A graphical view complements a compact format view where the results of *linguistic cue* checking are displayed, as we will explain below in more detail.

3.1. AAILE compact format

AAILE's goal is to assist lexicographers in the analysis of corpus data for deciding on the encoding of lexical items according to a particular linguistic classification. AAILE wants to offer an easy inspecting and compact view of the linguistic cues that the lexicographer wants to take into account for assigning such classes.

For example, *adjacent* should be given the class *transitive* because this class is the one assigned to adjectives that show the following syntactic characteristics: (i) they are to be found in pre-nominal, post-nominal and predicative positions, and (ii) they have a prepositional complement, in this particular case, headed by *to*. This is the information that the *signature* of *adjacent* should contain and that must show up clearly in the compact vector representation.[5]

Thus, in the AAILE platform, *signatures* are shown in tables that display the *profiles* of the vectors they are made of. Each profile is informed with its absolute and relative frequency information, as shown in Figure 2.



Signature Data

Freq.	Dist.	Profile	Options
72.0% (21)	0.0	10000000000000000000000	Cent. B.Ext.
10.0% (3)	1.0	001[illegible]00000000000000100	Cent. B.Ext.
7.0% (2)	0.667	00000000000000000000100	Cent. B.Ext.
3.0% (1)	0.667	00100000000000000000000	Cent. B.Ext.
3.0% (1)	0.8	00101000000000000000100	Cent. B.Ext.
3.0% (1)	0.0	11000000000000000000000	Cent. B.Ext.

Figure 2: Profile information for the English adjective *blind*. The keyword referring to the linguistic cue value can be retrieved by clicking on each component value

As shown in Table 2, *profile* components are located at the centre of the table. For easy interpreting the different components of each one, the keyword referring to the linguistic cues can be retrieved by clicking on each components value.

At the left hand side column of the *profile* compact representation, there is the information about similarity between different profiles. It offers a quantitative measure of the number of *linguistic cues*, that is components, shared by two vectors. In the example of Figure 2, **Dist** values show that the most similar *profiles* share the component referring to the predicative position and the presence of the preposition *to*. We will describe in detail the functionalities regarding similarity calculations in the following section.

The most left column, under the header **Freq,** refers to the absolute (in parenthesis) and relative frequency of profiles in the *signature*.

3.2. Graphical format

For the graphical representation of a *signature*, as shown in Figure 3, AAILE takes as reference the most frequent *profile* and computes the similarity of the other profiles with respect to it. Using *cosine distance*[6] as the similarity measure allows the plotting of different profiles in a 2-dimensional graphic.

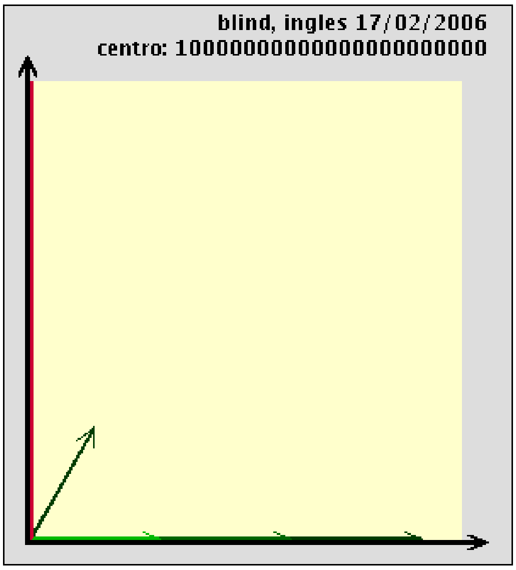


Figure 3: *blind* 2D graphic using its most frequent profile as center.

By default, the most frequent *profile* is a red arrow on the *y* axis at the graphic display. The *x* axis is for the profiles whose cosine distance to the one chosen for reference is 0. The profiles that are similar in some measure to the reference are plotted accordingly. Arrow's length represents frequency information. In case of resulting in equal similarity measures, the profiles are plotted sequentially, but in different green colored vectors.

The user can change the *profile* used as point of reference. the one shown in *y* axis. Thus, by clicking at the *Cent* sign (in the compact view) of another profile, the whole is re-calculated and plotted again. In Figure 4 we show the graphical display for the signature of *blind* again, but with the profile showing a pre-nominal location as the point of reference for similarity calculations. A comparison of the graphic displays allows to see quickly that there is more variance in the case of the predicative position.

In Figure 2 we can see that some invariant positive components, such as the one referring to its being after a predicative component and appearing together with the preposition *to*, indicates a consistent pattern that should be

[5] The examples shown until now are simple cases where the relation between the linguistic cues and the class assigned is almost straightforward. This is the case for adjectives, which have a very informative local context. For other grammatical categories, linguistic cues are less certain and specific combination of them have to be devised and put in relation with the linguistic classes in a *n* to 1 relation.

[6] The closer to 0, the more significantly different, the closer to 1, the more similar.

taken into account. We will refer again to that observation in the following section.

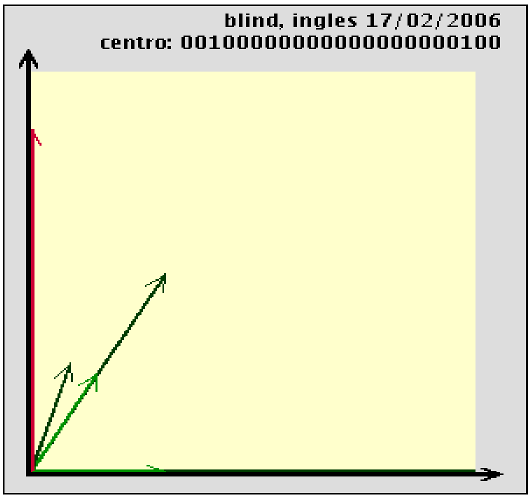


Figure 4: *blind* 2D graphic using *pred* profile as center.

4. Additional information in AAILE

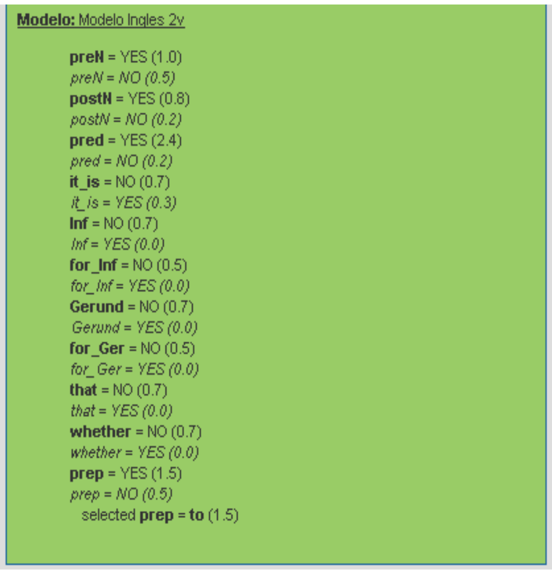
4.1. Syntactic Features assessment

AAILE incorporates a *Knowledge Module* (KM) that evaluates the syntactic behavior displayed in a word's *signature* on the light of the possible syntactic features it can be said to hold. By *syntactic features* we refer to the linguistic generalizations that can be drawn over the distributional characteristics a part of speech, that is, the potential contexts where a word can appear according to its grammatical category. This distributional information is the basis for deciding on the encoding of such a word, the goal of the lexicographer who has to compare empirical evidences, the *linguistic cues* with the *syntactic features* that a given class is supposed to display.

The quantitative information supplied by AAILE is intended to help the lexicographer to assess the degree of confidence that the information brought by *linguistic cues* deserves. It is beyond the scope of this paper to present how the *KM* works in detail, but we will explain now how it proceeds in a very intuitive way.

KM module calculates, on the basis of the *linguistic cues* captured, to what extent the behavior displayed in the *signature* corresponds to the one expected if the word could be said to hold a particular *syntactic feature*. For instance, the system can assess that the adjective *blind*, that displays the behavior reported in Figure 2, can be said to be *predicative* with enough confidence because there were *linguistic cues* confirming that feature. However, for the adjective *countless*, following the data in Table 2, should be considered *no predicative*, there are no cues that support such a consideration for *countless*.

In Figure 5, we show the assessment of all *linguistic cues* for the adjective *blind* as currently displayed in AAILE. We see how *KM* returns a value for each *linguistic cue* and possible value *yes* and *no*. The indicator, that computes the conditional probability of the *linguistic cues* ones depending on the others, is also sensitive to the invariant cues in different *profiles*. This means that the presence in different *profiles* raises the level of confidence on this *linguistic cue*.



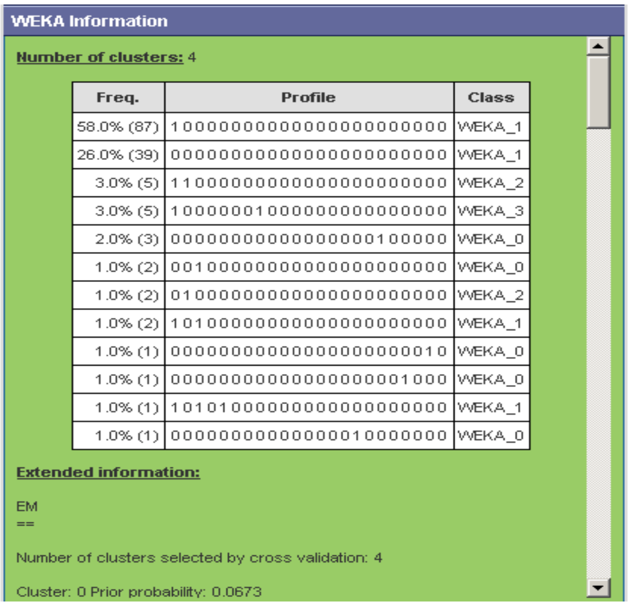
Modelo: Modelo Ingles 2v

preN = YES (1.0)
preN = NO (0.5)
postN = YES (0.8)
postN = NO (0.2)
pred = YES (2.4)
pred = NO (0.2)
it_is = NO (0.7)
it_is = YES (0.3)
Inf = NO (0.7)
Inf = YES (0.0)
for_Inf = NO (0.5)
for_Inf = YES (0.0)
Gerund = NO (0.7)
Gerund = YES (0.0)
for_Ger = NO (0.5)
for_Ger = YES (0.0)
that = NO (0.7)
that = YES (0.0)
whether = NO (0.7)
whether = YES (0.0)
prep = YES (1.5)
prep = NO (0.5)
selected **prep** = **to** (1.5)

Figure 5. Confidence on *linguistic cues*

4.2. Clustering with WEKA

AAILE has been provided with an interface to WEKA libraries (Witten, Frank, 2005.) It uses the Expectation Maximization (EM) algorithm a clustering analysis of the signatures on user demands. Figure 6 shows the results of a clustering exercise.



WEKA Information

Number of clusters: 4

Freq.	Profile	Class
58.0% (87)	10000000000000000000000	WEKA_1
26.0% (39)	00000000000000000000000	WEKA_1
3.0% (5)	11000000000000000000000	WEKA_2
3.0% (5)	10000001000000000000000	WEKA_3
2.0% (3)	00000000000000000100000	WEKA_0
1.0% (2)	00100000000000000000000	WEKA_0
1.0% (2)	01000000000000000000000	WEKA_2
1.0% (2)	10100000000000000000000	WEKA_1
1.0% (1)	00000000000000000000010	WEKA_0
1.0% (1)	00000000000000000001000	WEKA_0
1.0% (1)	10101000000000000000000	WEKA_1
1.0% (1)	00000000000000010000000	WEKA_0

Extended information:

```
EM
==

Number of clusters selected by cross validation: 4

Cluster: 0 Prior probability: 0.0673
```

Figure 6: Clustering with WEKA EM.

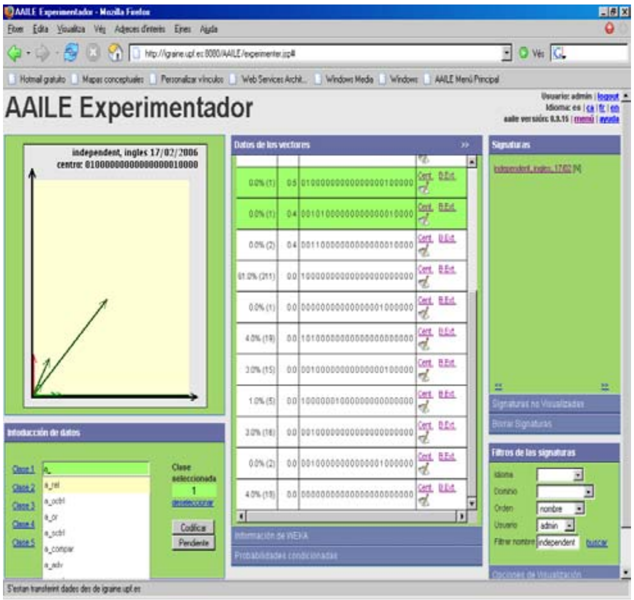
When inspecting the clustering results, it becomes clear that groups are made according to the quantity of information in the profile of linguistic cues and the number of vectors with the same profile.

5. Using AAILE platform for lexical encoding

Along the preceding sections, we have provided details on how the system has been designed to assist the lexicographer when taking decisions about what are the relevant characteristics of a word. We should see now how is the intended use of the platform.

The lexicographer using AAILE has been asked to relate lemmas belonging to a particular grammatical category and the lexical classes defined on linguistic grounds, for instance for the use by Natural Processing Tools (Lenci et al. 2000.) The lexicographer should, instead of looking at texts, receive enough analytical information to determine the class of the word under study.

A typical encoding exercise is based on the observation of the *profiles* that conform a *signature*, their frequency and the selection of profiles that correspond to a given class.


AAILE Experimentador - Mozilla Firefox
Fitxer Edita Visualitza Vés Adreces d'interès Eines Ajuda
http://igraine.upf.es:8080/AAILE/experimenter.jsp#
Hotmail gratuito Mapas conceptuales Personalizar vínculos Web Services Archit... Windows Media Windows AAILE Menú Principal
AAILE Experimentador
Datos de los vectores
Cent. B.Ext.
Signaturas
Signaturas no Visualizadas
Borrar Signaturas
Filtros de las signaturas
Introducción de datos
Clase 1
Clase 2
Clase 3
Clase 4
Clase 5
deseleccionar
Codificar
Pendiente
Información de WEKA
Probabilidades condicionadas
Opciones de Visualización

Usuario: admin | logout
Idioma: es | ca | fr | en
aaile versión: 0.3.15 | menú | ayuda

independent, ingles 17/02/2006
centro: 0100000000000000000010000

0.0% (1)	0.5	0100000000000000000100000	Cent. B.Ext.
0.0% (1)	0.4	0010100000000000000010000	Cent. B.Ext.
0.0% (2)	0.4	0011000000000000000010000	Cent. B.Ext.
81.0% (211)	0.0	1000000000000000000000000	Cent. B.Ext.
0.0% (1)	0.0	0000000000000000001000000	Cent. B.Ext.
4.0% (19)	0.0	1010000000000000000000000	Cent. B.Ext.
3.0% (15)	0.0	0010000000000000000100000	Cent. B.Ext.
1.0% (5)	0.0	1000000100000000000000000	Cent. B.Ext.
3.0% (18)	0.0	0010000000000000000000000	Cent. B.Ext.
0.0% (2)	0.0	0010000000000000001000000	Cent. B.Ext.
4.0% (18)	0.0	0000000000000000000000000	Cent. B.Ext.

independent, ingles, 17/02 [M]

a_
a_rel
a_octrl
a_or
a_sctrl
a_compar
a_adv

Clase seleccionada: 1

Idioma:
Dominio:
Orden: nombre
Usuario: admin
Filtrar nombre: independent

S'estan transferint dades des de igraine.upf.es

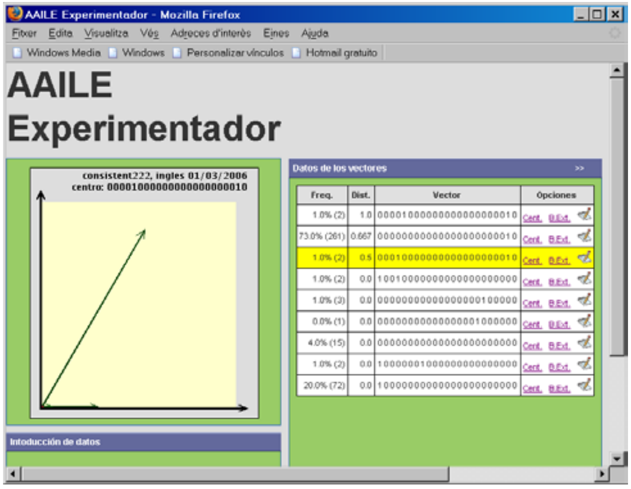
Figure 7: AAILE main window. At the left-bottom corner the lexicographer is assisted with completive writing for consistency encoding

Figure 7 shows, at the bottom left hand corner, the encoding section of the AAILE main view. The lexicographer should use it when entering the class or classes assigned. The system asks the lexicographer to mark (in the Figure, in green color) those profiles that are related to the class selected. The system allows to create two different lexical entries out of a unique signature. This could be the case when encoding different bound prepositions, for instance.

The lexicographer enters the selected lexical class help by AAILE that shows him or her the possible classes with a completive writing facility.

In the most simple case, the encoding exercise should take just few minutes. Analytical information is mostly required when dealing with high frequency words, that present a large number of profiles, with high variation in the *linguistic cues* shown. Then the lexicographer might find useful to check whether there is a particular profile that shows to be a better point of reference because it creates a big distance between two groups of profiles, for instance. This could be a sign that there is more than one lexical entry in the signature.

An example of such a case is the adjective *consistent*, where we could identify the *consistent-with* reading and the *be-consistent* one. Figure 8 shows the main view for *consistent*.


AAILE Experimentador - Mozilla Firefox
Fitxer Edita Visualitza Vés Adreces d'interès Eines Ajuda
Windows Media Windows Personalizar vínculos Hotmail gratuito
AAILE Experimentador
Datos de los vectores
Introducción de datos

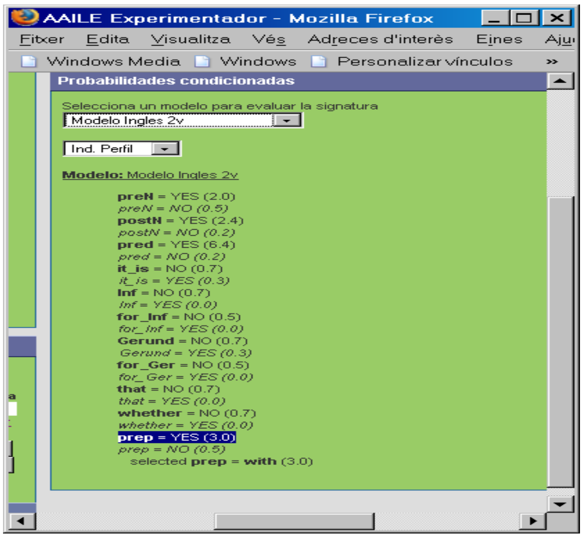
consistent222, ingles 01/03/2006
centro: 0000100000000000000000010

Freq.	Dist.	Vector	Opciones
1.0% (2)	1.0	0000100000000000000000010	Cent. B.Ext.
73.0% (261)	0.667	0000000000000000000000010	Cent. B.Ext.
1.0% (2)	0.5	0001000000000000000000010	Cent. B.Ext.
1.0% (2)	0.0	1001000000000000000000000	Cent. B.Ext.
1.0% (3)	0.0	0000000000000000000100000	Cent. B.Ext.
0.0% (1)	0.0	0000000000000000001000000	Cent. B.Ext.
4.0% (15)	0.0	0000000000000000000000000	Cent. B.Ext.
1.0% (2)	0.0	1000000010000000000000000	Cent. B.Ext.
20.0% (72)	0.0	1000000000000000000000000	Cent. B.Ext.

Figure 8: AAILE view for *consistent*

The most significant information comes from the clean cut between the three profiles with 0.5 distance, and the others going directly to 0. This is a sign that it is worth considering two different entries.

As already mentioned, invariant components across different signatures are very informative, specially for bound prepositions. We see in the example of Figure 8 that *with* is present in 3 profiles, while other prepositions like *in* or *from* are present only in 1. AAILE can also supply the lexicographer with a confidence measure on what are the relevant characteristics of the signature to classify the lexical entry. In the case of *consistent*, it shows a clear confidence on the bound character of preposition _with:


AAILE Experimentador - Mozilla Firefox
Fitxer Edita Visualitza Vés Adreces d'interès Eines Aju
Windows Media Windows Personalizar vínculos
Probabilidades condicionadas

Selecciona un modelo para evaluar la signatura
Modelo Ingles 2v

Ind. Perfil

Modelo: Modelo Ingles 2v

preN = YES (2.0)
preN = NO (0.5)
postN = YES (2.4)
postN = NO (0.2)
pred = YES (6.4)
pred = NO (0.2)
it_is = NO (0.7)
it_is = YES (0.3)
Inf = NO (0.7)
Inf = YES (0.0)
for_Inf = NO (0.5)
for_Inf = YES (0.0)
Gerund = NO (0.7)
Gerund = YES (0.3)
for_Ger = NO (0.5)
for_Ger = YES (0.0)
that = NO (0.7)
that = YES (0.0)
whether = NO (0.7)
whether = YES (0.0)
prep = YES (3.0)
prep = NO (0.5)
selected **prep** = **with** (3.0)

Figure 9: *Linguistic cues* confidence for *consistent*

6. Technical details

AAILE platform is divided into three tiers, following model view controller (MVC) pattern: user system interface, process management and database management. Database management tier is implemented using MySQL 4.1., MySQL JDBC driver and DAO pattern access to data.

Process management is performed in Java 5, and uses various open source libraries: Jakarta Commons DBUtils, Jakarta Commons FileUpload, Jakarta Log4j and Apache Axis 1.2.

The user interface tier is developed using AJAX, Javascript, CSS in the client side and Java Server Pages in the server side.

Logically, the architecture of the system can be divided as follows: A data acquisition and introduction module, a codification module, a visualization module and a knowledge module. These modules use data from a database and a wiki.

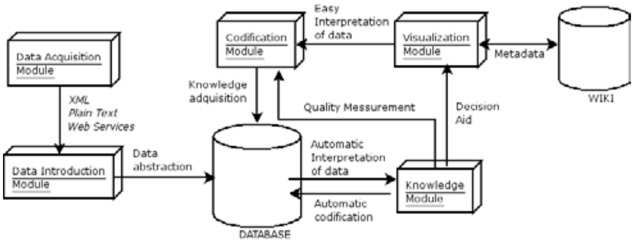


Figure 8: AAILE architecture

7. References

Bel, N. (2004). "Corpus representativeness for syntactic information acquisition" in Blache, Philipe; Rodríguez, Horacio [eds.] *Companion Volume to the Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*. Barcelona: Association for Computational Linguistics. Pàg. 138-141.

Cabré, M. T; Bach, C. (2004). "El Corpus Tècnic del IULA: corpus textual especializado plurilingüe" in *Panace@ - Boletín de Medicina y Traducción* V(16): MedTrad. Pàg. 173-176.

Lenci, A.; Bel, N.; Busa, F.; Calzolari, N.; Gola, E.; Monachini, M.; Ogonowski, A.; Peters, I.; Peters, W.; Ruimy, N.; Villegas, M.; Zampolli, A. (2000). "SIMPLE: A General Framework for the Development of Multilingual Lexicons", *International Journal of Lexicography* 13(4). Oxford: Oxford University. Pàg. 249-263

Marimon, M.; Bel, N. (2004). "Lexical Entry Templates for Robust Deep Parsing" in *LREC 2004 Fourth International Conference on Language Resources and Evaluation*. Lisboa: European Languages Resources Association

Rosenfeld, R. (1997). A whole sentence maximum entropy language model. In Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding

Witten, Ian H. and Eibe Frank (2005). Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann, San Francisco.