CS11-737 Multilingual NLP

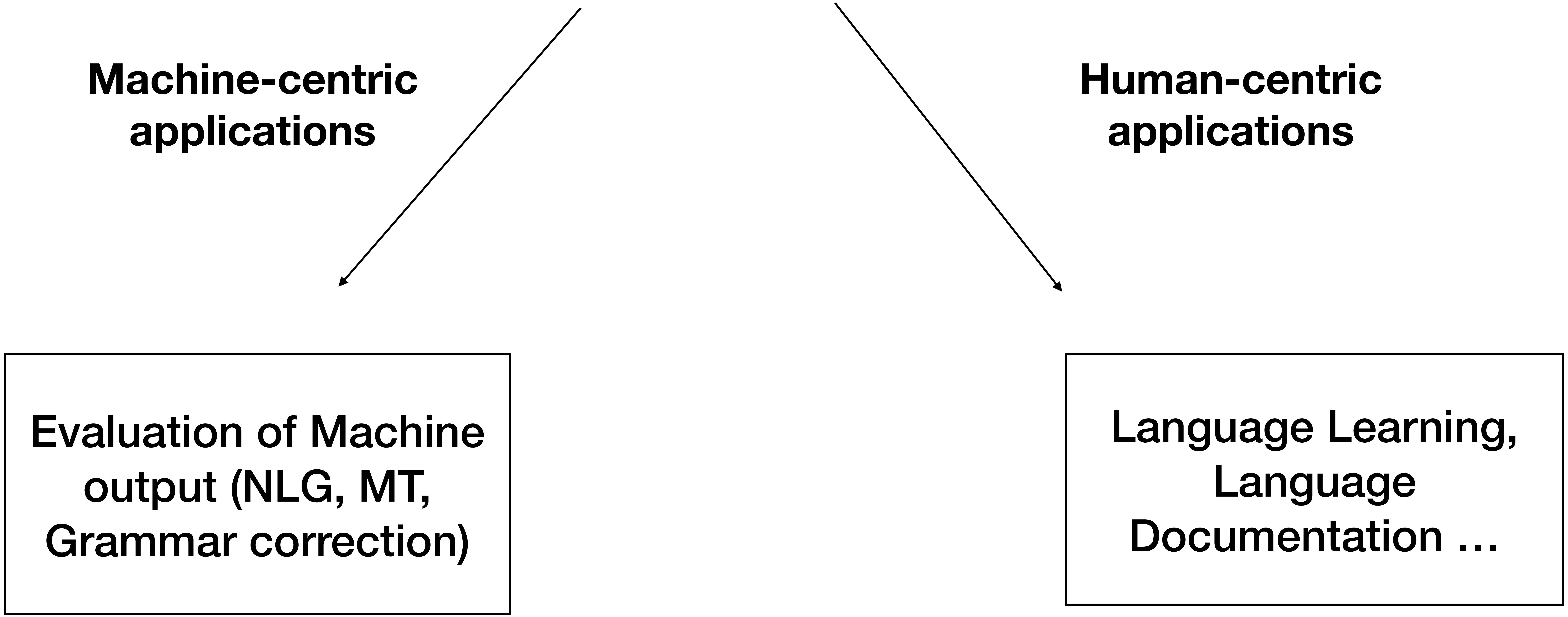
Predicting Linguistic Insights

Aditi Chaudhary

Carnegie Mellon University
Language Technologies Institute

Use of dependencies?

- Understand complex linguistic phenomena e.g.
  morphological agreement, word order, case marking, suffix usage …



Machine-centric applications

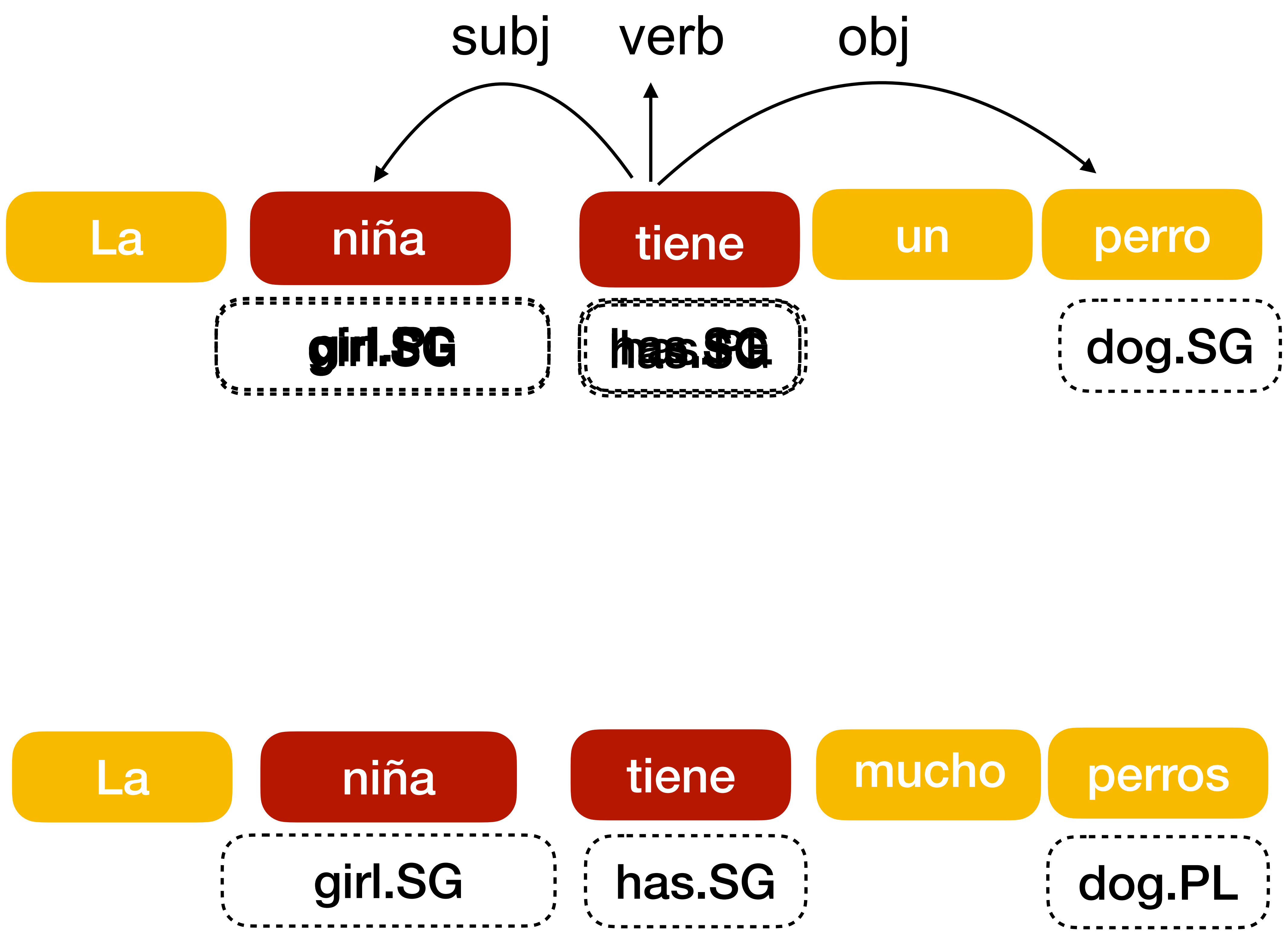
Human-centric applications

Evaluation of Machine output (NLG, MT, Grammar correction)

Language Learning, Language Documentation …

Morphological Agreement

- *Agreement* is the process when one word/morpheme changes form based on other word/morpheme's grammar categories (e.g. number)



Subject-verb required-agreement

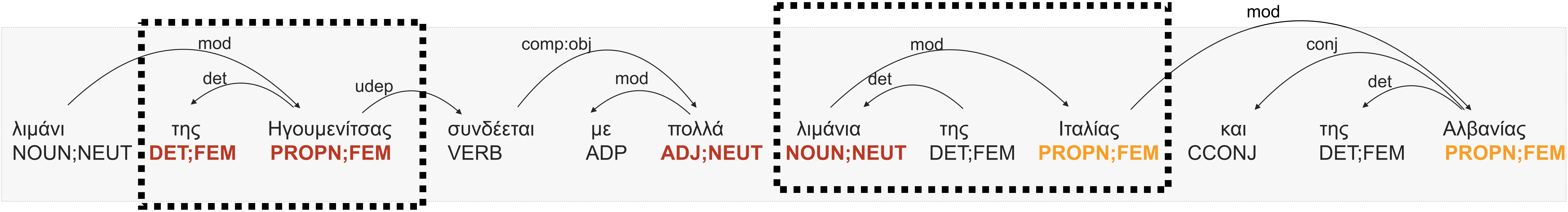
Object-verb chance-agreement

Number agreement in Spanish

Problem Formulation

- Devise a task of predicting **required-agreement** vs **chance-agreement**

λιμάνι	της	Ηγουμενίτσας	συνδέεται	με	πολλά	λιμάνια	της	Ιταλίας	και	της	Αλβανίας
port.SG	DET	Igoumenítsa.GEN	connect.SG	with	many	port.PL	DET	Italy.GEN	and	DET	Albania.GEN



Training Sample	Agree?
PROPN det DET	Yes
NOUN mod ADJ	Yes
PROPN mod NOUN	No

Leaf -1:
relation = **det**, head-POS = **NOUN, PROPN**, child-POS = *

Leaf -2:
relation = **mod**, head-POS = **NOUN, PROPN**, child-POS = **ADJ,PROPN**

Leaf-1:
Required-Agreement

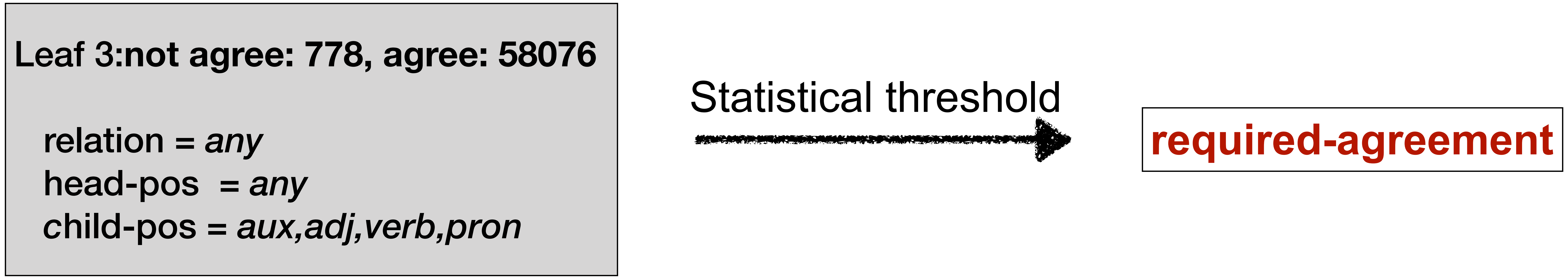
Leaf-2:
Chance-Agreement

Raw text → Extract syntactic features → Create Training data → Learn Model → Extract Rules

Rule Labeling

How do we assign a label of **required-agreement** to a leaf?

Each leaf induces a distribution of agreement over examples

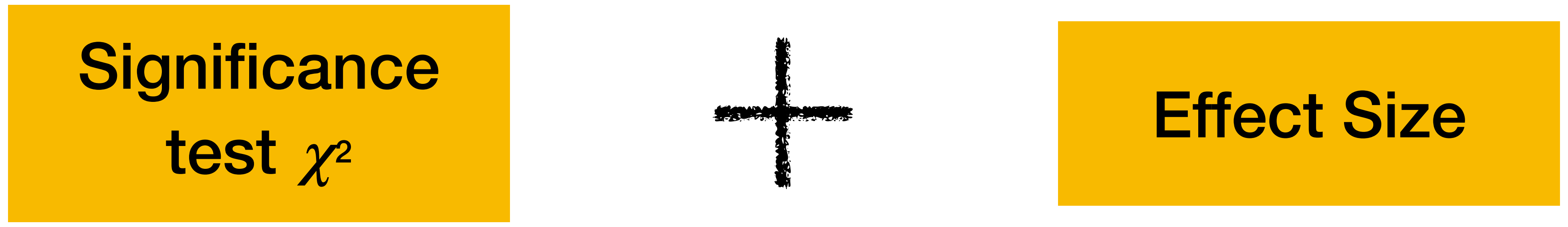


Leaf 3:not agree: 778, agree: 58076

relation = *any*
head-pos = *any*
child-pos = *aux,adj,verb,pron*

Statistical threshold

required-agreement



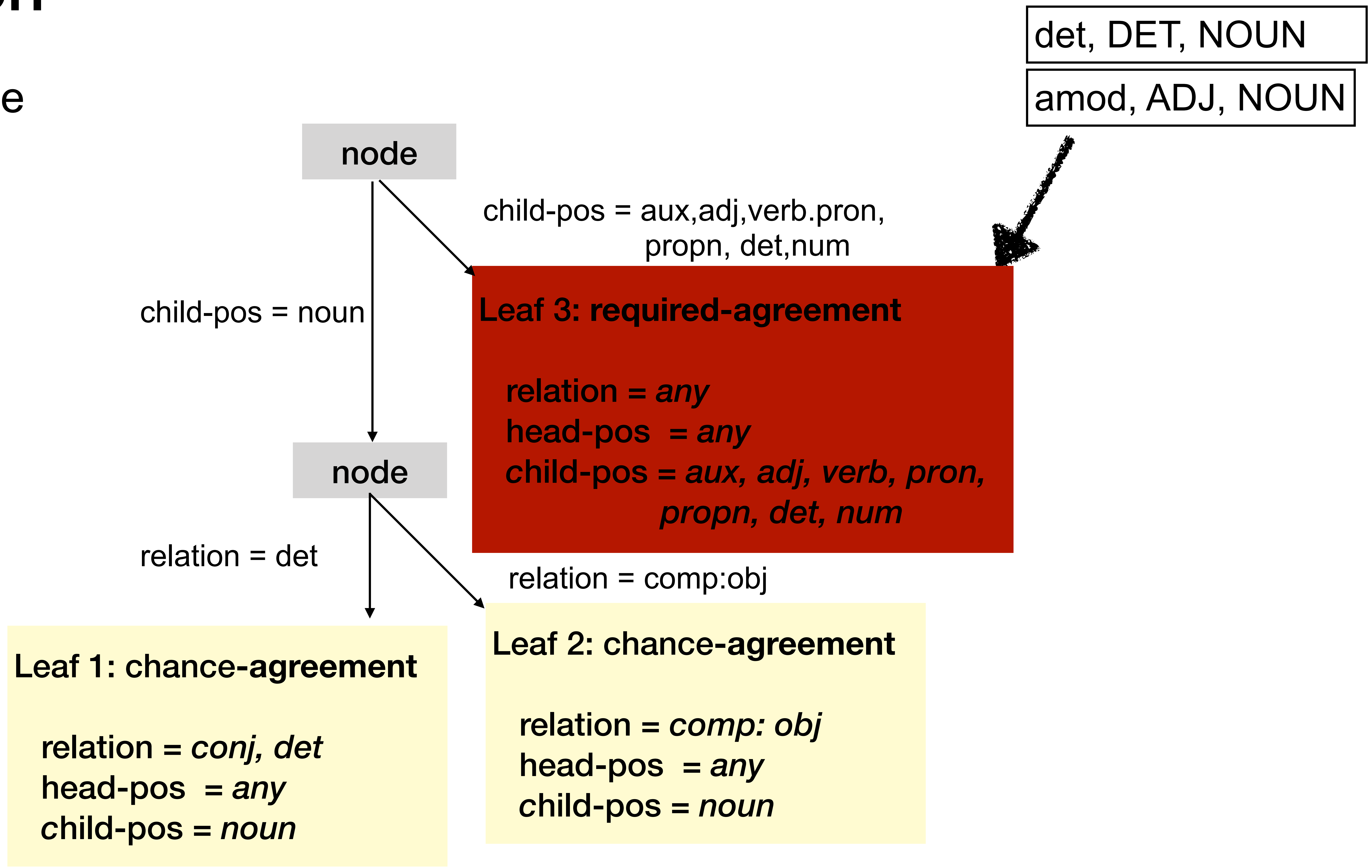
Significance test χ^2 + **Effect Size**

Observed agreement distribution is significant

Magnitude of significance is large

Rule Extraction

Labeled Decision Tree



det, DET, NOUN

amod, ADJ, NOUN

node

child-pos = aux,adj,verb.pron,
propn, det,num

Leaf 3: required-agreement

relation = *any*
head-pos = *any*
child-pos = *aux, adj, verb, pron,*
propn, det, num

child-pos = noun

node

relation = det

relation = comp:obj

Leaf 1: chance-agreement

relation = *conj, det*
head-pos = *any*
child-pos = *noun*

Leaf 2: chance-agreement

relation = *comp: obj*
head-pos = *any*
child-pos = *noun*

Spanish Gender Agreement

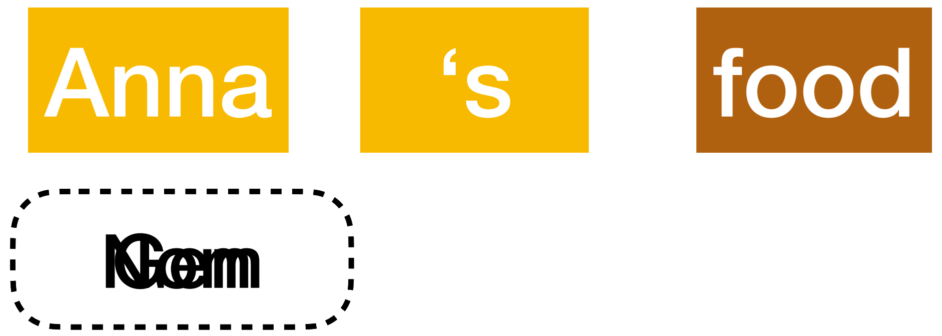
Formulate each linguistic question as a prediction task!

Agreement

When do syntactic heads show morphological agreement (e.g. gender agreement) with their dependents

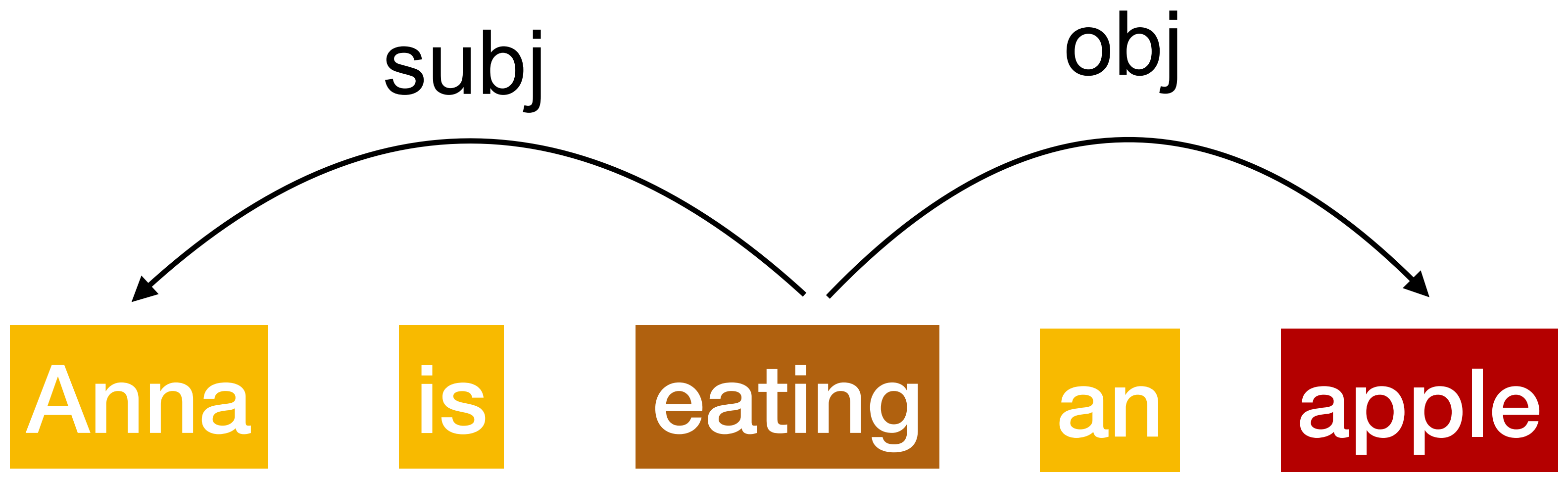
Case Marking

When does a particular class of words (e.g. nouns) take one value (e.g. nominative) over the other?

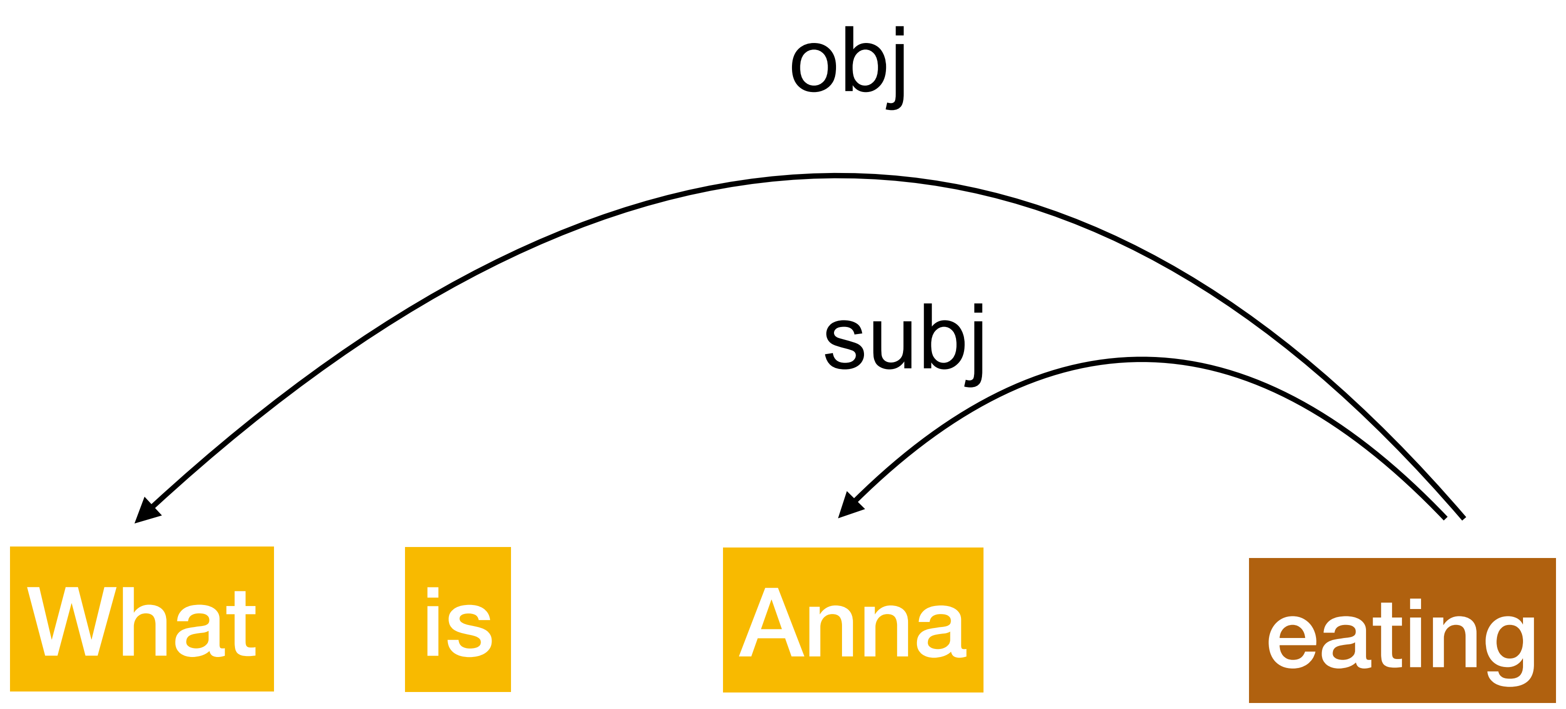


Word Order

When is one word order (e.g. subject-verb) predominant over the other?



Subject Verb Object

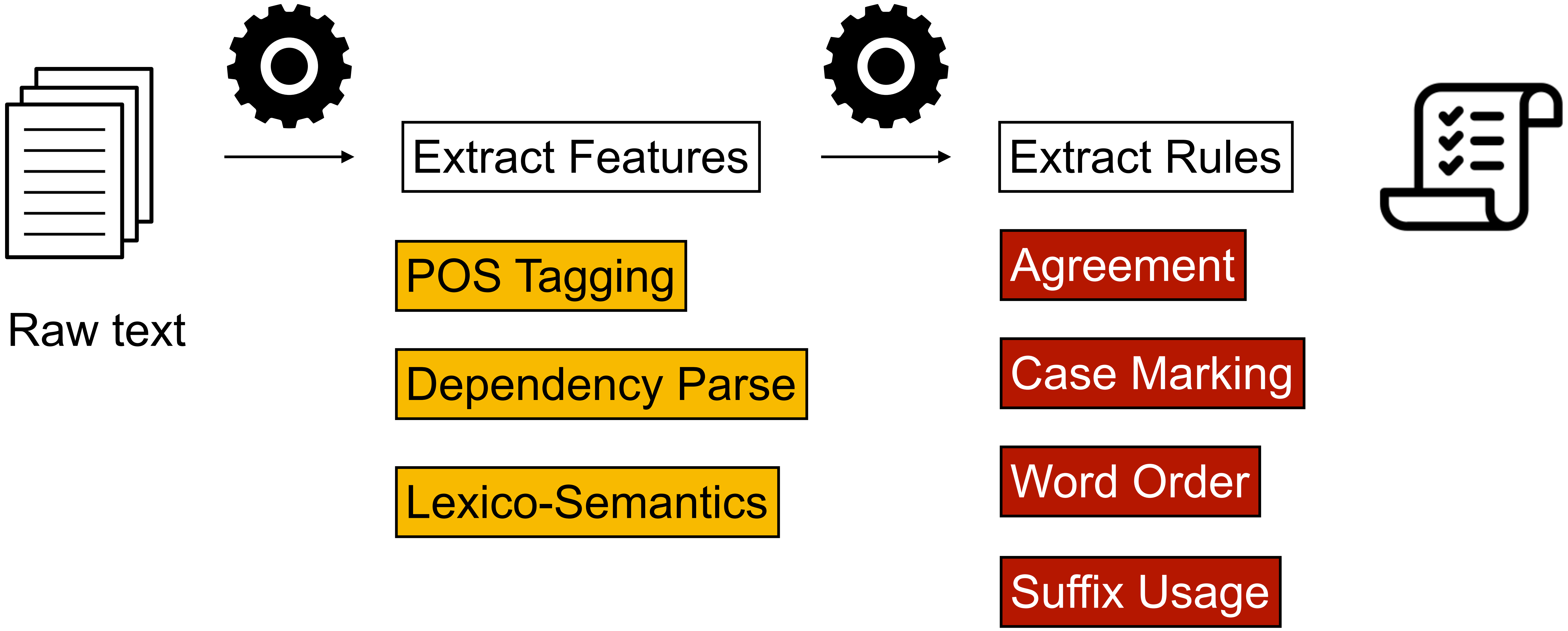


Object Subject Verb

Deriving Linguistic Insights

General framework to extract descriptions for different linguistic phenomena

Assumption: Linguistic phenomena can be explained by syntactic/semantic criteria



Raw text

Extract Features

POS Tagging

Dependency Parse

Lexico-Semantics

Extract Rules

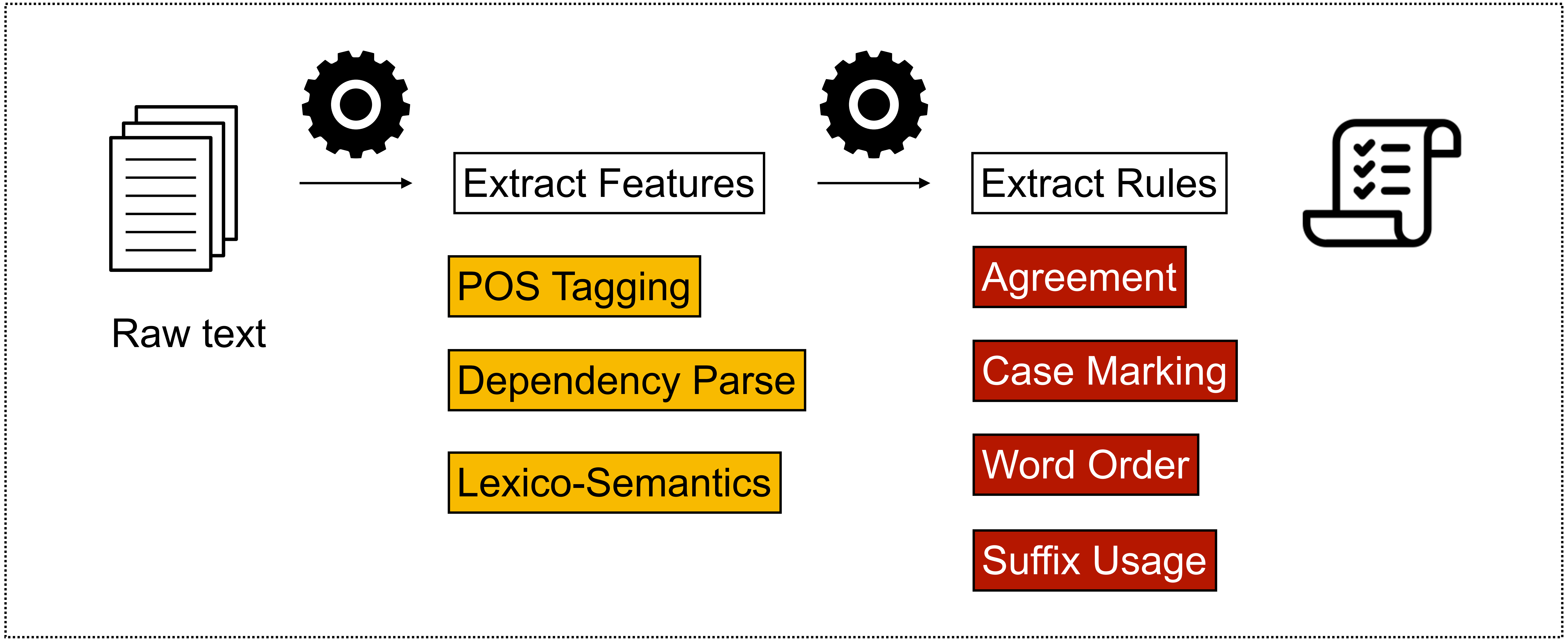
Agreement

Case Marking

Word Order

Suffix Usage

Deriving Linguistic Insights



Raw text

Extract Features

POS Tagging

Dependency Parse

Lexico-Semantics

Extract Rules

Agreement

Case Marking

Word Order

Suffix Usage

If expert syntactic analysis available

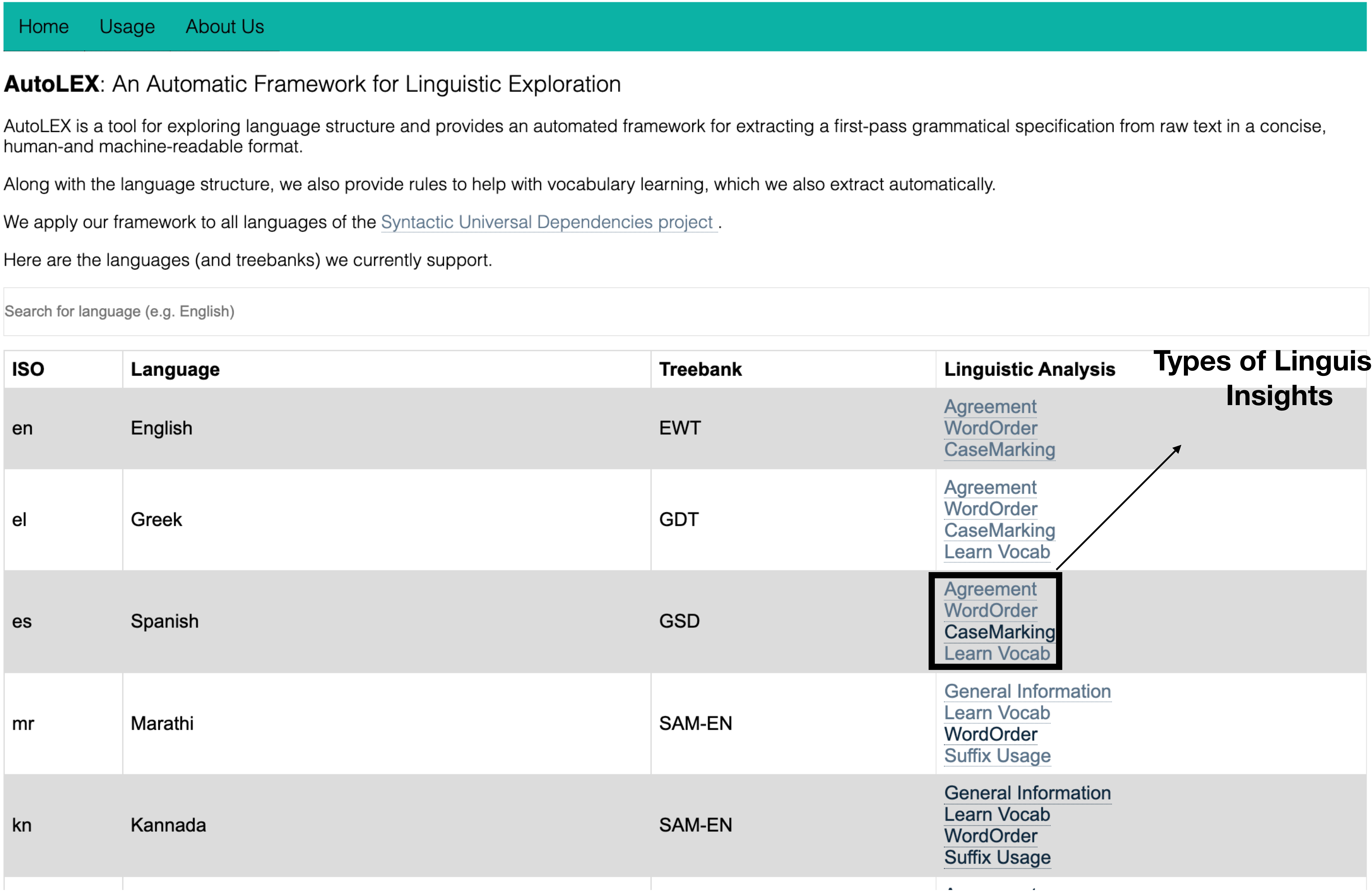
Yes

Syntactic Universal Dependency (SUD) Treebanks

No

Automatic Parser Multilingual model (e.g. UDIFY)

Carnegie Mellon University



AutoLEX: An Automatic Framework for Linguistic Exploration

AutoLEX is a tool for exploring language structure and provides an automated framework for extracting a first-pass grammatical specification from raw text in a concise, human-and machine-readable format.

Along with the language structure, we also provide rules to help with vocabulary learning, which we also extract automatically.

We apply our framework to all languages of the Syntactic Universal Dependencies project .

Here are the languages (and treebanks) we currently support.

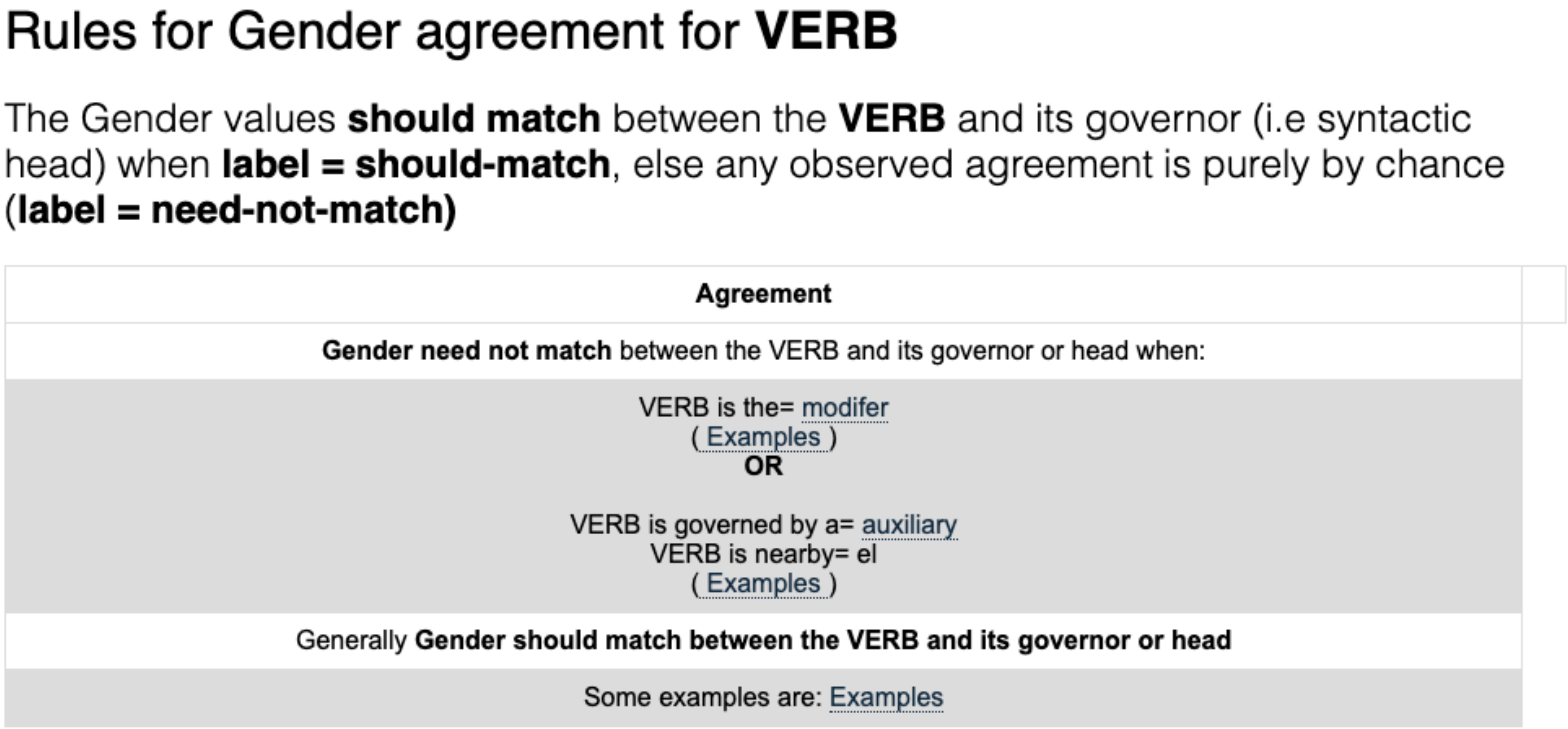
Search for language (e.g. English)

ISO	Language	Treebank	Linguistic Analysis
en	English	EWT	Agreement WordOrder CaseMarking
el	Greek	GDT	Agreement WordOrder CaseMarking Learn Vocab
es	Spanish	GSD	Agreement WordOrder CaseMarking Learn Vocab
mr	Marathi	SAM-EN	General Information Learn Vocab WordOrder Suffix Usage
kn	Kannada	SAM-EN	General Information Learn Vocab WordOrder Suffix Usage

Types of Linguistic Insights

Example Grammar Rules

- Automatically extracted rules for Spanish gender agreement



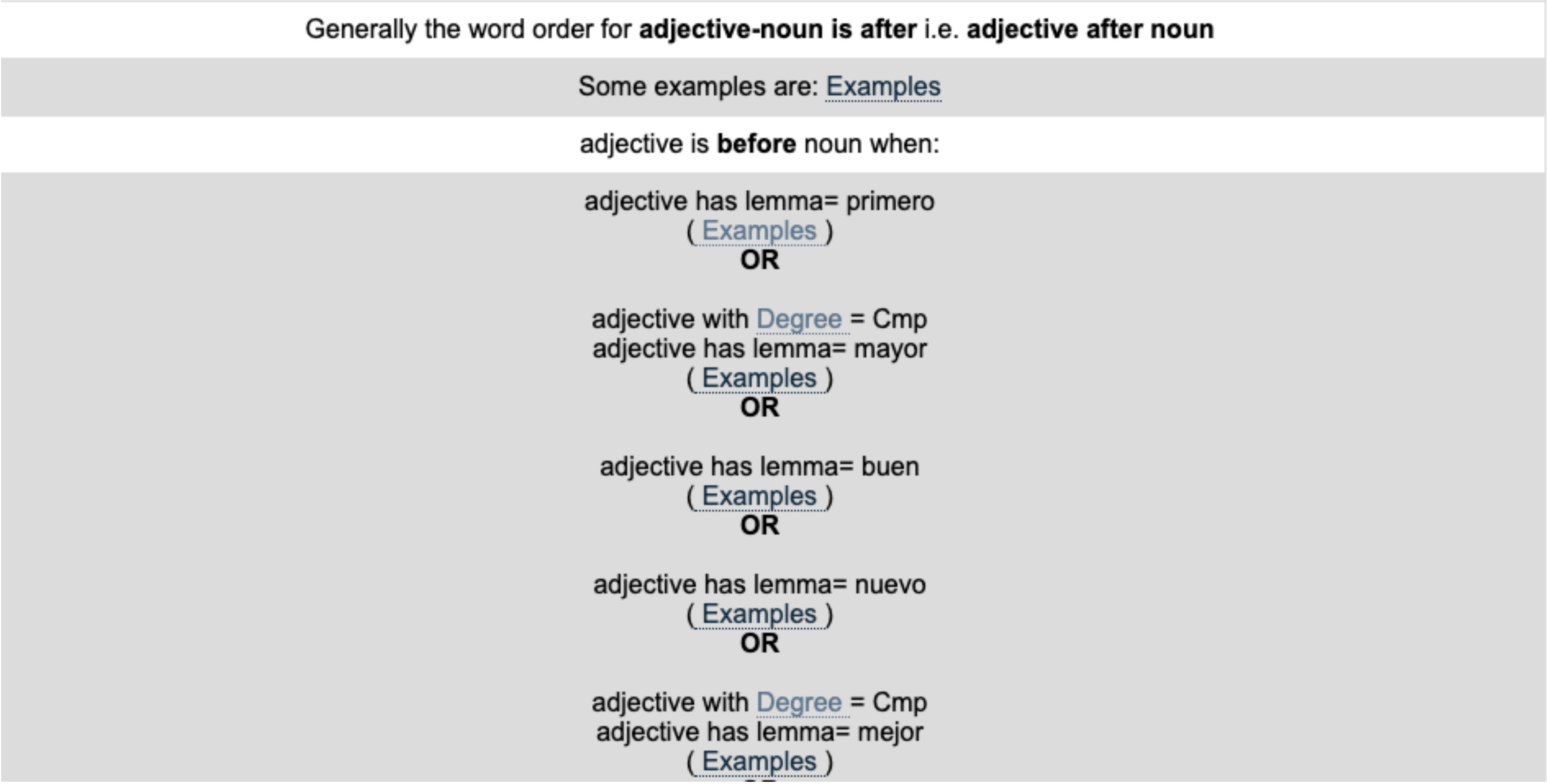
Rules for Gender agreement for **VERB**

The Gender values **should match** between the **VERB** and its governor (i.e syntactic head) when **label = should-match**, else any observed agreement is purely by chance **(label = need-not-match)**

Agreement
Gender need not match between the VERB and its governor or head when:
VERB is the= modifer (Examples) **OR** VERB is governed by a= auxiliary VERB is nearby= el (Examples)
Generally **Gender should match between the VERB and its governor or head**
Some examples are: Examples

Example Grammar Rules

- Automatically extracted rules for Spanish adjective-noun word order



Generally the word order for **adjective-noun is after** i.e. **adjective after noun**

Some examples are: Examples

adjective is **before** noun when:

adjective has lemma= primero
(Examples)
OR

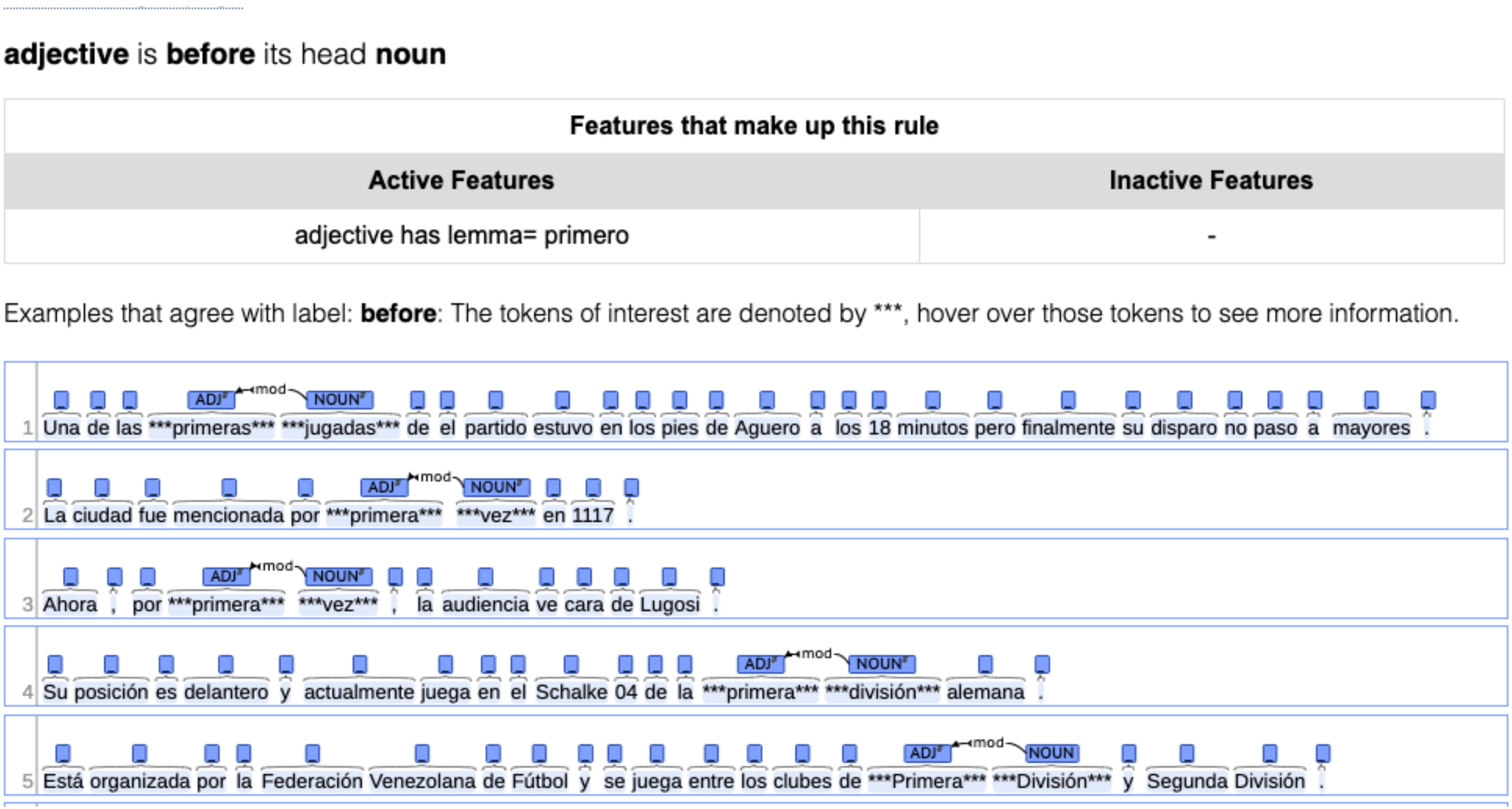
adjective with Degree = Cmp
adjective has lemma= mayor
(Examples)
OR

adjective has lemma= buen
(Examples)
OR

adjective has lemma= nuevo
(Examples)
OR

adjective with Degree = Cmp
adjective has lemma= mejor
(Examples)

Illustrative Examples



adjective is **before** its head **noun**

Features that make up this rule	
Active Features	**Inactive Features**
adjective has lemma= primero	-

Examples that agree with label: **before**: The tokens of interest are denoted by ***, hover over those tokens to see more information.

1 Una de las ***primeras*** ***jugadas*** de el partido estuvo en los pies de Aguero a los 18 minutos pero finalmente su disparo no paso a mayores .

2 La ciudad fue mencionada por ***primera*** ***vez*** en 1117 .

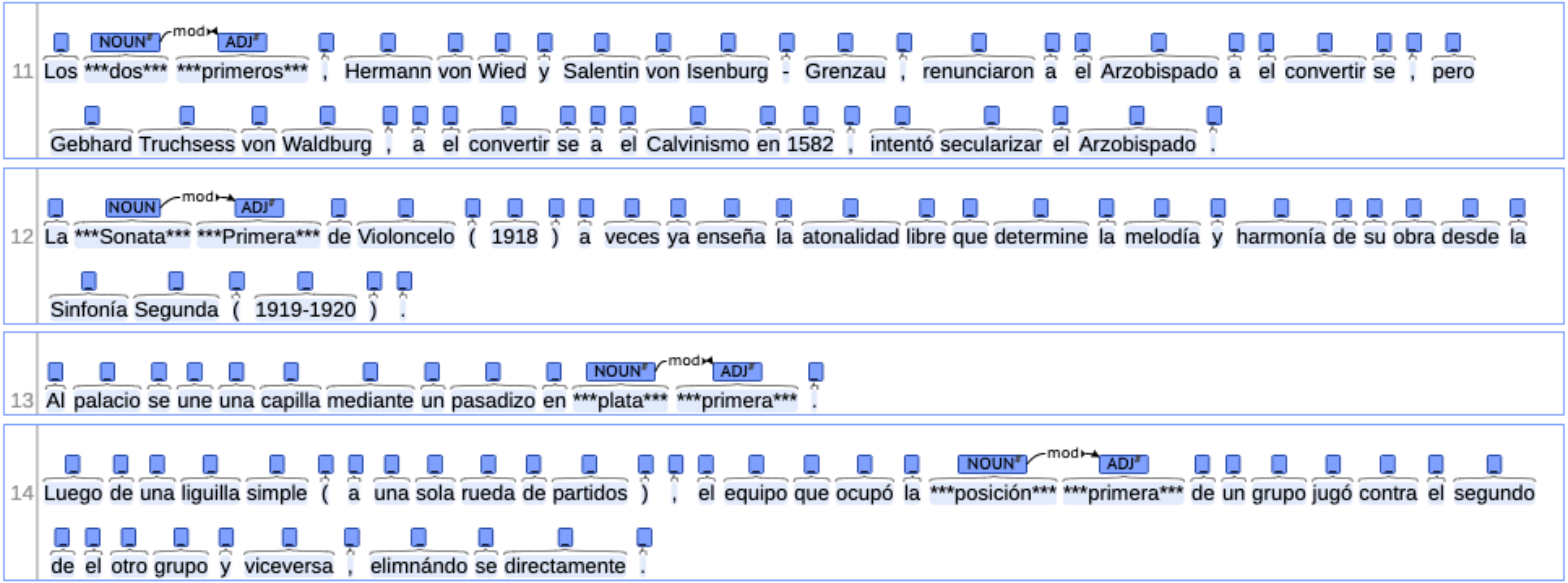
3 Ahora , por ***primera*** ***vez*** , la audiencia ve cara de Lugosi .

4 Su posición es delantero y actualmente juega en el Schalke 04 de la ***primera*** ***división*** alemana .

5 Está organizada por la Federación Venezolana de Fútbol y se juega entre los clubes de ***Primera*** ***División*** y Segunda División .

Illustrative Examples

Examples that disagree with the label: **before**



11 Los ***dos*** ***primeros*** , Hermann von Wied y Salentin von Isenburg - Grenzau , renunciaron a el Arzobispado a el convertir se , pero Gebhard Truchsess von Waldburg , a el convertir se a el Calvinismo en 1582 , intentó secularizar el Arzobispado .

12 La ***Sonata*** ***Primera*** de Violoncelo (1918) a veces ya enseña la atonalidad libre que determine la melodía y harmonía de su obra desde la Sinfonía Segunda (1919-1920) .

13 Al palacio se une una capilla mediante un pasadizo en ***plata*** ***primera*** .

14 Luego de una liguilla simple (a una sola rueda de partidos) , el equipo que ocupó la ***posición*** ***primera*** de un grupo jugó contra el segundo de el otro grupo y viceversa , elimnándo se directamente .

Other Applications

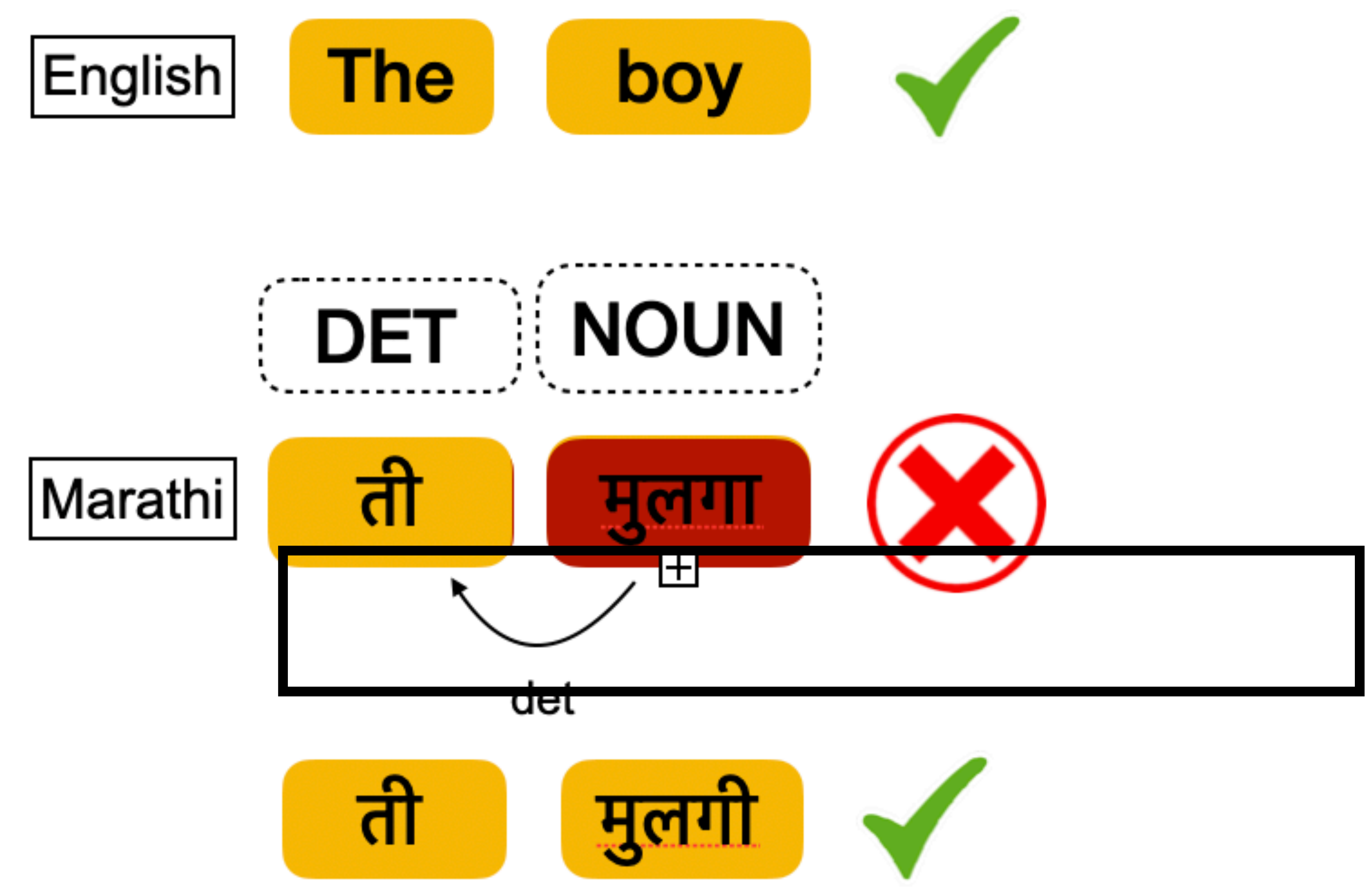
Automatic Grammar Rule Extraction

Image credit: Adithya Pratapa

Errors

case assi

e assignment



Machine Output

Determiners need to agree with nouns on gender

2020

of Rules Governing Morphological Agreement
sopoulos, Pratapa, Mortensen, Sheikh, Tsvetkov, Neubig. ***EMNLP 2020***

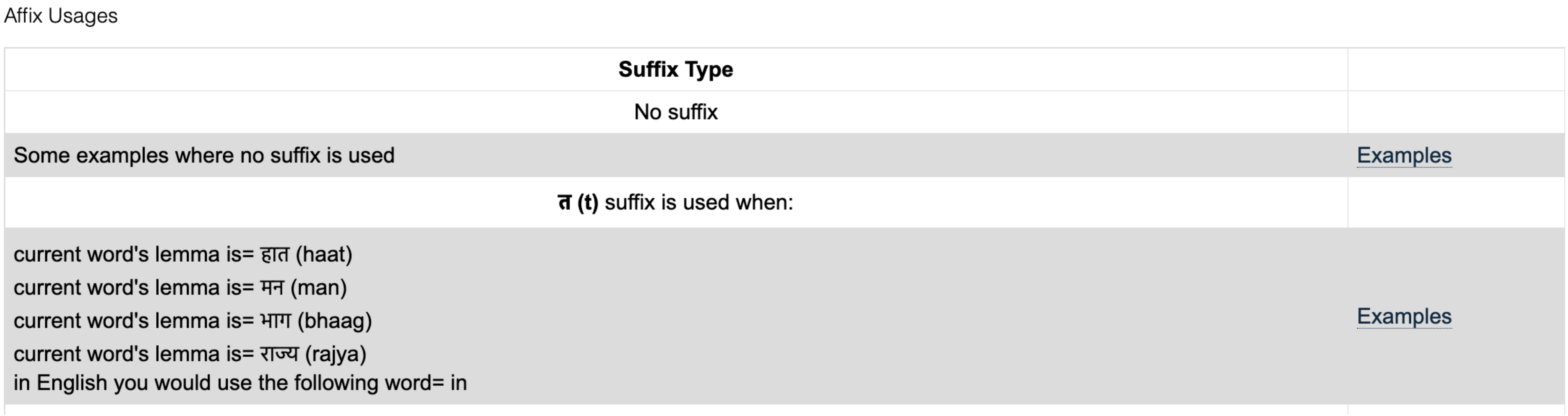
LANGUAGE DOCUMENTATION & CONSERVATION

AutoLEX: An Automatic Framework for Linguistic Exploration
Chaudhary**, Sheikh, Mortensen, Anastasopoulos, Neubig.* ***In Submission

Carnegie Mellon University

Teach a Language

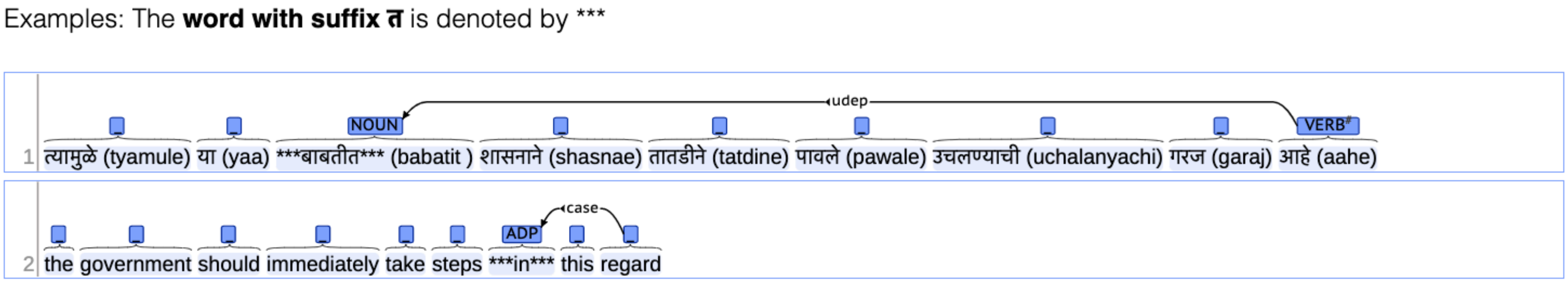
What suffix to use?



Affix Usages

Suffix Type	
No suffix	
Some examples where no suffix is used	Examples
त (t) suffix is used when:	
current word's lemma is= हात (haat) current word's lemma is= मन (man) current word's lemma is= भाग (bhaag) current word's lemma is= राज्य (rajya) in English you would use the following word= in	Examples

Tells you that (t) suffix is used when you need to say “in … “ English!



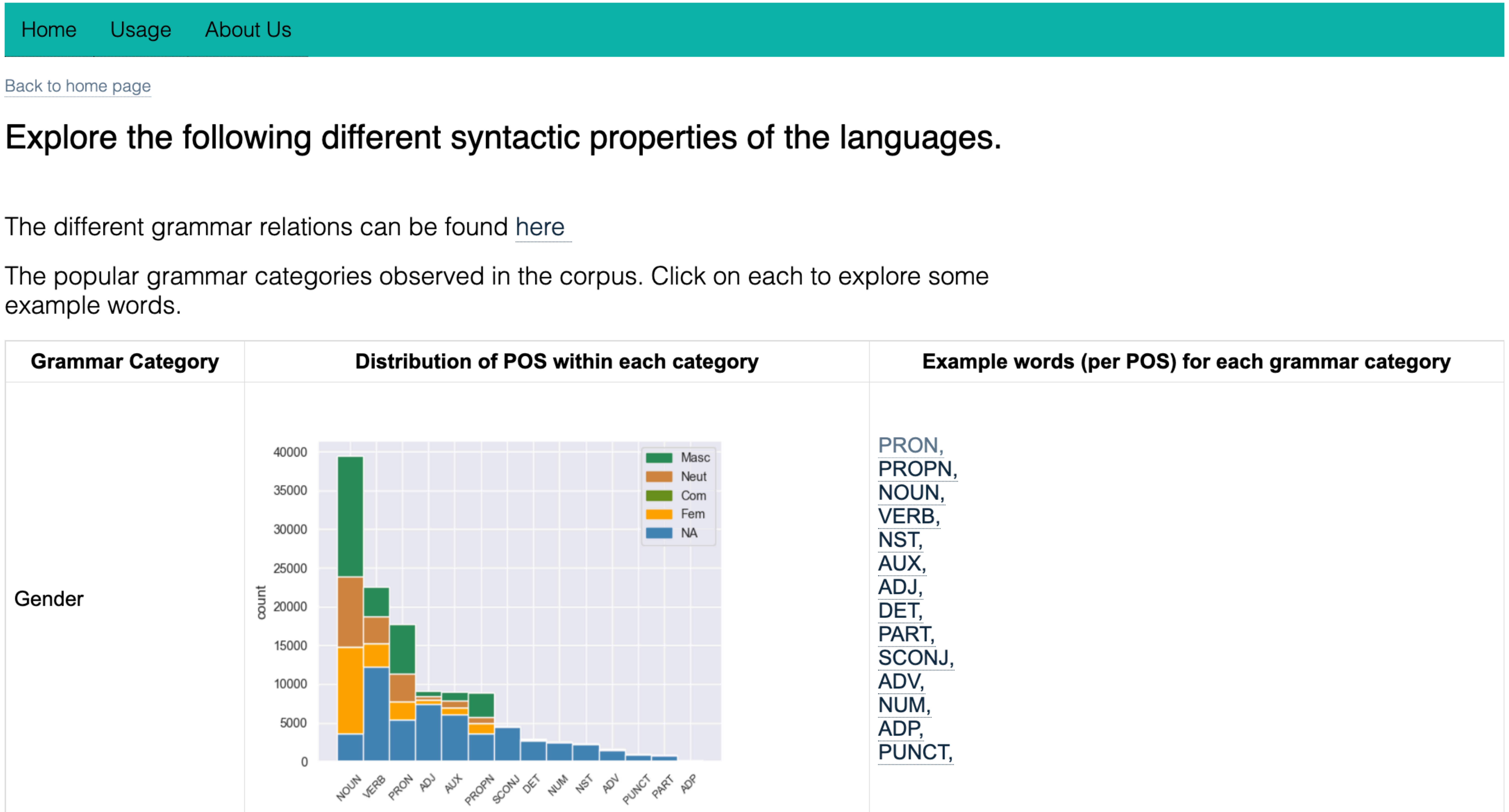
Examples: The **word with suffix त** is denoted by ***

1 त्यामुळे (tyamule) या (yaa) ***बाबतीत*** (babatit) शासनाने (shasnae) तातडीने (tatdine) पावले (pawale) उचलण्याची (uchalanyachi) गरज (garaj) आहे (aahe)

2 the government should immediately take steps ***in*** this regard

General Information of the Language

Understanding the language properties at a glance!



Explore the following different syntactic properties of the languages.

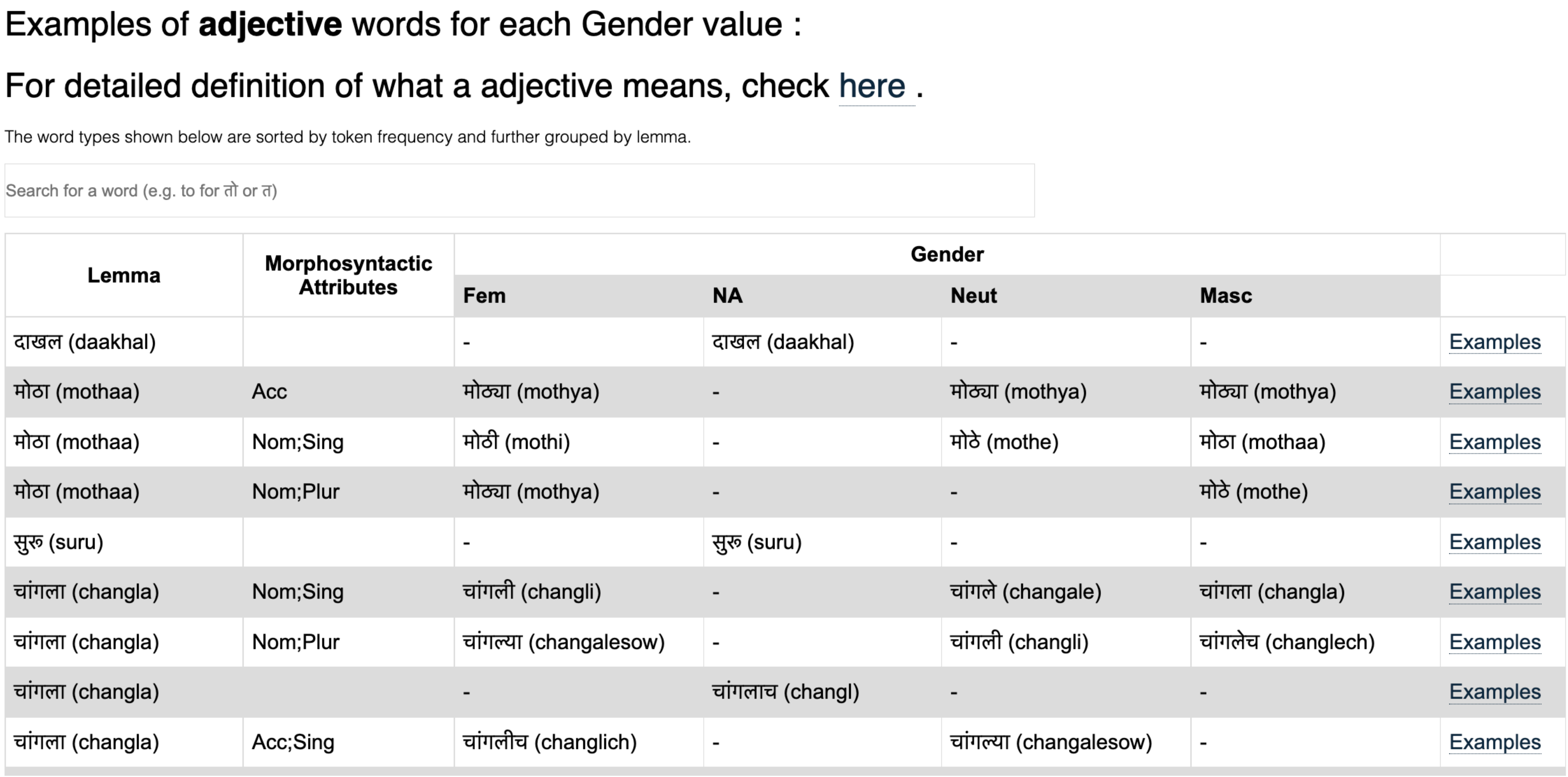
The different grammar relations can be found here

The popular grammar categories observed in the corpus. Click on each to explore some example words.

Grammar Category	Distribution of POS within each category	Example words (per POS) for each grammar category
Gender		PRON, PROPN, NOUN, VERB, NST, AUX, ADJ, DET, PART, SCONJ, ADV, NUM, ADP, PUNCT,

Informs us that Marathi nouns, verbs and pronouns have 4 genders, and esp. nouns exhibit 3 of those almost equally!

General Information of the Language



Examples of **adjective** words for each Gender value :

For detailed definition of what a adjective means, check here .

The word types shown below are sorted by token frequency and further grouped by lemma.

Search for a word (e.g. to for तो or त)

Lemma	Morphosyntactic Attributes	Gender				
		Fem	NA	Neut	Masc	
दाखल (daakhal)		-	दाखल (daakhal)	-	-	Examples
मोठा (mothaa)	Acc	मोठ्या (mothya)	-	मोठ्या (mothya)	मोठ्या (mothya)	Examples
मोठा (mothaa)	Nom;Sing	मोठी (mothi)	-	मोठे (mothe)	मोठा (mothaa)	Examples
मोठा (mothaa)	Nom;Plur	मोठ्या (mothya)	-	-	मोठे (mothe)	Examples
सुरू (suru)		-	सुरू (suru)	-	-	Examples
चांगला (changla)	Nom;Sing	चांगली (changli)	-	चांगले (changale)	चांगला (changla)	Examples
चांगला (changla)	Nom;Plur	चांगल्या (changalesow)	-	चांगली (changli)	चांगलेच (changlech)	Examples
चांगला (changla)		-	चांगलाच (changl)	-	-	Examples
चांगला (changla)	Acc;Sing	चांगलीच (changlich)	-	चांगल्या (changalesow)	-	Examples

Tells you about the different lexical variations for each gender!

AutoLEX: https://aditi138.github.io/auto-lex-learn/index.html

Contact: aschaudh@andrew.cmu.edu