CORCODE: A Corpus on Definitional Contexts as a Lexicography Resource

Rodrigo Alarcón & Gerardo Sierra

Language Engineering Group, National Autonomous University of Mexico, Mexico City
{ralarconm,gsierram}@ii.unam.mx

AACL 2009 - American Association for Corpus Linguistics

Edmonton, Alberta

October 8 - 11, 2009

Outline

- Problem description
- Objectives
- Definitional Contexts Extraction
- Corpus Annotation
- Current & future work

Problem description (1)

- An important goal for lexicography and terminology is to identify and classify possible terms and definitions in large-collections of specialised texts.

- There are several methods and automatic systems to search and extract relevant information in texts:

  · Term extraction

  · Term and definition extraction

  · Ontology construction

Problem description (2)

- There is a relevant success in the automatic extraction of terms, because their syntactic organization: in more cases, it is clear the presence of regular linguistic patterns.

- Nevertheless, the automatic extraction of definitions is an awkward task, because the complex and richness of how definitions can occur in texts.

Problem description (3)

Abbreviation is also an important component of text mining algorithms. In one study of computational approaches to generate a lexicon for biomedical natural language processing applications, the researchers noted that not handling abbreviations in the text is a major source of error…

Speaking precisely, abbreviation is a broad term that describes a shortened form of a word or phrase. The term acronym is also commonly used and generally means a shortened form created from the initial letters of the word in the phrase. Some people also require acronyms to be pronounceable. In this chapter, we will refer to the most general problem as abbreviation identification, and we consider an acronym to be a type of abbreviation.

Problem description (3)

Abbreviation is also an important component of text mining algorithms. In one study of computational approaches to generate a lexicon for biomedical natural language processing applications, the researchers noted that not handling abbreviations in the text is a major source of error…

Speaking precisely, abbreviation is a broad term that describes a shortened form of a word or phrase. The term acronym is also commonly used and generally means a shortened form created from the initial letters of the word in the phrase. Some people also require acronyms to be pronounceable. In this chapter, we will refer to the most general problem as abbreviation identification, and we consider an acronym to be a type of abbreviation.

Problem description (3)

Abbreviation is also an important component of text mining

consider an acronym to be a type of abbreviation.

Definitional Contexts

Speaking precisely, abbreviation is a broad term that describes a shortened form of a word or phrase.

Term + Verbal Pattern + Definition

Problem description (3)

abbreviation

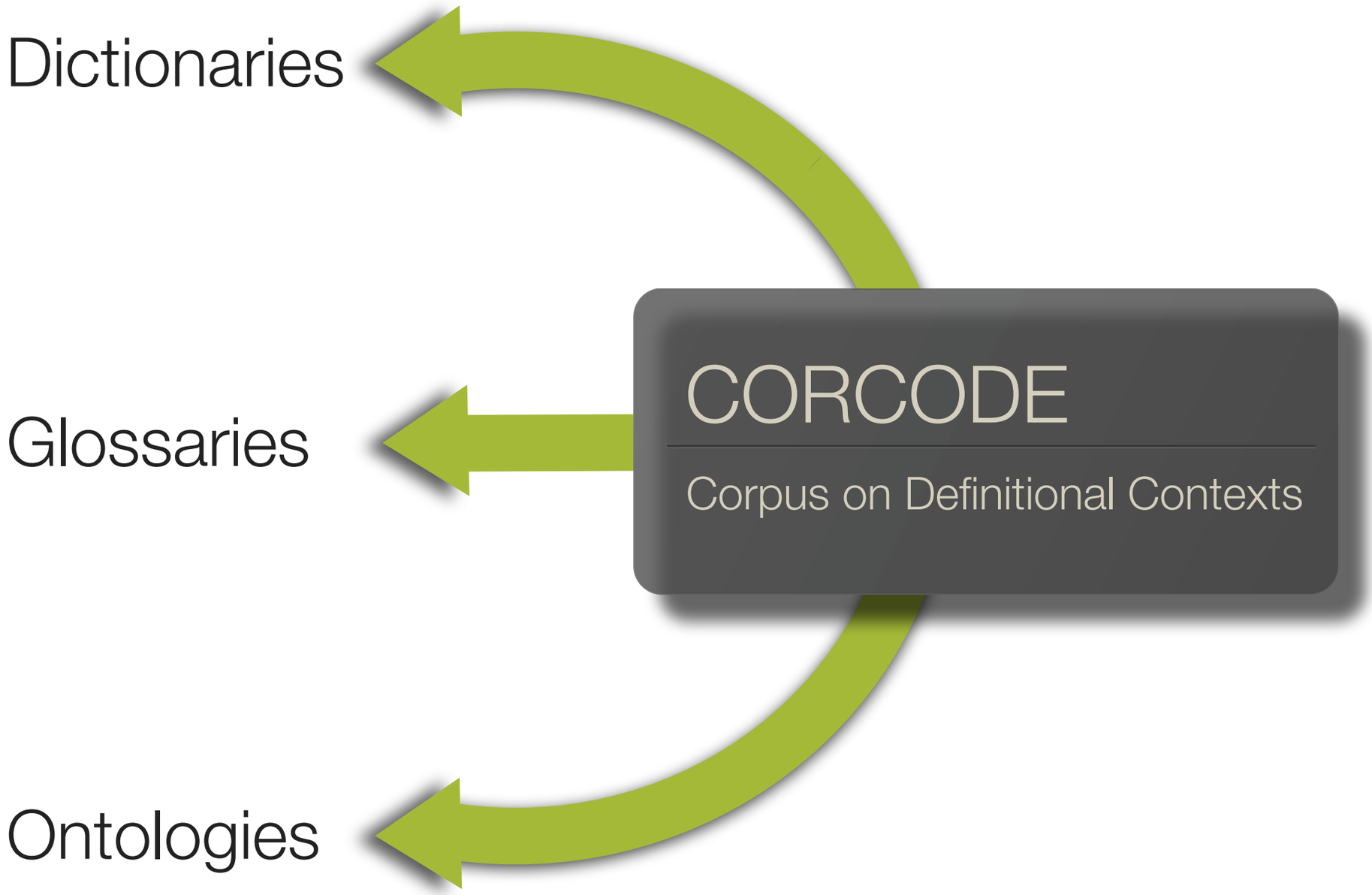
~ a shortened form of a word or phrase

→ acronym

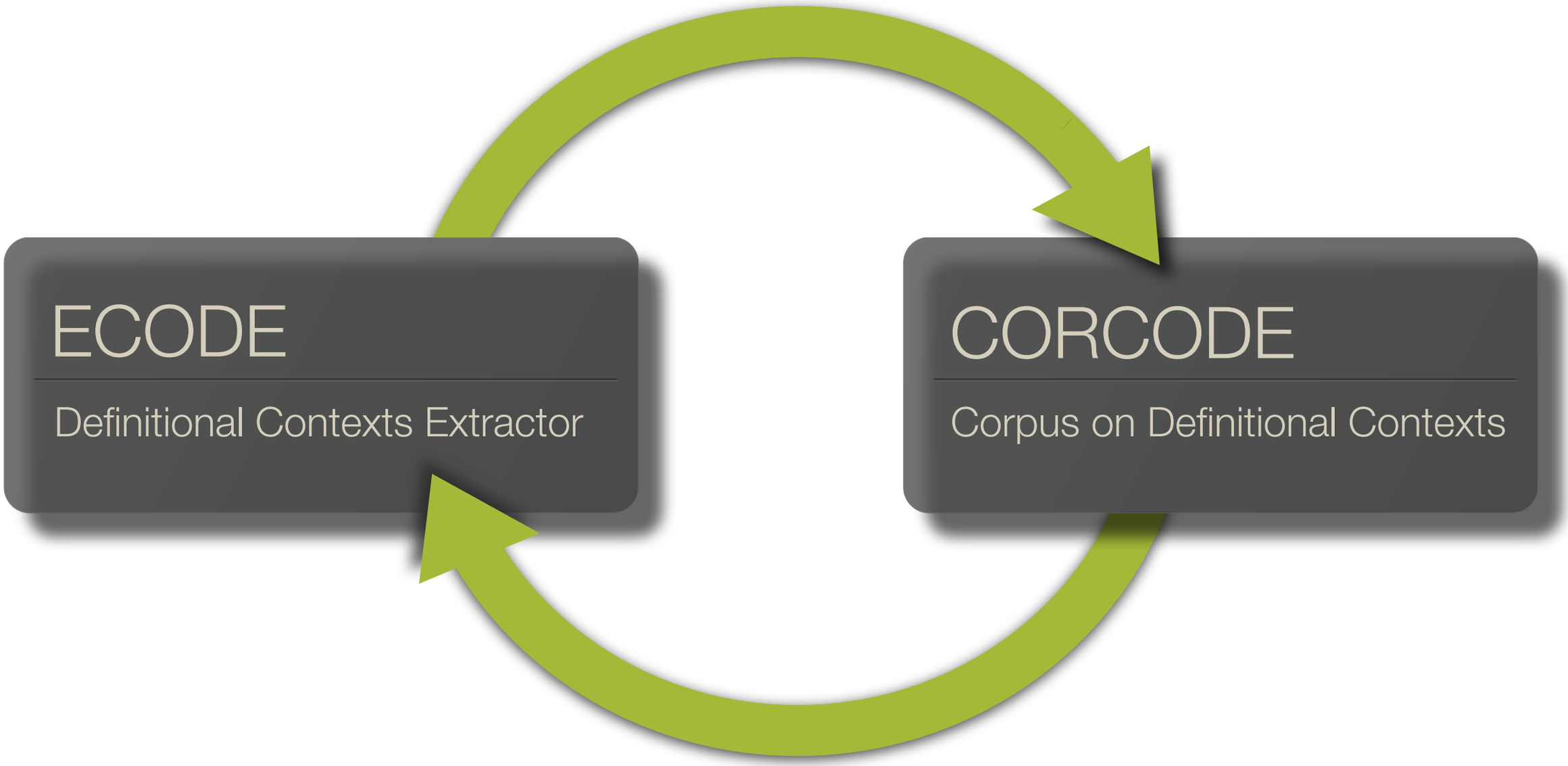
~ a shortened form created from the initial letters of the word in a phrase

~ a type of abbreviation

Objectives


Dictionaries
Glossaries
Ontologies
CORCODE
Corpus on Definitional Contexts

Objectives


ECODE
Definitional Contexts Extractor
CORCODE
Corpus on Definitional Contexts

Definitional Contexts Extraction

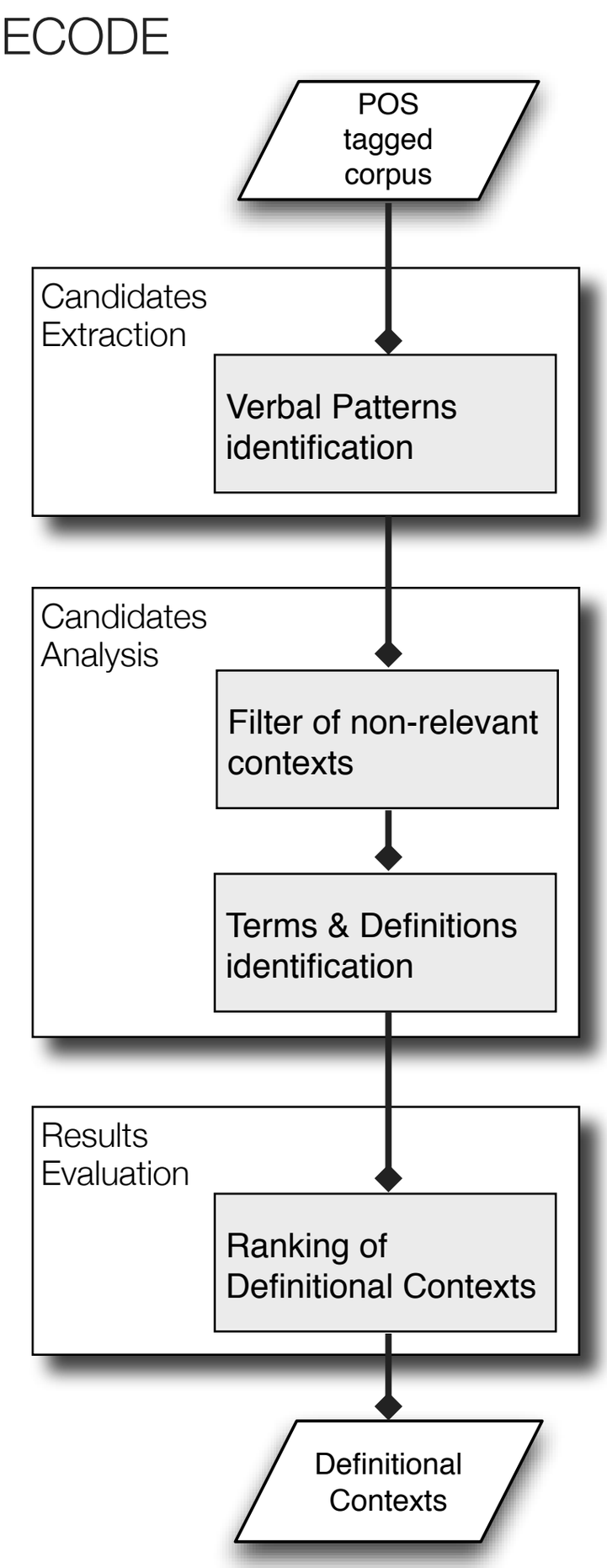
- Pattern-based approach
- Linguistic rules

ECODE

Definitional Contexts Extractor

- Verbal patterns
- Different types of definitions:
  - Analytical
  - Extensional
  - Functional
  - Synonymical

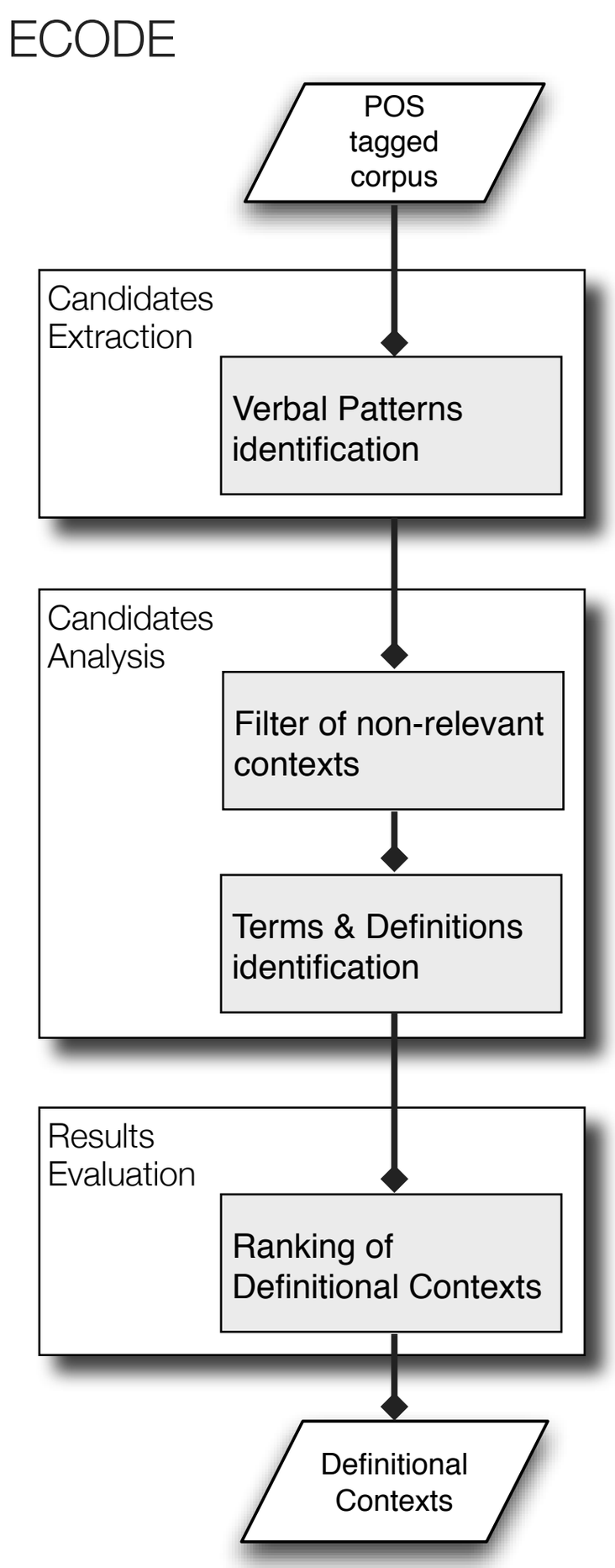
Definitional Contexts Extraction



Definitional Contexts Extraction

- **IULA's Technical Corpus**
  Universidad Pompeu Fabra, Spain
  - Medicine, Genome

- **Corpus on Engineering**
  Universidad Nacional Autónoma de México



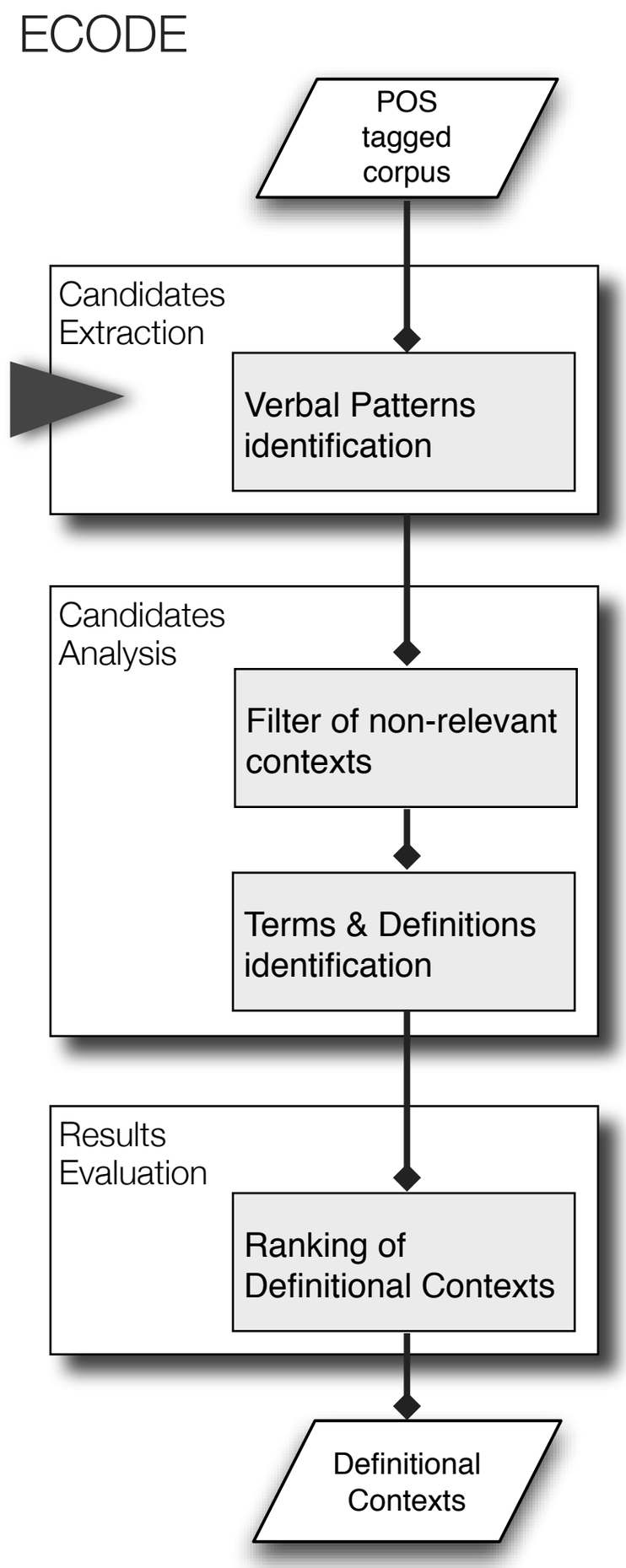
Definitional Contexts Extraction

- **Analytic**
  is a
  defined as

- **Extensional**
  contains
  formed by

- **Functional**
  used for
  serves as

- ***Synonymic***
  also called
  known also



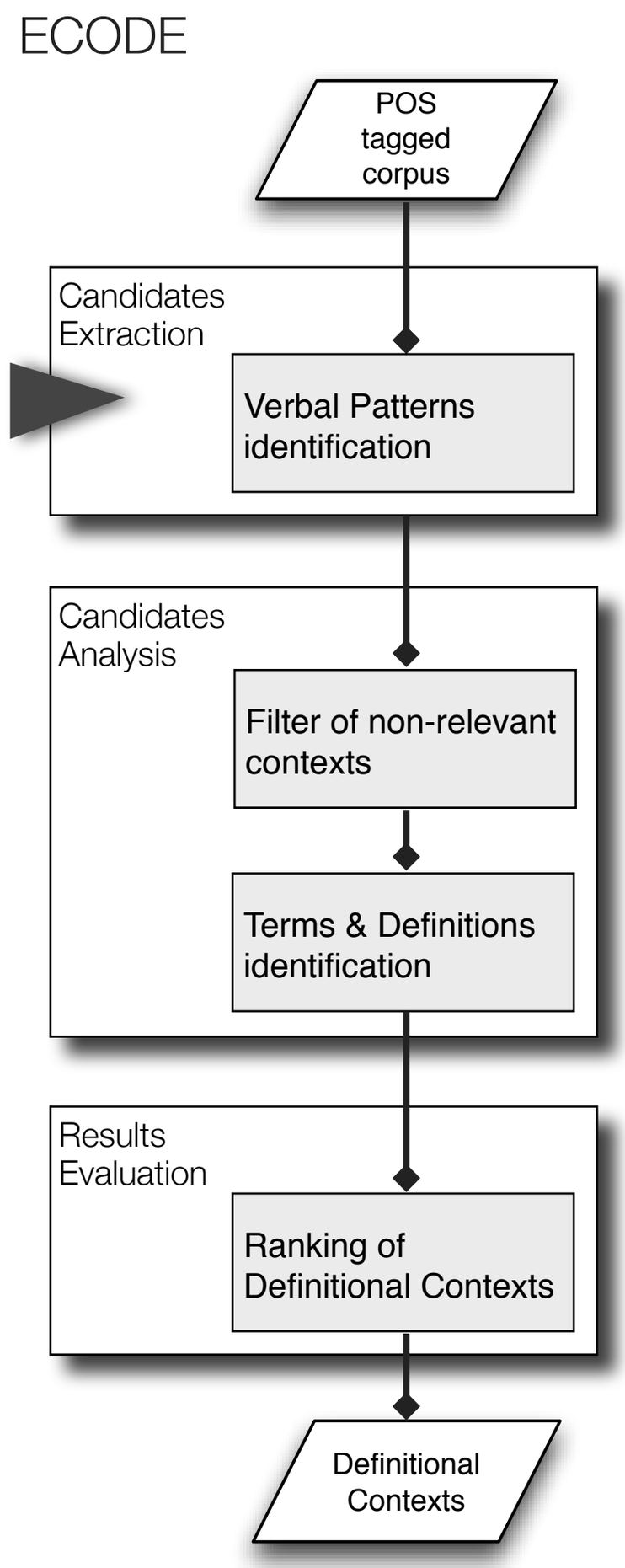
Definitional Contexts Extraction

In this case we **<dvp>**know the initial and final velocities, as well as**</dvp>** the aceleration.

In general terms, a gene **<dvp>**is defined as**</dvp>** a secuence of DNA that codifies a protein RNAt or RNAr.

The reason of that **<dvp>**is a**</dvp>** problem.

The ADN **<dvp>**is a**</dvp>** double helix.



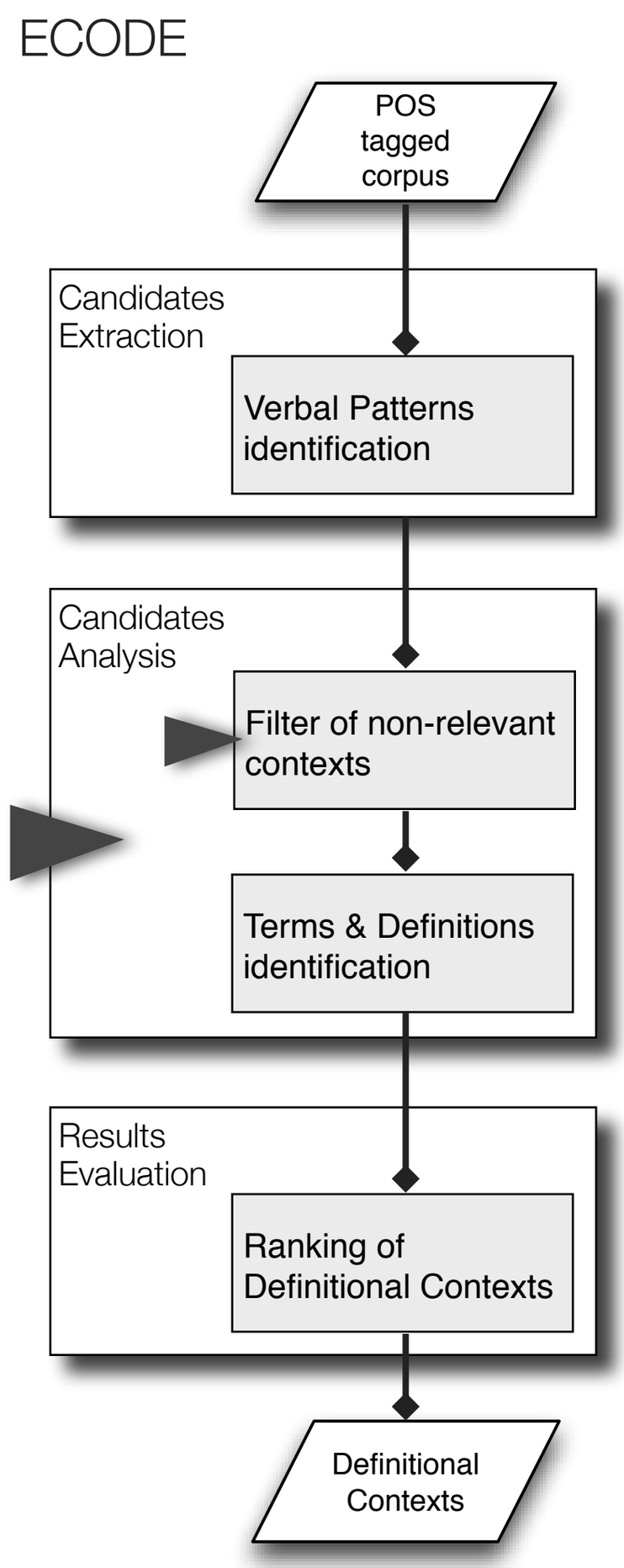
Definitional Contexts Extraction

Filtering rules

“as well as”

“like as”

“no + <dvp>”



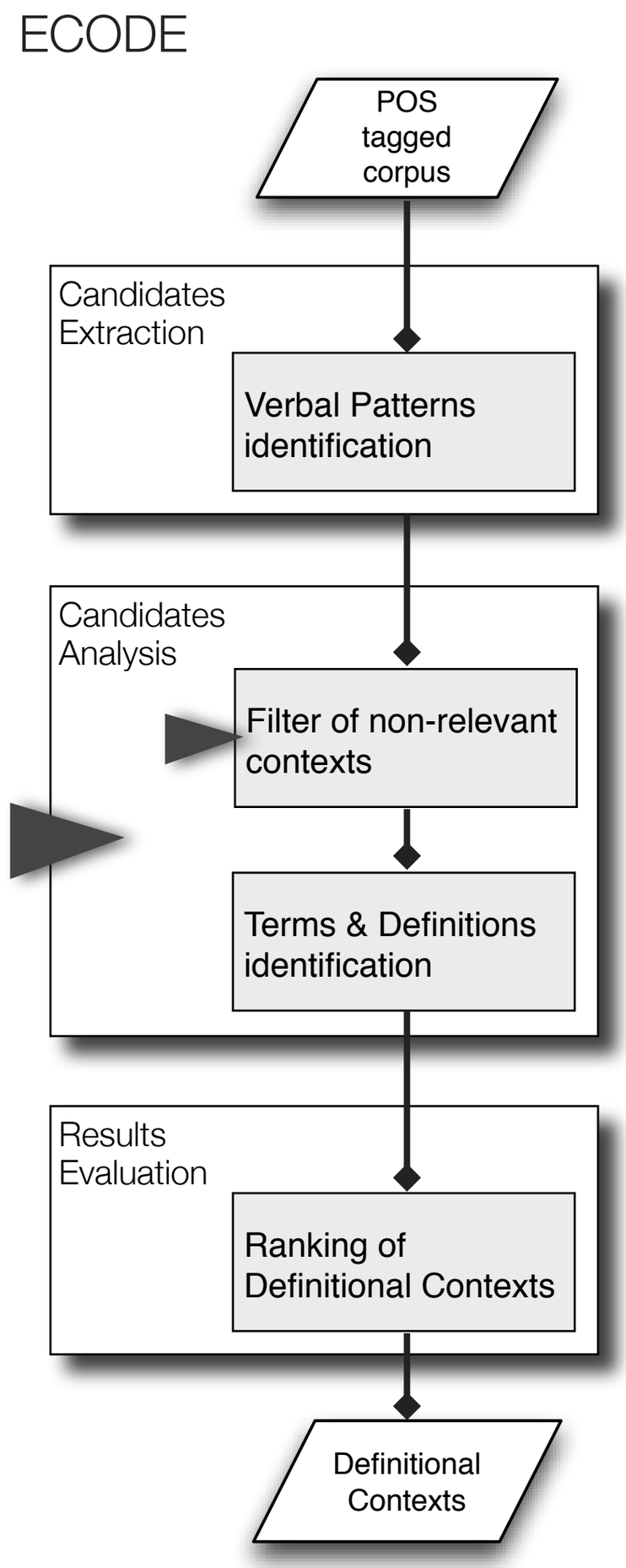
Definitional Contexts Extraction

In this case we **<dvp>**know the initial and final velocities, **as well as</dvp>** the aceleration.

In general terms, a gene **<dvp>**is defined as**</dvp>** a secuence of DNA that codifies a protein RNAt or RNAr.

The reason of that **<dvp>**is a**</dvp>** problem.

The ADN **<dvp>**is a**</dvp>** double helix.

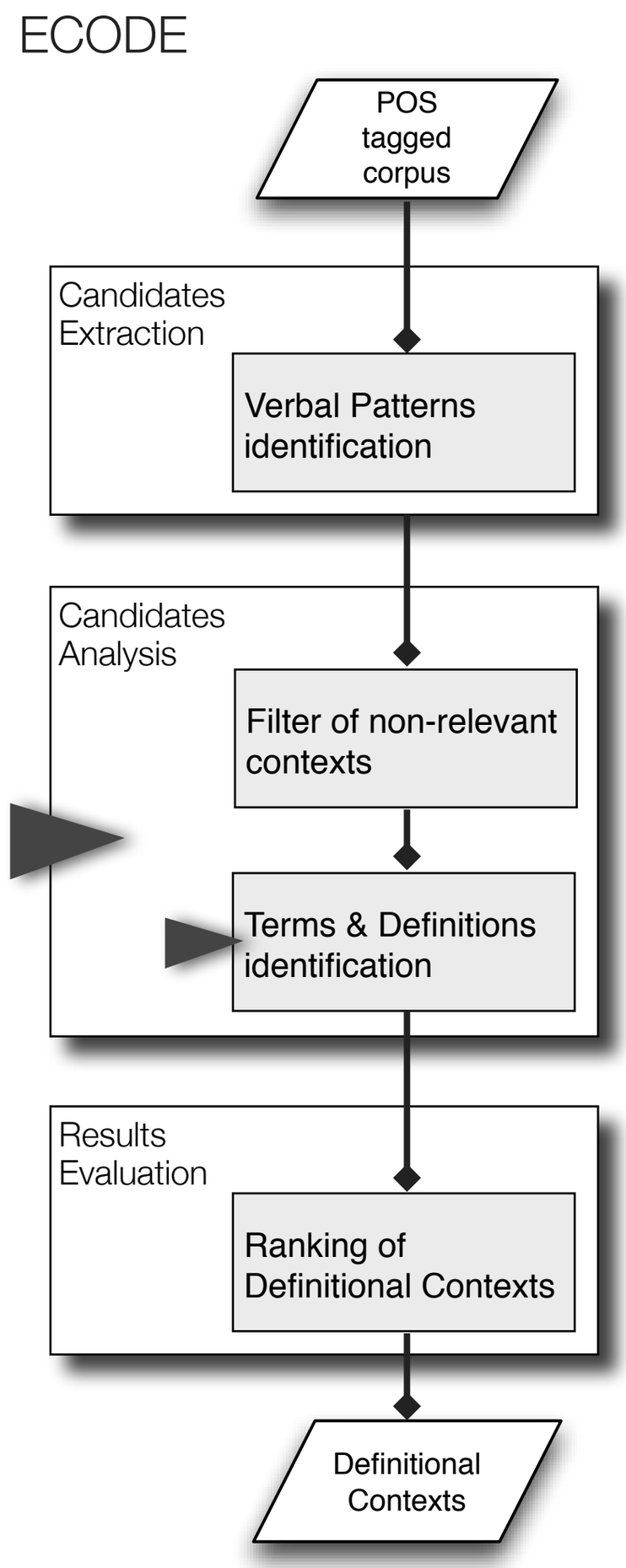


Definitional Contexts Extraction

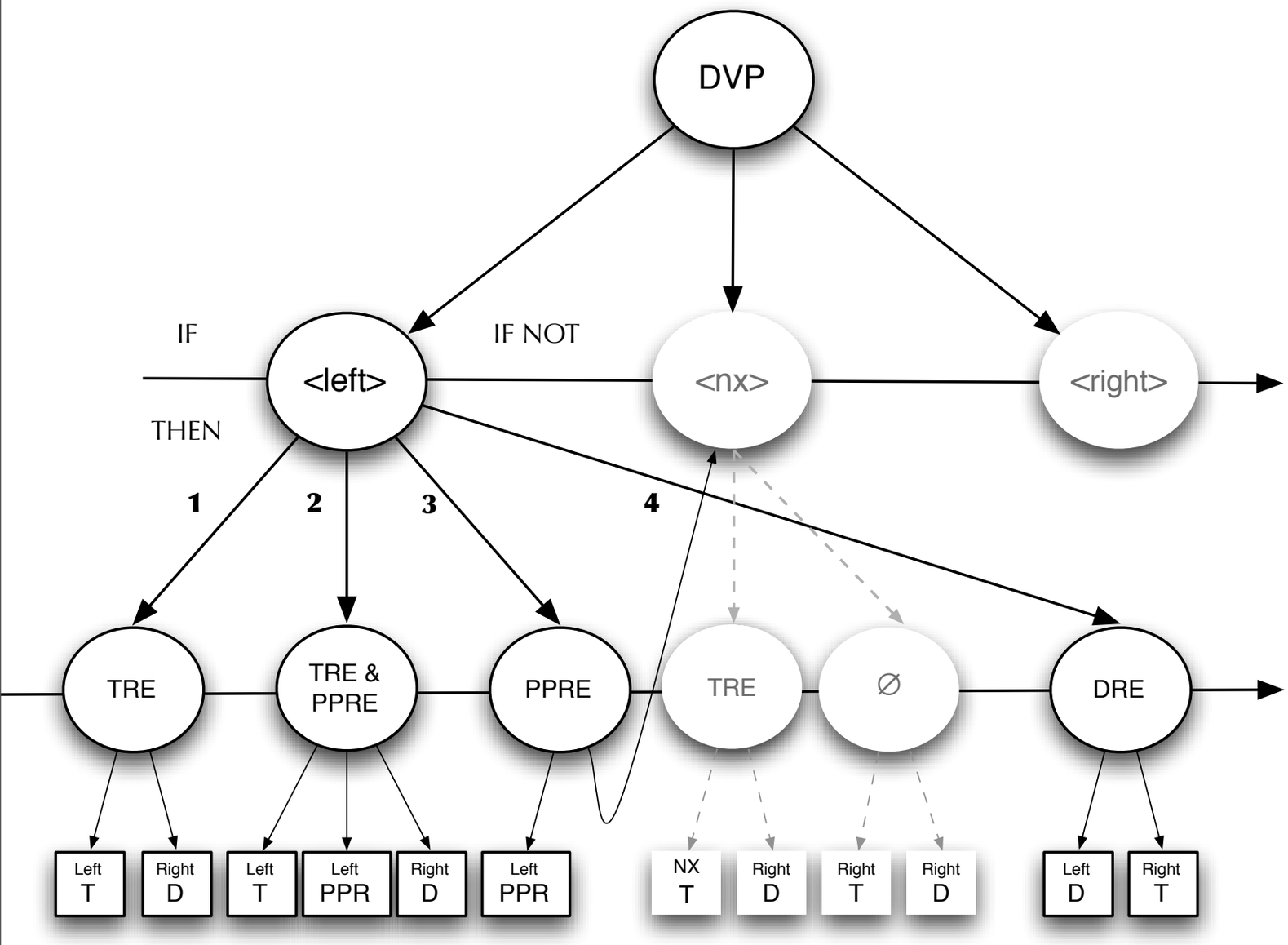
In general terms, a gene **<dvp>**is defined as**</dvp>** a secuence of DNA that codifies a protein RNAt or RNAr.

The reason of that **<dvp>**is a**</dvp>** problem.

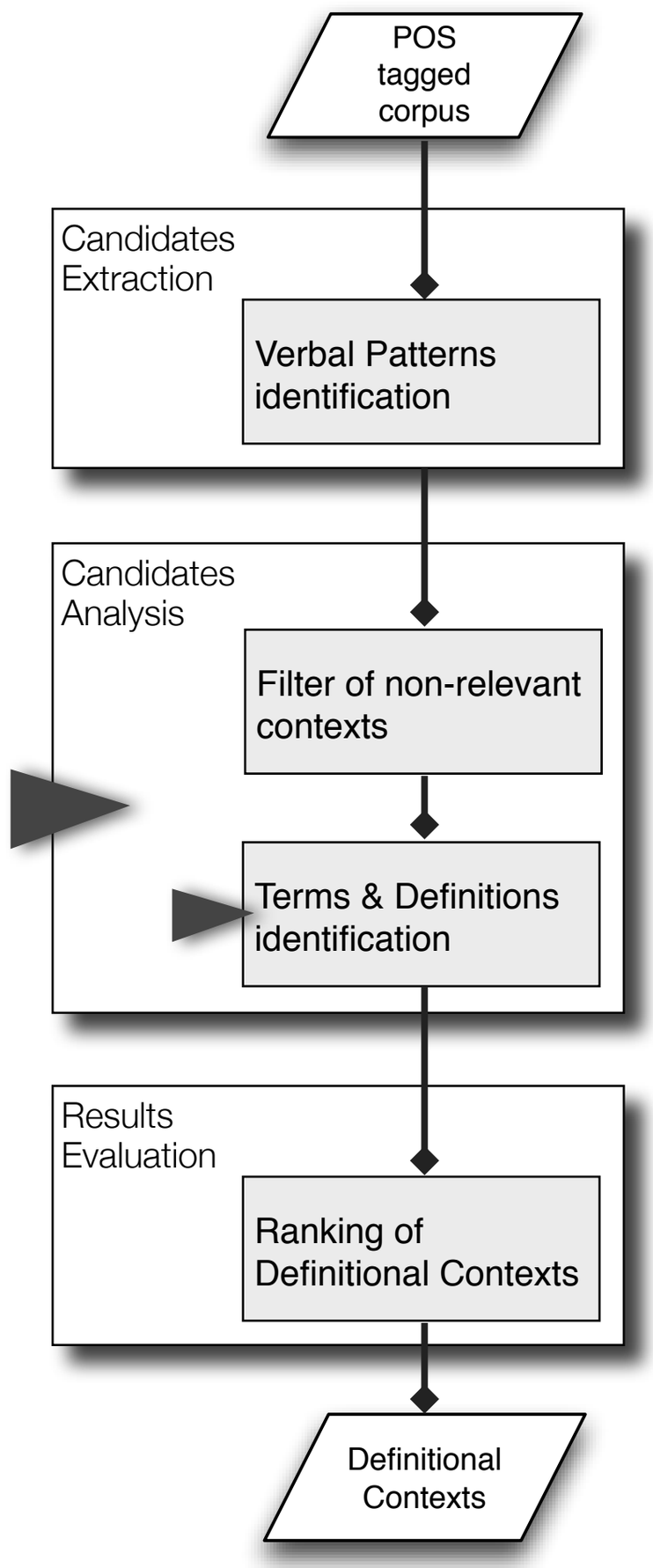
The ADN **<dvp>**is a**</dvp>** double helix.



Definitional Contexts Extraction



ECODE

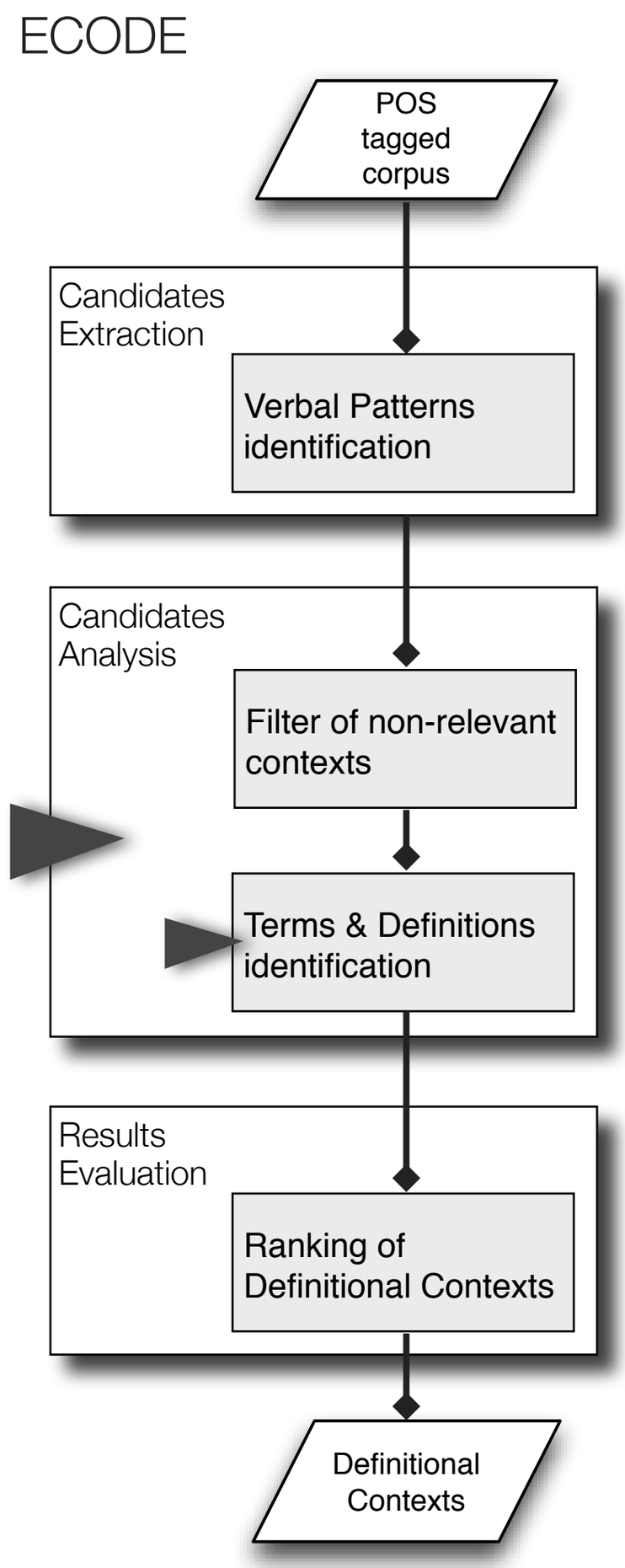


Definitional Contexts Extraction

In general terms, **<term>**a gene**</term>** **<dvp>**is defined as**</dvp>** **<definition>**a sequence of DNA that codifies a protein RNAt or RNAr.**</definition>**

<term>The reason of that**</term>** **<dvp>**is a**</dvp>** **<definition>**problem.**</definition>**

<term>The ADN**</term>** **<dvp>**is a**</dvp>** **<definition>**double helix.**</definition>**

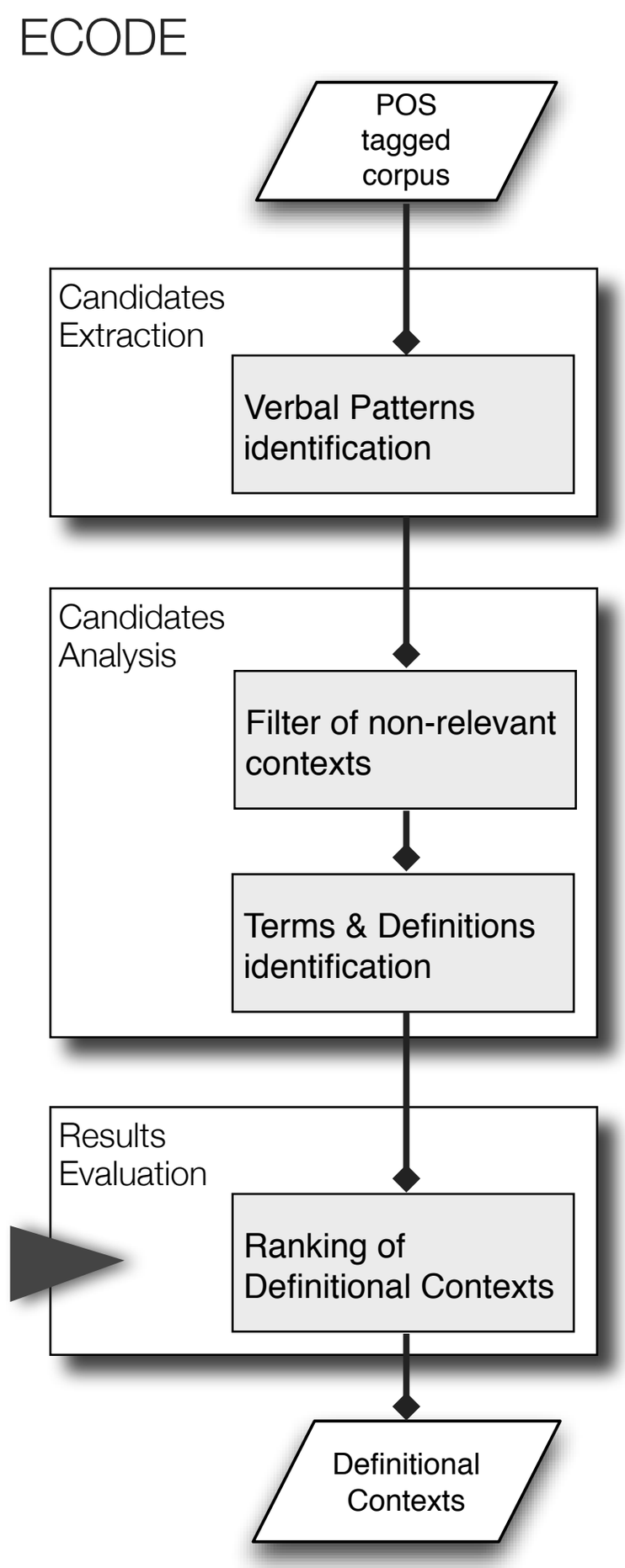


Definitional Contexts Extraction

Ranking rules

v=3 Term has **“pronoun”**

v=3 Term has **“conjugated verb”**

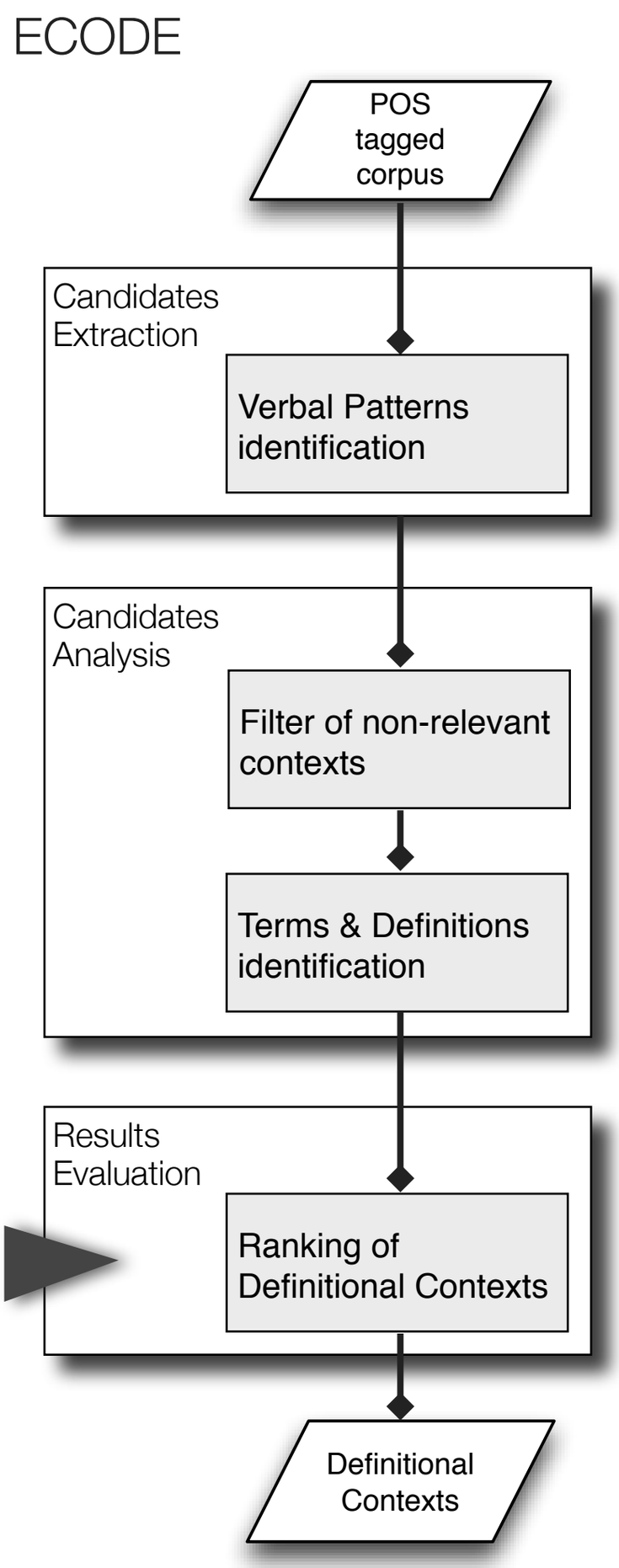


Definitional Contexts Extraction

In general terms, **<term=1>**a gene**</term> <dvp>**is defined as**</dvp> <definition>**a sequence of DNA that codifies a protein RNAt or RNAr.**</definition>**

<term=3>The reason of that**</term> <dvp>**is a**</dvp> <definition>**problem.**</definition>**

<term=1>The ADN**</term> <dvp>**is a**</dvp> <definition>**double helix.**</definition>**

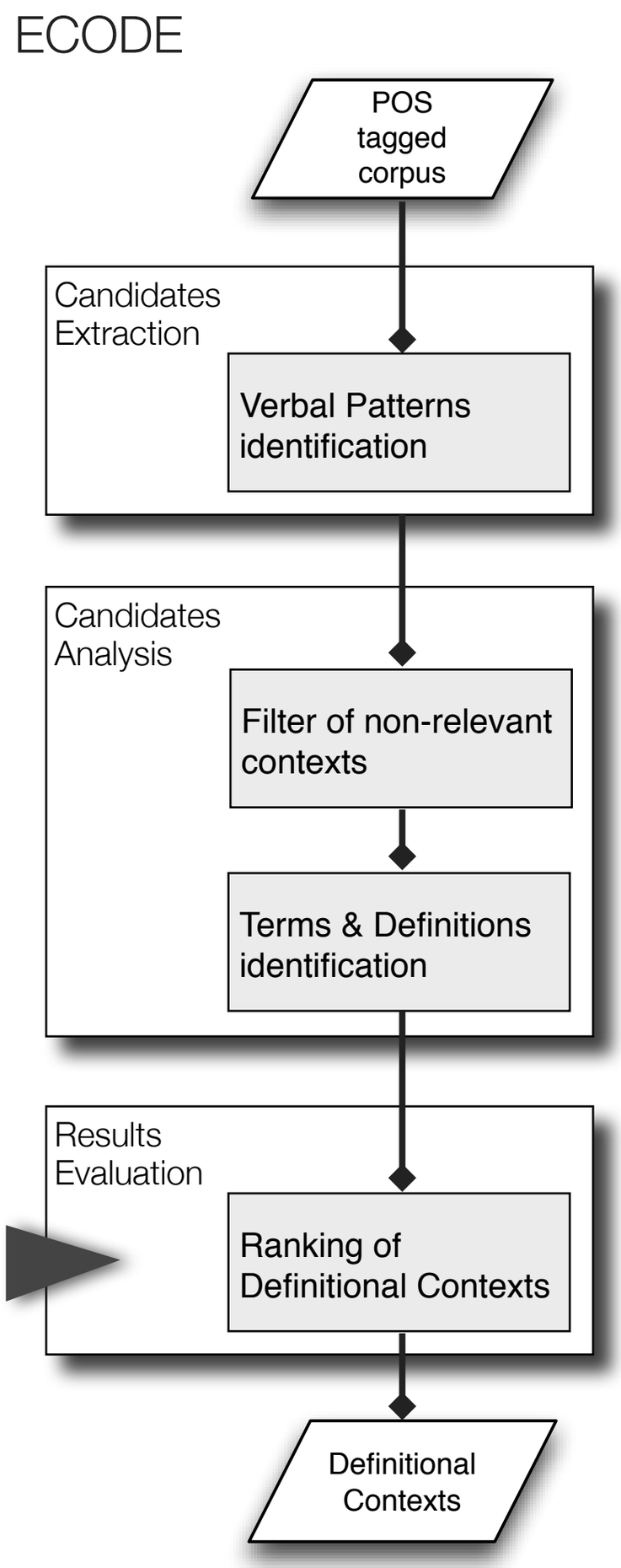


Definitional Contexts Extraction

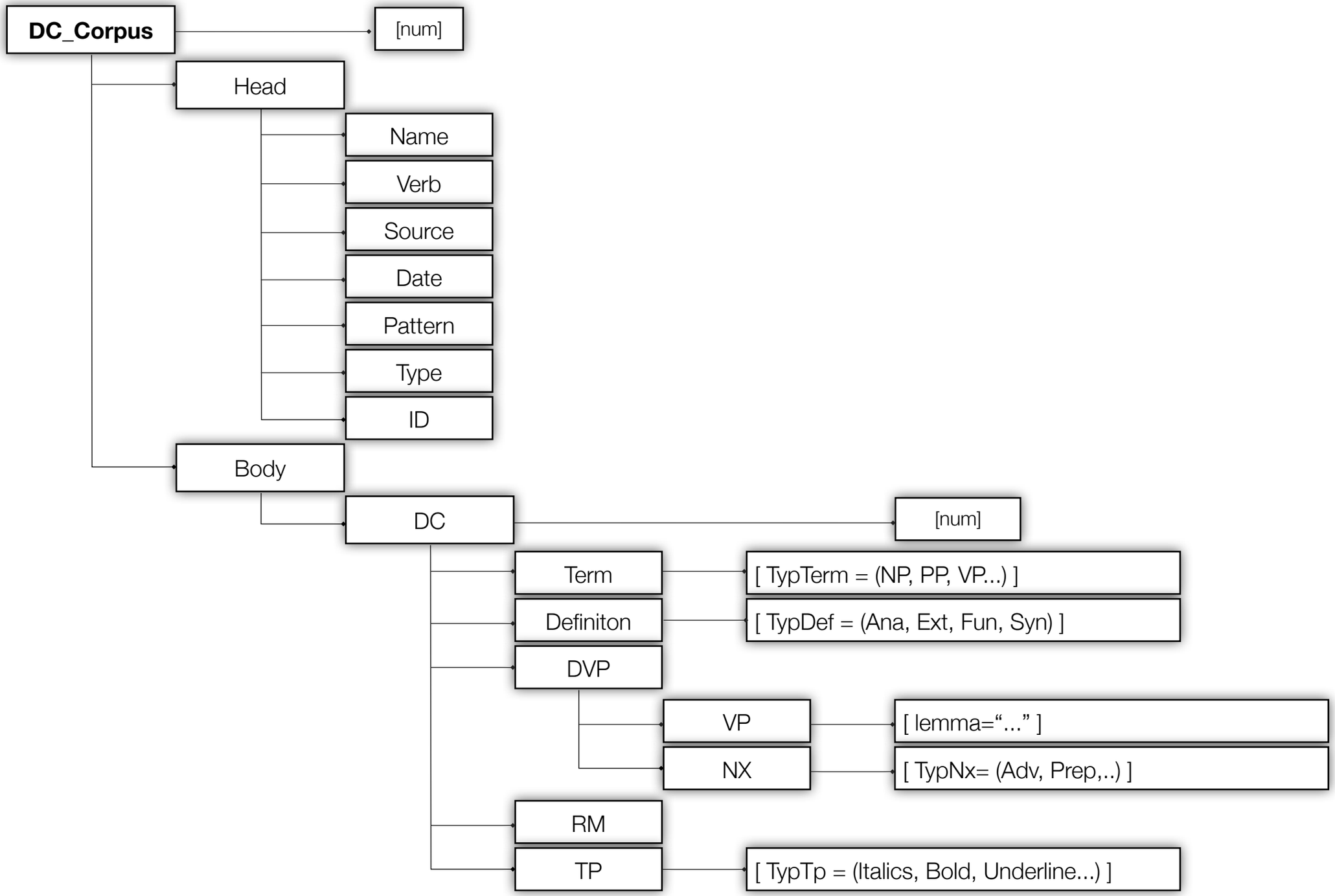
In general terms, **<term=1>**a gene**</term>** **<dvp>**is defined as**</dvp>** **<definition>**a sequence of DNA that codifies a protein RNAt or RNAr.**</definition>**

<term=1>The ADN**</term>** **<dvp>**is a**</dvp>** **<definition>**double helix.**</definition>**

<term=3>The reason of that**</term>** **<dvp>**is a**</dvp>** **<definition>**problem.**</definition>**



Corpus Annotation



Corpus Annotation

Genome is defined as the haploid set of chromosomes in a gamete or microorganism, or in each cell of a multicellular organism.

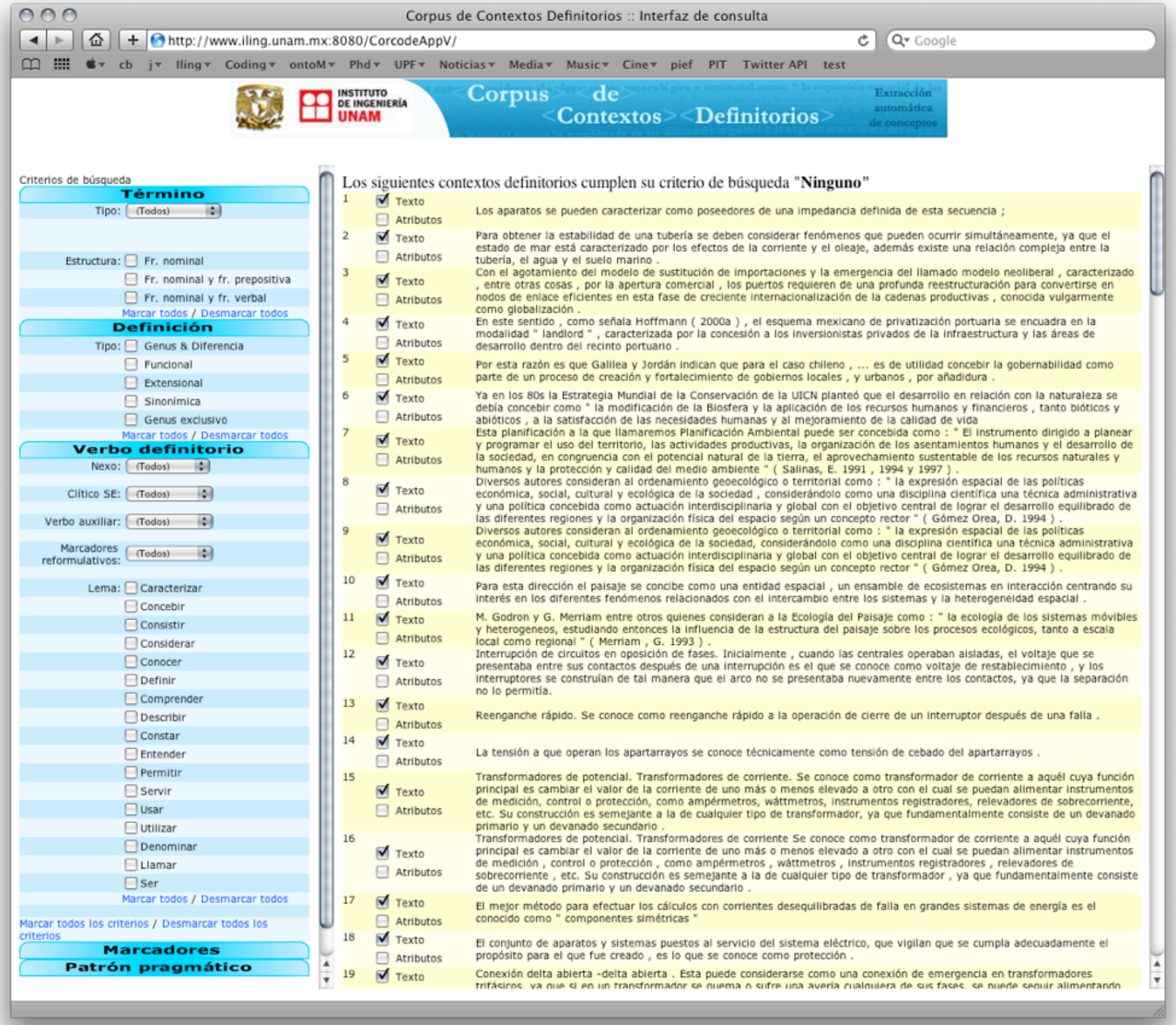
<DC num="908"><TERM typterm="np">Genome</TERM> <DVP><AUXverb>is</AUXverb> <DV lemma="define">defined</DV> <NX typnx="adv">as</NX></DVP> <DEFINITION typdef="Ana">the haploid set of chromosomes in a gamete or microorganism, or in each cell of a multicellular organism.</DEFINITION></DC>

CORCODE

Corpus de Contextos Definitorios — Extracción automática de conceptos

http://www.iling.unam.mx:8080/CorcodeAppV/

2215 Definitional Contexts in Spanish



CORCODE

Corpus de Contextos Definitorios — Extracción automática de conceptos

http://www.iling.unam.mx:8080/CorcodeAppV/

2215 Definitional Contexts in Spanish

Criterios de búsqueda

Término

Tipo: (Todos)

Estructura:
- ☐ Fr. nominal
- ☐ Fr. nominal y fr. prepositiva
- ☐ Fr. nominal y fr. verbal

Marcar todos / Desmarcar todos

Definición

Tipo:
- ☐ Genus & Diferencia
- ☐ Funcional
- ☐ Extensional
- ☐ Sinonímica
- ☐ Genus exclusivo

Marcar todos / Desmarcar todos

Verbo definitorio

Nexo: (Todos)

Clítico SE: (Todos)

Verbo auxiliar: (Todos)

Marcadores reformulativos: (Todos)

Lema:
- ☐ Caracterizar
- ☐ Concebir
- ☐ Consistir
- ☐ Considerar
- ☐ Conocer
- ☐ Definir
- ☐ Comprender
- ☐ Describir
- ☐ Constar
- ☐ Entender
- ☐ Permitir
- ☐ Servir
- ☐ Usar
- ☐ Utilizar
- ☐ Denominar
- ☐ Llamar

Los siguientes contextos definitorios cumplen su criterio de búsqueda "**Tipo definición=sinonímica. Lema=conocer.** "

1

☑ Texto
☑ Atributos

Matemáticamente , la Teoría Lineal , también conocida como Teoría de Airy , puede ser considerada como una primera aproximación de una descripción teórica completa acerca del comportamiento del oleaje

Tipo término: **lingüístico**
Estructura término: **frase nominal**
Tipo definición: **sinonímica**
Tipo nexo: **adverbio**
Tiene marcador SE: **no**
Tiene verbo auxiliar: **no**
Tiene marcadores reformulativos: **no**
Lema del verbo: **conocer**
Tipos de predicaciones pragmáticas: **instruccionales**

Current & Future Work

- Improvement of the corpus
    - New types of definitions
- Multilingual corpus
    - Improve definition extraction on different languages

Thank you

CORCODE:

http://www.iling.unam.mx:8080/CorcodeAppV//

ECODE examples:

http://brangaene.upf.es/ecode/medimit

1st Workshop on Definition Extraction (RANLP 2009):

http://www.iling.unam.mx/wde/

My email:

ralarconm@ii.unam.mx