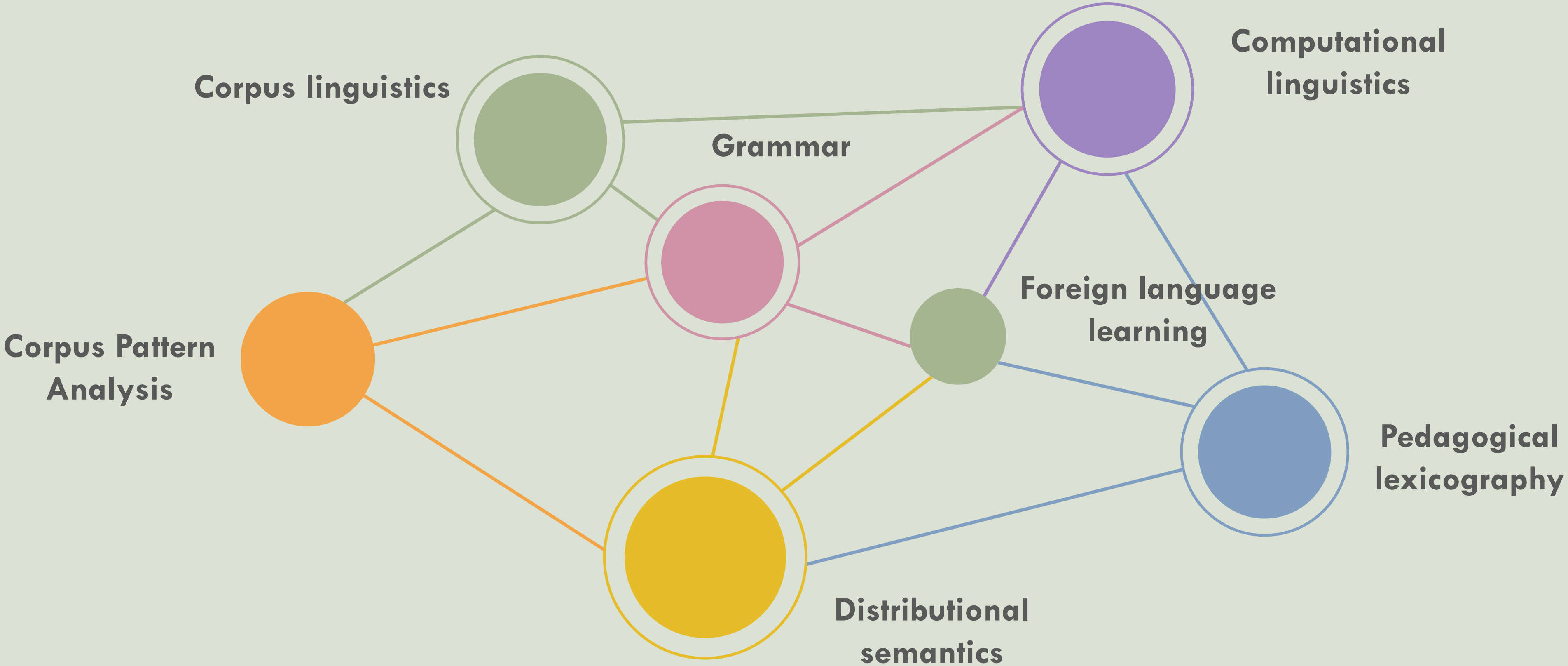


VERBARIO
A DATABASE OF SPANISH VERBS AS SOURCE FOR PEDAGOGICAL LEXICOGRAPHY

Irene Renau & Rogelio Nazar
Pontificia Universidad Católica de Valparaíso

PhrasaLex | 19-20 September 2019 | Univ. degli Studi di Modena

OUTLINE

Starting points

Introduction to the Verbario project

Theoretical background

Applying CPA to Spanish verbs

Towards a pedagogical dictionary based on Verbario’s data

Final remarks

STARTING POINTS

The data problem

- Lack of empirical data
- Difficulty to systematically and carefully analyse corpus data

The lexicographic problem

- Conversion of empirical data into understandable, pedagogical, user-centered… data
- There is still room for reflection about the model of pedagogical dictionary that we want, e.g. monolingual / bilingual / semi-bilingual…

(Cf. Atkins & Rundell 2008 for the steps of dictionary making)

STARTING POINTS

The data problem

- Lack of empirical data
- Difficulty to systematically and carefully analyse corpus data

The lexicographic problem

- Conversion of empirical data into understandable, pedagogical, user-centered… data
- There is still room for reflection about the model of pedagogical dictionary that we want, e.g. monolingual / bilingual / semi-bilingual…

(Cf. Atkins & Rundell 2008 for the steps of dictionary making)

The Verbario project adresses the data problem and (in progress) is trying to be compatible with lexicographic projects.

THEORETICAL BACKGROUND

“The meaning of a single word is to a very high degree dependent on its **context**” (Malinowski, 1923: 306).

“The complete meaning of a word is always **contextual**, and no study of meaning apart from a complete context can be taken seriously” (Firth, 1935: 37).

“The opportunity to observe **recurrent patterns of language in corpora** has shown how choices at word rank co-ordinate with other choices round about in an intricate fashion, suggesting a hierarchy of units of different sizes sharing the **realization of meaning**” (Sinclair, 2004: 140).

“It is reasonable to assume that in the everyday use of language, meanings are events, not entities […]. It is a convenient shorthand to talk about ‘the meanings of a word in a dictionary,’ but strictly speaking these are not meanings at all. Rather, they are **meaning potentials**” (Hanks, 2013: 73).

THEORETICAL BACKGROUND

The Theory of Norms and Exploitations (Hanks 2004, 2013) postulates that word meanings are linked to patterns of usage of verbs in context.

Normal patterns of usage are syntagmatic structures which can be identified by the argument structure and by the semantic types of the arguments.

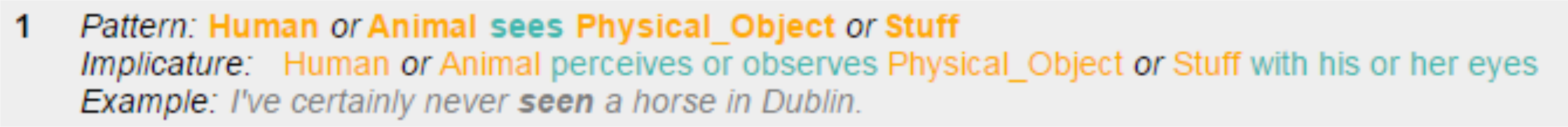
In contrast with norms, exploitations are creative, non conventional uses of patterns.

The lexicon works as a double helix system of norms and exploitations.

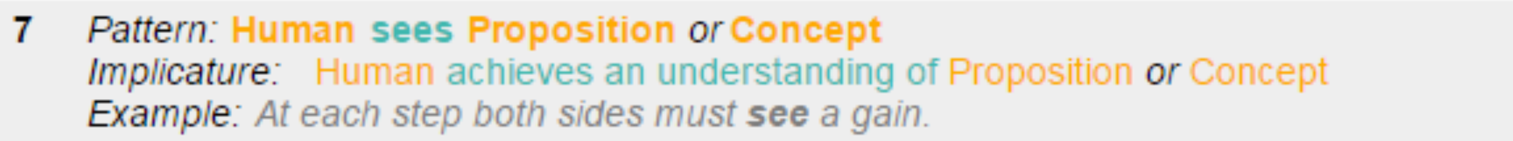
CPA is the technique for detecting these normal patterns and exploitations in corpora.

APPLYING CPA TO SPANISH VERBS

Example from the *Pattern Dictionary of English Verbs* (PDEV), Hanks, in progress (http://www.pdev.org.uk):



1 *Pattern:* **Human** *or* **Animal** **sees** **Physical_Object** *or* **Stuff**
Implicature: Human *or* Animal perceives or observes Physical_Object *or* Stuff with his or her eyes
Example: *I've certainly never* ***seen*** *a horse in Dublin.*



7 *Pattern:* **Human** **sees** **Proposition** *or* **Concept**
Implicature: Human achieves an understanding of Proposition *or* Concept
Example: *At each step both sides must* ***see*** *a gain.*

Key points:

- Taking into account the syntactic / argument structure
- A semantic analysis of arguments
- Differentiation between norm and exploitation when analysing corpus data

The Verbario project replicates the exact same procedure than PDEV project.

APPLYING CPA TO SPANISH VERBS

01 **Creation of corpus sample**

Around 200 (iterative sampling)

02 **Annotation of each concordance in the sample**

- Semantic labelling of arguments
- SPOCA analysis
- Detection of exploitation (abnormal arguments, syntactic structures, collocations, etc.

03 **Creation of patterns**

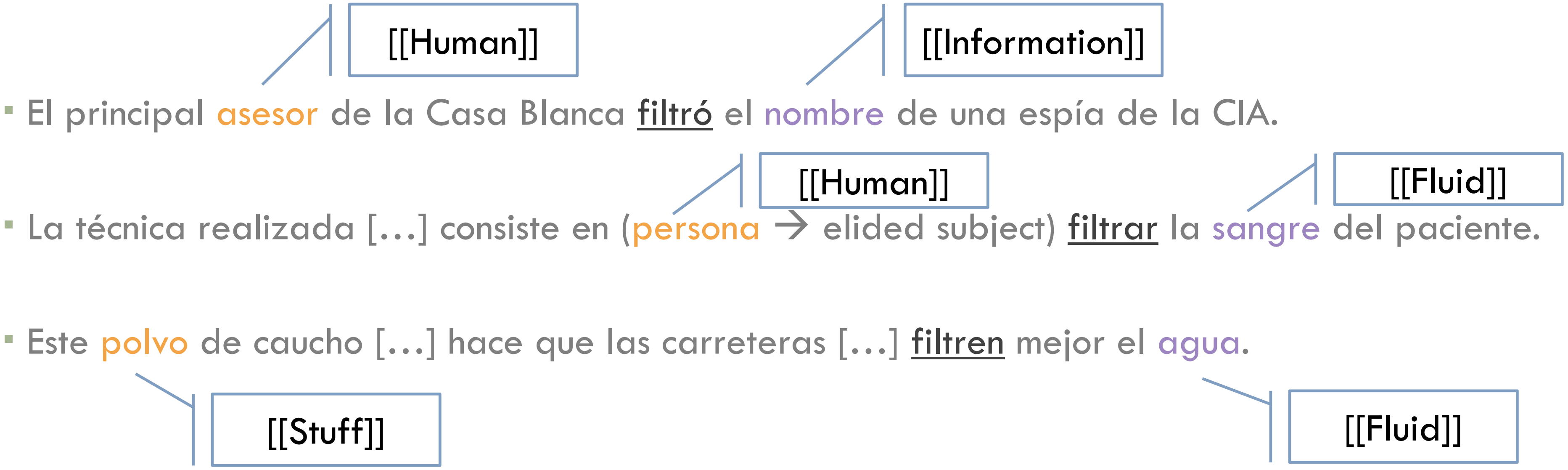
A synthesis of the analyzed sample(s).

04 **Writing the implicatures**

Paraphrases of the patterns (similar to definitions).

APPLYING CPA TO SPANISH VERBS

Example: verb filtrar (= filter)



[[Human]] [[Information]]

- El principal asesor de la Casa Blanca filtró el nombre de una espía de la CIA.

[[Human]] [[Fluid]]

- La técnica realizada […] consiste en (persona → elided subject) filtrar la sangre del paciente.

- Este polvo de caucho […] hace que las carreteras […] filtren mejor el agua.

[[Stuff]] [[Fluid]]

APPLYING CPA TO SPANISH VERBS

Example: verb filtrar (= filter)

🟩 **PATTERN 1** [[Human]] filtrar [[Information]]

[[Human]] makes public a confidential or secret [[Information]].

El principal asesor de la Casa Blanca filtró el nombre de una espía de la CIA.

🟧 **PATTERN 2** [[Human]] filtrar [[Fluid]] ({con [[Physical Object | Stuff]]})

[[Human]] makes [[Fluid]] go through [[Physical Object | Stuff]]
to clean it or separate it from other [[Stuff]].

La técnica realizada […] consiste en filtrar la sangre del paciente.

🟨 **PATTERN 3** [[Physical Object]] filtrar [[Fluid]]

[[Fluid]] pass through [[Physical Object]].

Este polvo de caucho […] hace que las carreteras […] filtren mejor el agua.

APPLYING CPA TO SPANISH VERBS

Verbario database today (work in progress!):

- 228 verbs analysed
- 1.233 patterns
  - mean: 5,41 patterns per verb
- 84.227
  - mean: 369,42 concordances per verb
  - 68,31 concordances per pattern

A lot of work! ☺
Verbario project has tried an automatic procedure which can help with part of this process.
Basic idea: replicate the manual procedure using corpus statistics.

See
www.verbario.com

for the online version of the database

APPLYING CPA TO SPANISH VERBS

1 **Ontology building**

We use CPA Ontology for the most general nodes of the taxonomy, and connect nouns with these nodes.

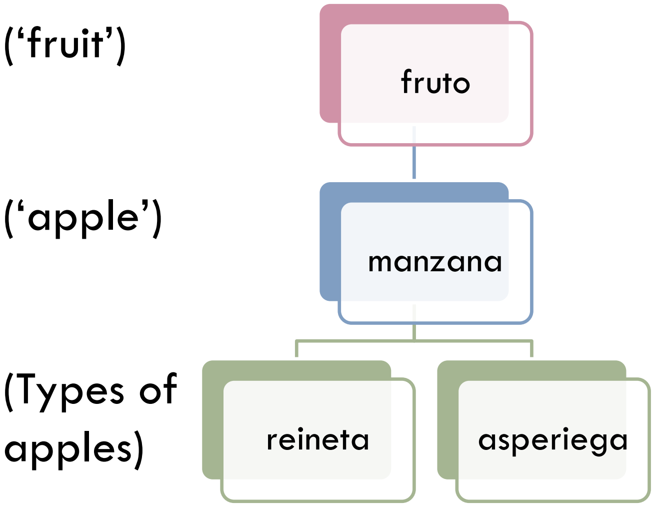
We apply a set of algorithms to different general corpora such as a press corpus, Wikipedia, Google Books Ngrams Corpus, etc.

English and French in progress!

This part of Verbario has been created as a separate project: Kind (aka The Taxonomy project): see www.tecling.com/kind (Nazar & Renau, 2016, LREC proceedings)

APPLYING CPA TO SPANISH VERBS

1 Ontology building



- **Alg 1: Dicco/Castellón:** hypernymy relations from definiens-definiendum co-occurrence in multiple dictionary definitions (Renau & Nazar, 2012).
- **Alg 2: Distrsimi/Getafe:** paradigmatic similarity between words. If *gouda* is distributionaly similar to other words of the *cheese* category, then *gouda* is a type of *cheese* (Nazar & Renau 2013).
- **Alg 3: Asimi/Oslo:** Co-occurrence graphs. Analysis of asymmetric co-occurrence: *bicycle* appears with *vehicle* but the relation is not reciprocal (Nazar & Renau 2012).
- **Alg 4: Morfsimi:** orthographic similarity of affixes (sequences of initial and final letters): it learns to associate affixes with semantic categories (Nazar & Renau, 2016).
- **Alg 5: Final decision:** for each input noun, each algorithm proposes a list of 10 hypernymy candidates and they are selected by weighted voting (Oslo’s vote counts more because it is the best classifier).

APPLYING CPA TO SPANISH VERBS

2 Creation of the patterns

El principal asesor **[[Human]]** *de la Casa Blanca filtró el nombre* **[[Information]]** *de una espía de la CIA.* →
[[Human]] filtrar [[Información]]

For each analysed verb (e.g. *acosar* ‘to harass’), do:

- Extract all sentences in the corpus where the verb occurs.
- Replace in all possible cases the nouns that appear in the sentences by their corresponding semantic types using the taxonomy. E.g.: *problema, dificultad, peligro* are replaced with the semantic type [[Eventuality]], and *hombres, mujer, habitantes,* with [[Human]].
- Detect the prepositions used after the verb.
- Build a data structure mapping each pattern with its example-sentences.
- Detect patterns with pronominal uses of the verb (in Spanish this is marked by the pronominal paradigm *me, te, se, nos, os, se* or by the enclitic).
- Sort the patterns by decreasing order of frequency.

APPLYING CPA TO SPANISH VERBS

Current state of the automatic part of Vebario:

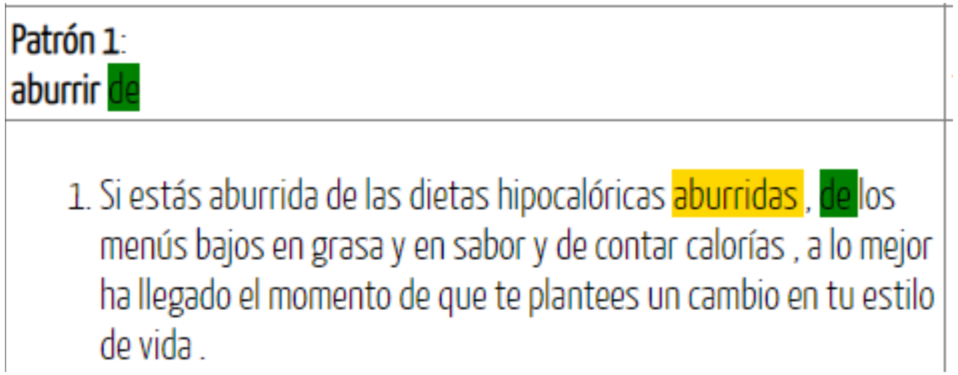
- It is useful for a first guide for the lexicographer, BUT it still does not detect *patterns,* but 'pseudo-patterns' = fragments of patterns (e.g. verb + preposition).
- It detects semantic neology (= new meanings that were never registered in dictionaries before)
- The ontology part detects lexical neology (= new words)
- We still have a lot of work to do with precision!

(Renau, Nazar et al., 2019, *Signos* journal)

→ See www.verbario.com for the automatic attempt.

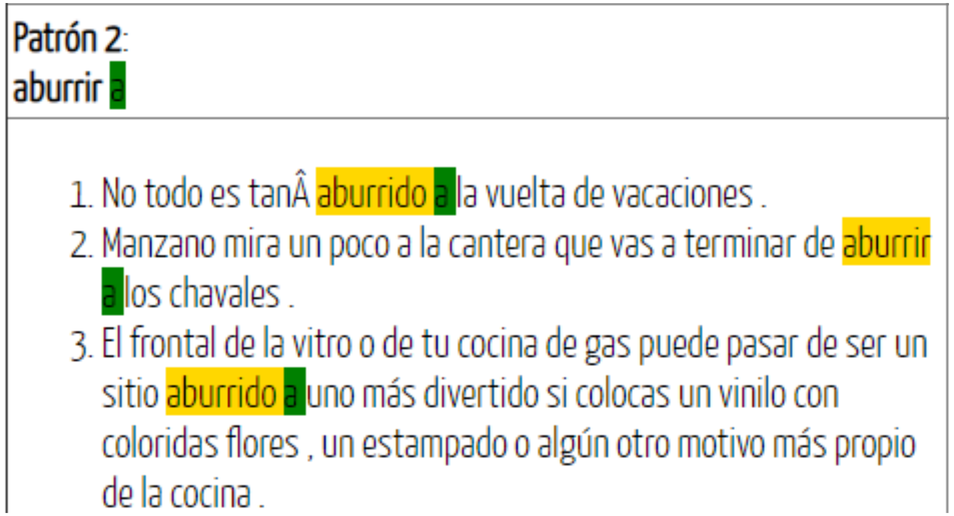
APPLYING CPA TO SPANISH VERBS

An example of results with the automatic part of Verbario (verb *aburrir* ‘to bore’):



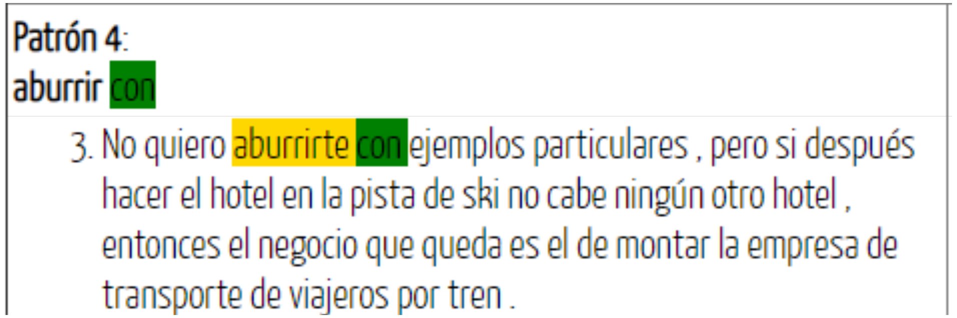
Patrón 1:
aburrir de

1. Si estás aburrida de las dietas hipocalóricas aburridas , de los menús bajos en grasa y en sabor y de contar calorías , a lo mejor ha llegado el momento de que te plantees un cambio en tu estilo de vida .



Patrón 2:
aburrir a

1. No todo es tanÂ aburrido a la vuelta de vacaciones .
2. Manzano mira un poco a la cantera que vas a terminar de aburrir a los chavales .
3. El frontal de la vitro o de tu cocina de gas puede pasar de ser un sitio aburrido a uno más divertido si colocas un vinilo con coloridas flores , un estampado o algún otro motivo más propio de la cocina .

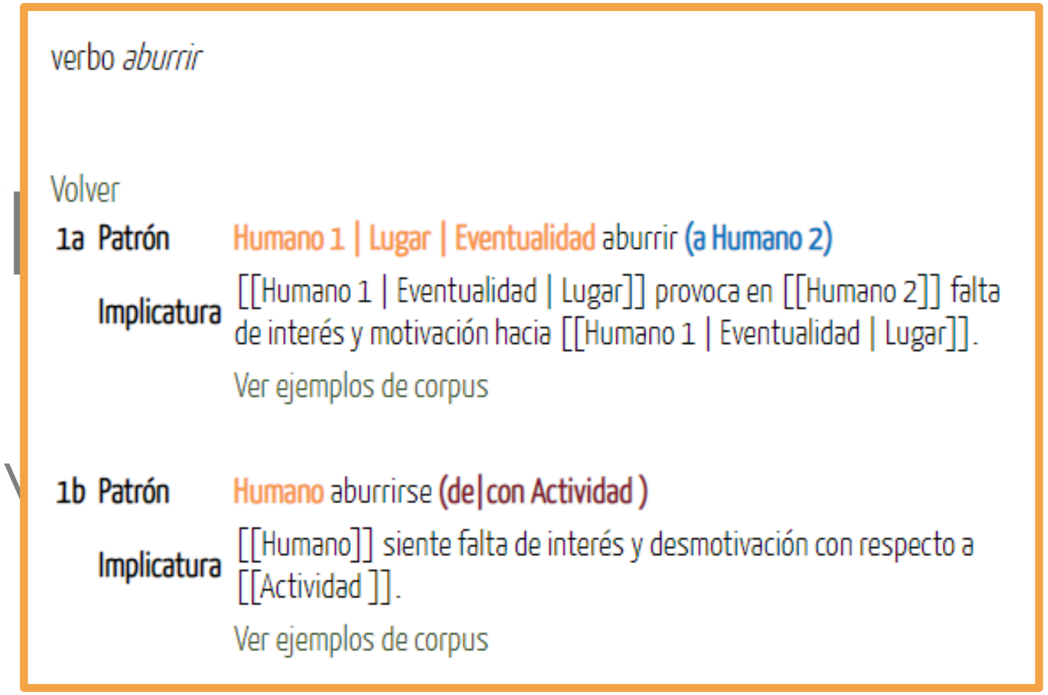


Patrón 4:
aburrir con

3. No quiero aburrirte con ejemplos particulares , pero si después hacer el hotel en la pista de ski no cabe ningún otro hotel , entonces el negocio que queda es el de montar la empresa de transporte de viajeros por tren .

APPLYING CPA TO SPANISH V

An example of results with the automatic part of V



verbo *aburrir*

Volver

1a Patrón Humano 1 | Lugar | Eventualidad aburrir (a Humano 2)

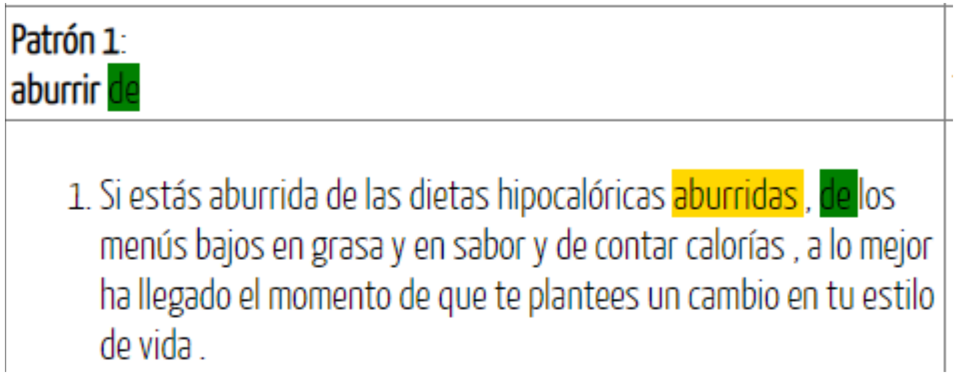
Implicatura [[Humano 1 | Eventualidad | Lugar]] provoca en [[Humano 2]] falta de interés y motivación hacia [[Humano 1 | Eventualidad | Lugar]].

Ver ejemplos de corpus

1b Patrón Humano aburrirse (de|con Actividad)

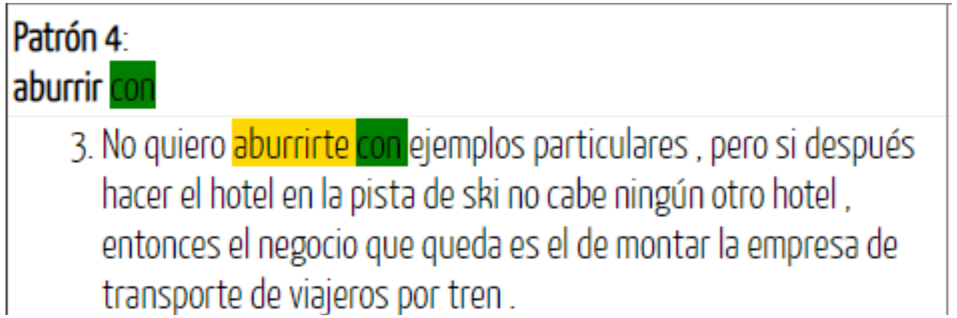
Implicatura [[Humano]] siente falta de interés y desmotivación con respecto a [[Actividad]].

Ver ejemplos de corpus



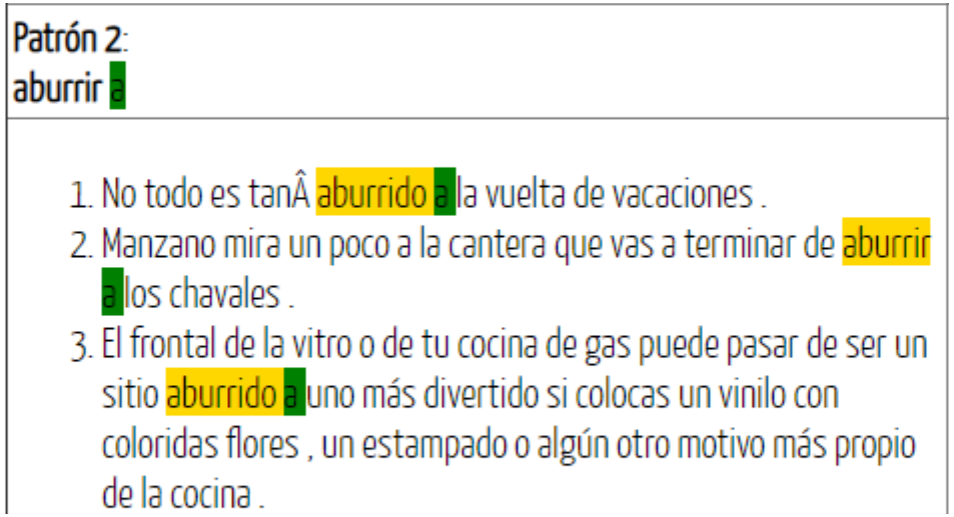
Patrón 1:
aburrir de

1. Si estás aburrida de las dietas hipocalóricas aburridas , de los menús bajos en grasa y en sabor y de contar calorías , a lo mejor ha llegado el momento de que te plantees un cambio en tu estilo de vida .



Patrón 4:
aburrir con

3. No quiero aburrirte con ejemplos particulares , pero si después hacer el hotel en la pista de ski no cabe ningún otro hotel , entonces el negocio que queda es el de montar la empresa de transporte de viajeros por tren .

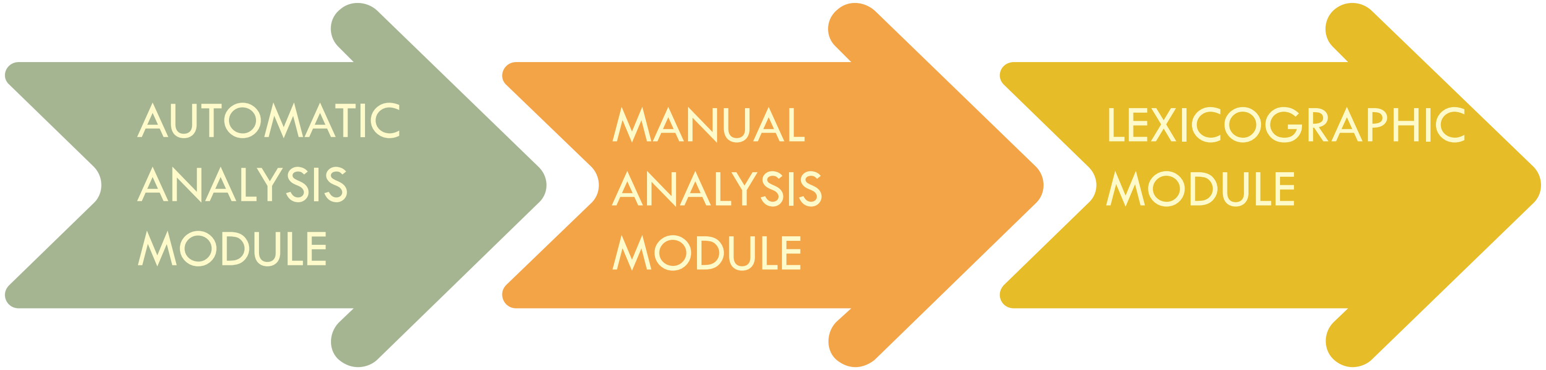


Patrón 2:
aburrir a

1. No todo es tanÂ aburrido a la vuelta de vacaciones .
2. Manzano mira un poco a la cantera que vas a terminar de aburrir a los chavales .
3. El frontal de la vitro o de tu cocina de gas puede pasar de ser un sitio aburrido a uno más divertido si colocas un vinilo con coloridas flores , un estampado o algún otro motivo más propio de la cocina .

TOWARDS A PEDAGOGICAL DICTIONARY BASED ON VERBARIO'S DATA

Process from the non-supervised data to a pedagogical dictionary (in progress)



We conduct a first automatic corpus analysis and offer hypotheses of patterns to the lexicographers, linked to the corpus data

We conduct manual analysis of the automatic data.

Patterns are transferred to the lexicographic database and put in the form of a pedagogical dictionary.

TOWARDS A PEDAGOGICAL DICTIONARY BASED ON VERBARIO'S DATA

Strategies for pattern → definition conversion:

01 Adapting the technical CPA terms into standard vocabulary

Pattern in Verbario:
[[Humano]] cortar ([[Objeto Físico | Parte de Objeto Físico]])

Pattern in a pedagogical Verbario-based dictionary:
Una persona corta un objeto o parte de este cuando lo divide en dos o más partes usando un cuchillo u otro instrumento afilado.

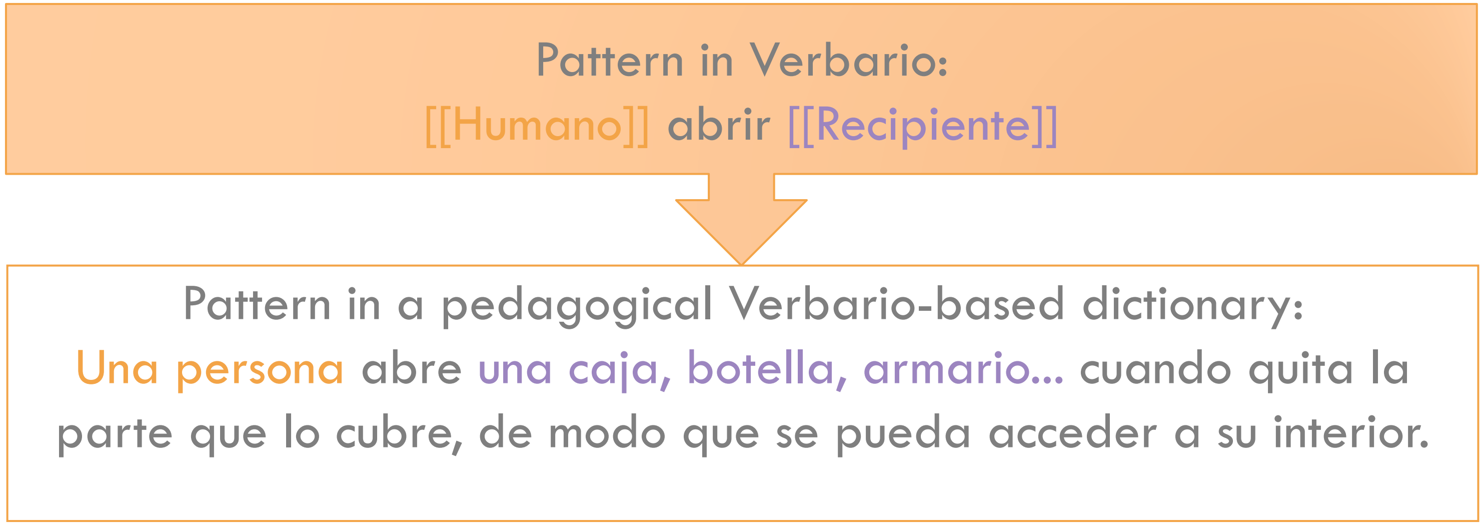
'[[Human]] cuts [[Physical Object | Physical Object part'

'A person cuts an object or a part of it when divides it into two or more parts using a knife or another sharp instrument'.

TOWARDS A PEDAGOGICAL DICTIONARY BASED ON VERBARIO'S DATA

Strategies for pattern → definition conversion:

02 Substitute a semantic type for a lexical set when it is clearer this way



Pattern in Verbario:
[[Humano]] abrir [[Recipiente]]

Pattern in a pedagogical Verbario-based dictionary:
Una persona abre una caja, botella, armario... cuando quita la parte que lo cubre, de modo que se pueda acceder a su interior.

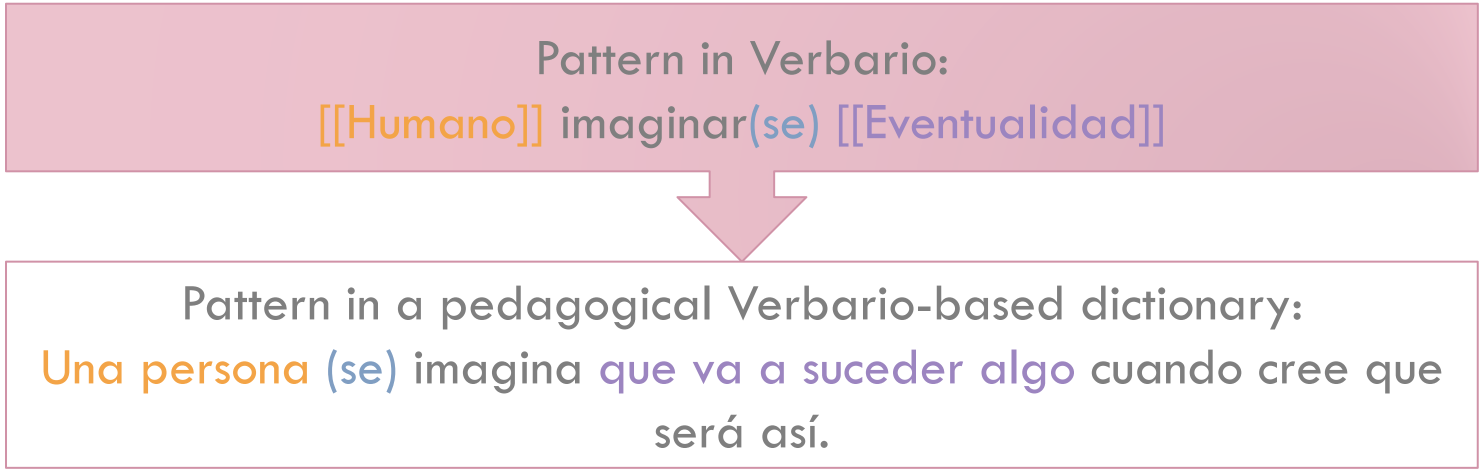
'[[Human]] opens [[Container]]'.

'A person opens a box, bottle, closet… when removes the part covering it, so there is access to what it is inside'.

TOWARDS A PEDAGOGICAL DICTIONARY BASED ON VERBARIO'S DATA

Strategies for pattern → definition conversion:

03 Making the syntactic structure explicit (often specifying the meaning)



Pattern in Verbario:
[[Humano]] imaginar(se) [[Eventualidad]]

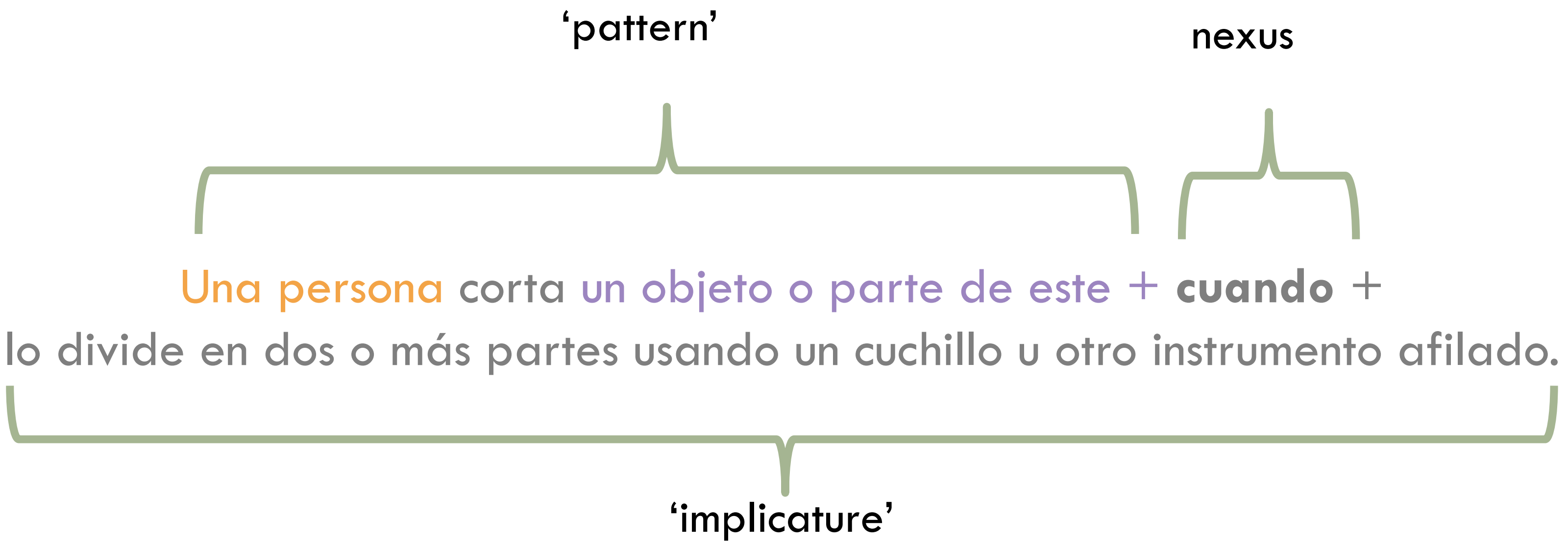
Pattern in a pedagogical Verbario-based dictionary:
Una persona (se) imagina que va a suceder algo cuando cree que será así.

'[[Human]] imagine [[Eventuality]]'.

'A person imagines that something is going to happen when he believes it is going to happen'.

TOWARDS A PEDAGOGICAL DICTIONARY BASED ON VERBARIO'S DATA

Strategies for pattern → definition conversion:



'pattern'

nexus

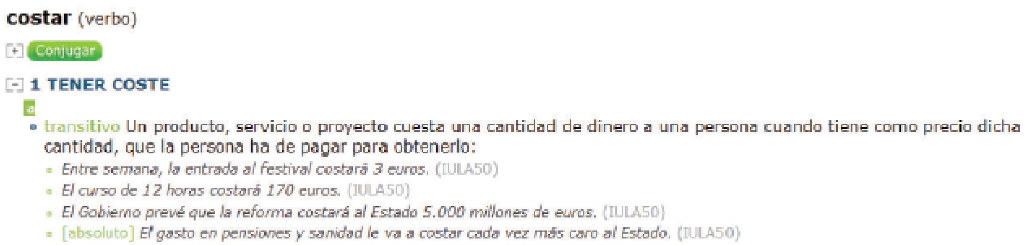
Una persona corta un objeto o parte de este + **cuando** +
lo divide en dos o más partes usando un cuchillo u otro instrumento afilado.

'implicature'

'A person cuts an object or a part of it **when** divides it into two or more parts using a knife or another sharp instrument'.

TOWARDS A PEDAGOGICAL DICTIONARY BASED ON VERBARIO'S DATA

There are very few precedents of using patterns à la CPA in Spanish lexicography. However…



costar (verbo)

[+] Conjugar

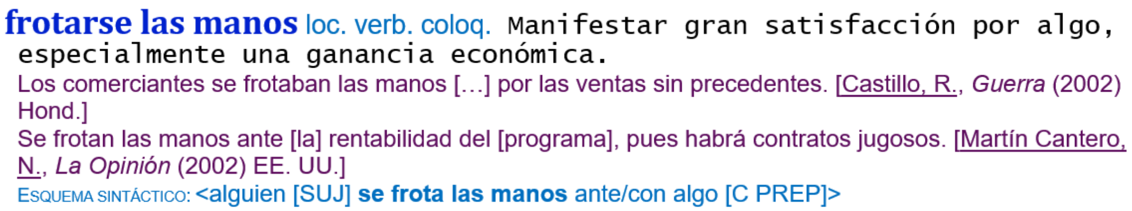
[-] **1 TENER COSTE**

a

- transitivo Un producto, servicio o proyecto cuesta una cantidad de dinero a una persona cuando tiene como precio dicha cantidad, que la persona ha de pagar para obtenerlo:
  - *Entre semana, la entrada al festival costará 3 euros.* (IULA50)
  - *El curso de 12 horas costará 170 euros.* (IULA50)
  - *El Gobierno prevé que la reforma costará al Estado 5.000 millones de euros.* (IULA50)
  - [absoluto] *El gasto en pensiones y sanidad le va a costar cada vez más caro al Estado.* (IULA50)

Diccionario de aprendizaje del español como lengua extranjera, DAELE, dir. P. Battaner

(See Arias Badia, Bernal &Alonso, 2014)



frotarse las manos loc. verb. coloq. Manifestar gran satisfacción por algo, especialmente una ganancia económica.

Los comerciantes se frotaban las manos […] por las ventas sin precedentes. [Castillo, R., *Guerra* (2002) Hond.]

Se frotan las manos ante [la] rentabilidad del [programa], pues habrá contratos jugosos. [Martín Cantero, N., *La Opinión* (2002) EE. UU.]

ESQUEMA SINTÁCTICO: <alguien [SUJ] **se frota las manos** ante/con algo [C PREP]>

Diccionario fraseológico panhispánico, in progress, Associarion of Spanish Academies

FINAL REMARKS

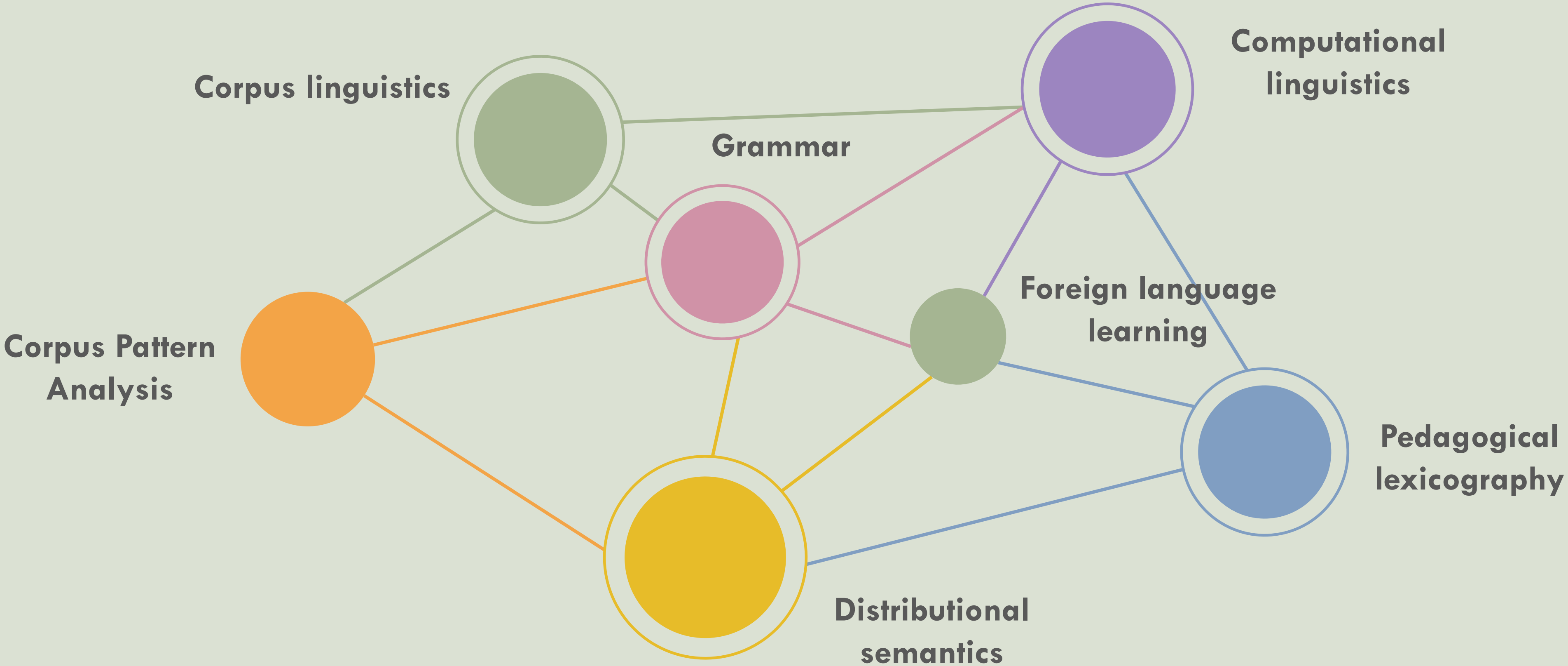
Verbario is a lexical database of Spanish verbs containing many corpus evidence of the meaning and function of verbs in context.

This information can be used for **lexicographic purposes**.

Kind is a parallel project with its own future tasks and applications (pattern building, semantic labelling, neology extraction, etc.).

Automatic pattern building is still in progress, as we have not yet achieved enough precision. Among the automatic patterns we find many cases of **'pseudo-patterns'** = a fragment of the pattern rather than a complete analysis, e.g. verb + semantic type / preposition.
We consider this data useful anyway and the challenge now is to integrate these parts. Other lines of future work:

- Improving dependency parsing
- Improving Kind
- Test more or different corpora



THANK YOU!

@TeclingGroup

www.tecling.com

Irene Renau & Rogelio Nazar
Pontificia Universidad Católica de Valparaíso

PhrasaLex | 19-20 September 2019 | Univ. degli Studi di Modena