PanLex

THE LONG NOW FOUNDATION

Tabularization using xml parsing in Python

Gary Patterson

August 8, 2013

Common bilingual source formats

revivir *vi* chédlane', chela' ši'i
revocar *vt* checa'a
revolcar 1. *vt* chchix̱ chtoḻe
2. *v prnl* chas chata', chbix̱ chtoḻe
revolver *vt* 1. ch̲chix̱e, chyitje
2. chta *(líquido)*
rey *m* rey
rezar *vi* chon rsar

Observations

- Five different 'categories' of information:
  - L1 expression (e.g. Spanish)
  - part(s)-of-speech
  - L2 expressions (e.g. Zapoteco), including synonym delimiters
  - numerical meaning delimiter
  - parenthetical information
- A fixed set of possible transitions between the different categories
- Each category associated with a consistent typeface

Observations

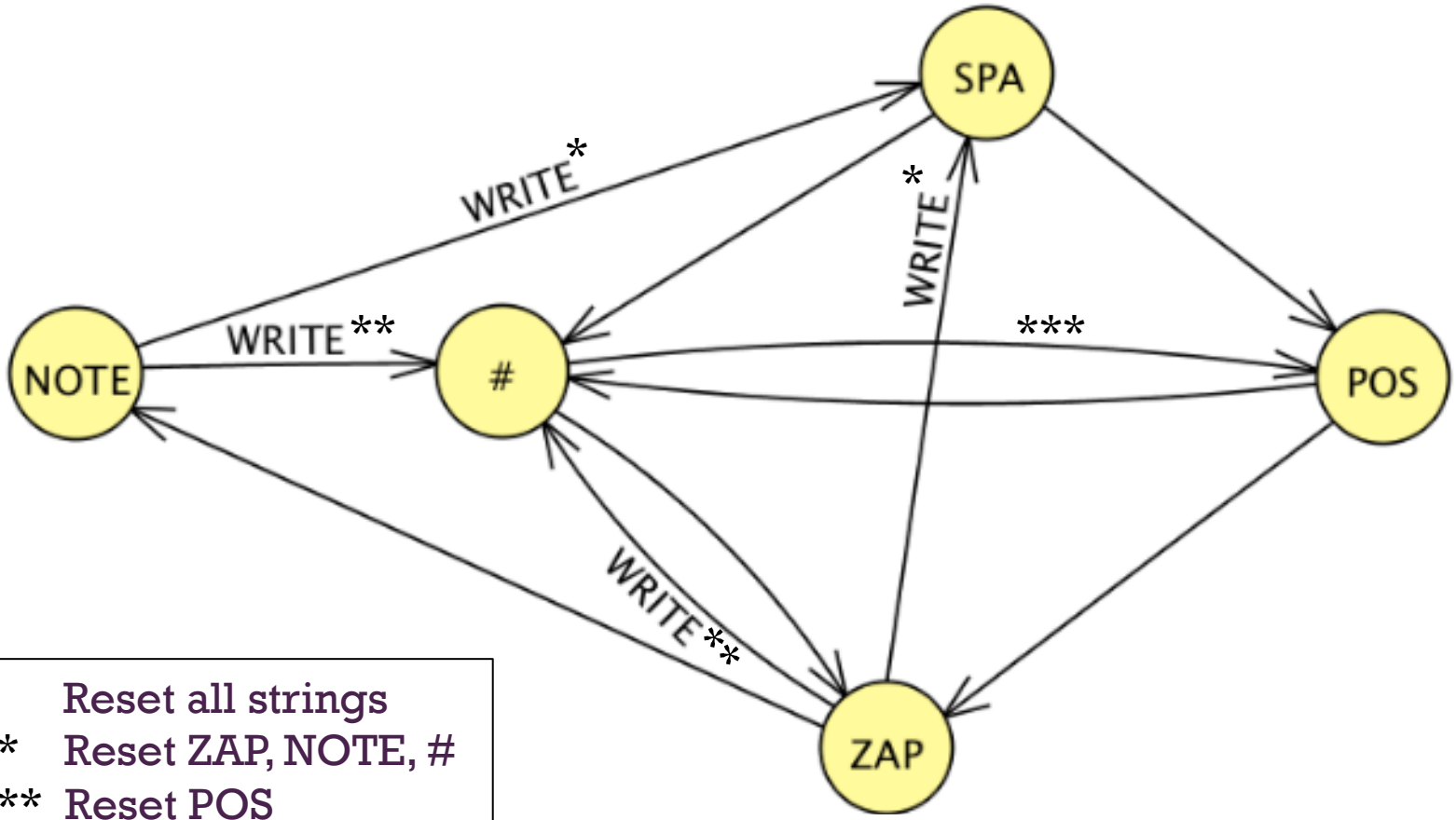
- Five different 'categories' of information:
  - L1 expression (e.g. Spanish)
  - part(s)-of-speech
  - L2 expressions (e.g. Zapoteco), including synonym delimiters
  - numerical meaning delimiter
  - parenthetical information
- A fixed set of possible transitions between the different categories
- Each category associated with a consistent typeface

Can we use this typographical information (available in xml dump of pdf file) and the known transitions between categories to parse the file more efficiently than by using regular expressions?

xml file

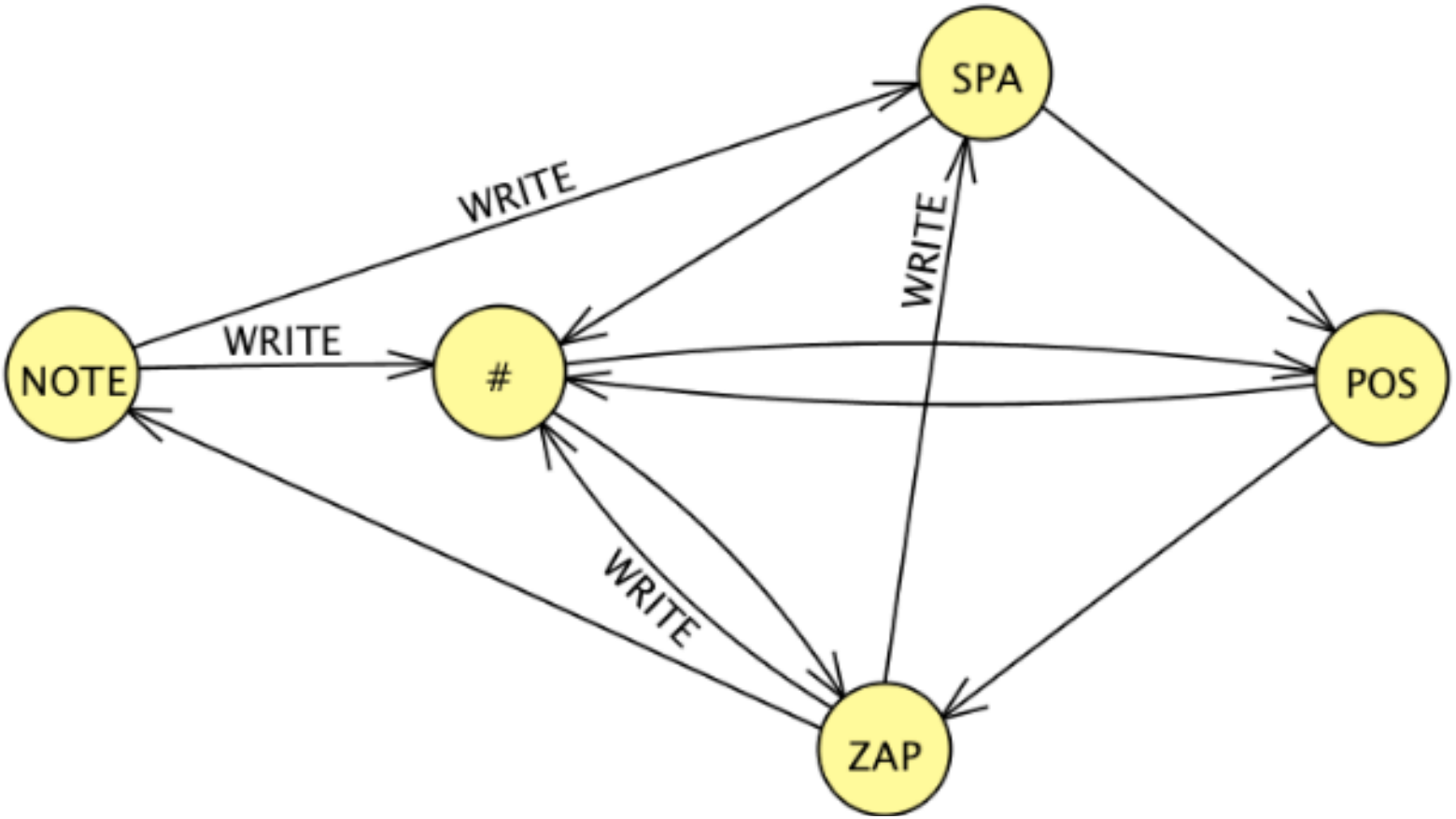

```
<text font="LongZapSILDoulosL,Bold" bbox="132.117,369.738,137.560,383.538" size="13.800">v</text>
<text font="LongZapSILDoulosL,Bold" bbox="137.560,369.738,143.004,383.538" size="13.800">o</text>
<text font="LongZapSILDoulosL,Bold" bbox="143.004,369.738,146.118,383.538" size="13.800">l</text>
<text font="LongZapSILDoulosL,Bold" bbox="146.118,369.738,150.977,383.538" size="13.800">c</text>
<text font="LongZapSILDoulosL,Bold" bbox="150.977,369.738,156.420,383.538" size="13.800">a</text>
<text font="LongZapSILDoulosL,Bold" bbox="156.420,369.738,161.279,383.538" size="13.800">r</text>
<text font="LongZapSILDoulosL" bbox="161.280,369.658,164.304,382.560" size="12.902"> </text>
<text font="LongZapSILDoulosL" bbox="164.400,369.658,167.424,382.560" size="12.902"> </text>
<text font="LongZapSILDoulosL" bbox="167.520,369.658,170.544,382.560" size="12.902"> </text>
<text font="LongZapSILDoulosL" bbox="170.640,369.658,176.184,382.560" size="12.902">1</text>
<text font="LongZapSILDoulosL" bbox="176.150,369.658,179.074,382.560" size="12.902">.</text>
<text font="LongZapSILDoulosL" bbox="179.040,369.658,182.064,382.560" size="12.902"> </text>
<text font="LongZapSILDoulosL,Italic" bbox="182.160,369.392,186.647,381.111" size="11.719">v</text>
<text font="LongZapSILDoulosL,Italic" bbox="186.669,369.392,189.578,381.111" size="11.719">t</text>
<text font="LongZapSILDoulosL" bbox="189.600,369.446,192.336,381.120" size="11.674"> </text>
<text font="LongZapSILDoulosL" bbox="192.240,369.446,194.976,381.120" size="11.674"> </text>
<text font="LongZapSILDoulosL" bbox="194.880,369.658,197.904,382.560" size="12.902"> </text>
<text font="LongZapSILDoulosL" bbox="198.000,369.658,202.959,382.560" size="12.902">c</text>
<text font="LongZapSILDoulosL" bbox="202.953,369.658,208.497,382.560" size="12.902">h</text>
<text font="LongZapSILDoulosL" bbox="208.491,369.658,213.451,382.560" size="12.902">c</text>
<text font="LongZapSILDoulosL" bbox="213.445,369.658,218.989,382.560" size="12.902">h</text>
<text font="LongZapSILDoulosL" bbox="218.983,369.658,222.198,382.560" size="12.902">i</text>
<text font="LongZapSILDoulosL" bbox="222.192,369.658,227.585,382.560" size="12.902">(cid:158)</text>
<text font="LongZapSILDoulosL" bbox="227.579,369.658,230.603,382.560" size="12.902"> </text>
<text font="LongZapSILDoulosL" bbox="230.699,369.658,235.658,382.560" size="12.902">c</text>
<text font="LongZapSILDoulosL" bbox="235.652,369.658,241.196,382.560" size="12.902">h</text>
<text font="LongZapSILDoulosL" bbox="241.190,369.658,244.406,382.560" size="12.902">t</text>
<text font="LongZapSILDoulosL" bbox="244.400,369.658,249.843,382.560" size="12.902">o</text>
<text font="LongZapSILDoulosL" bbox="249.837,369.658,253.053,382.560" size="12.902">(cid:153)</text>
```

Schematic Representation

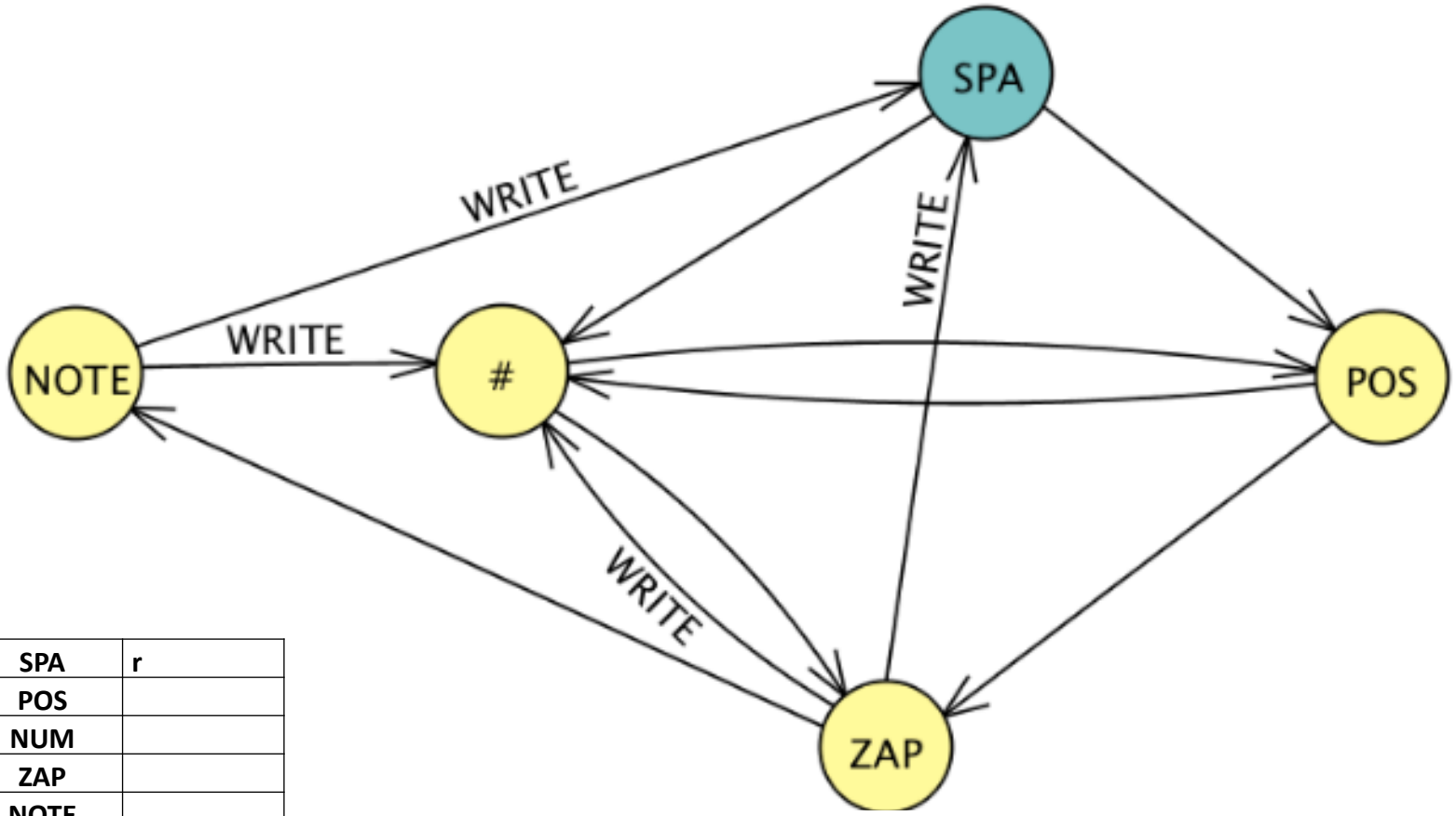


* Reset all strings
** Reset ZAP, NOTE, #
*** Reset POS

+ **revocar** *vt* checa'a
  revolcar 1. *vt* chtol 2. *v prnl* chata'
  revolver *vt* 1. chix 2. chta

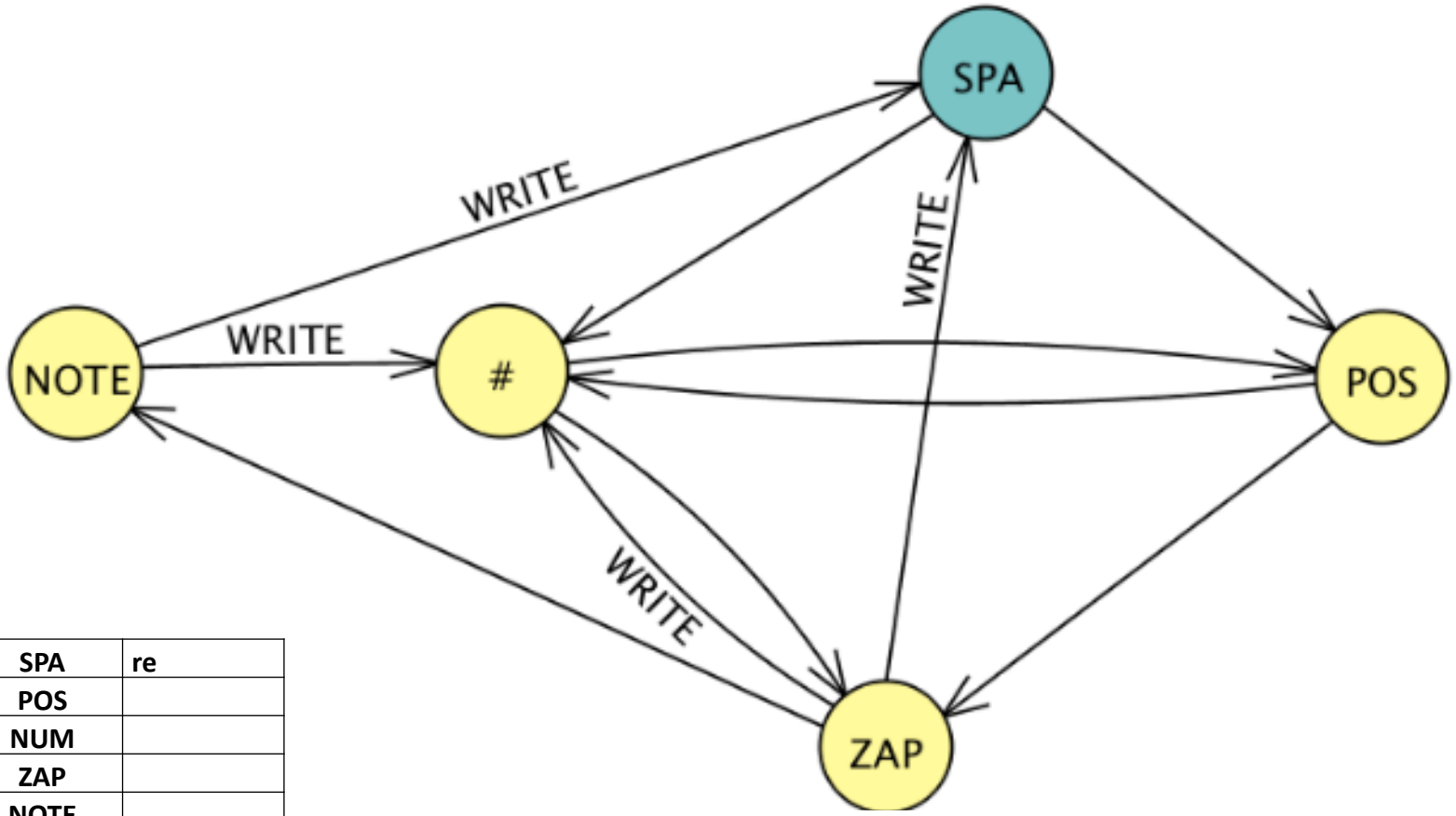


revocar *vt* checa'a **revolcar** …



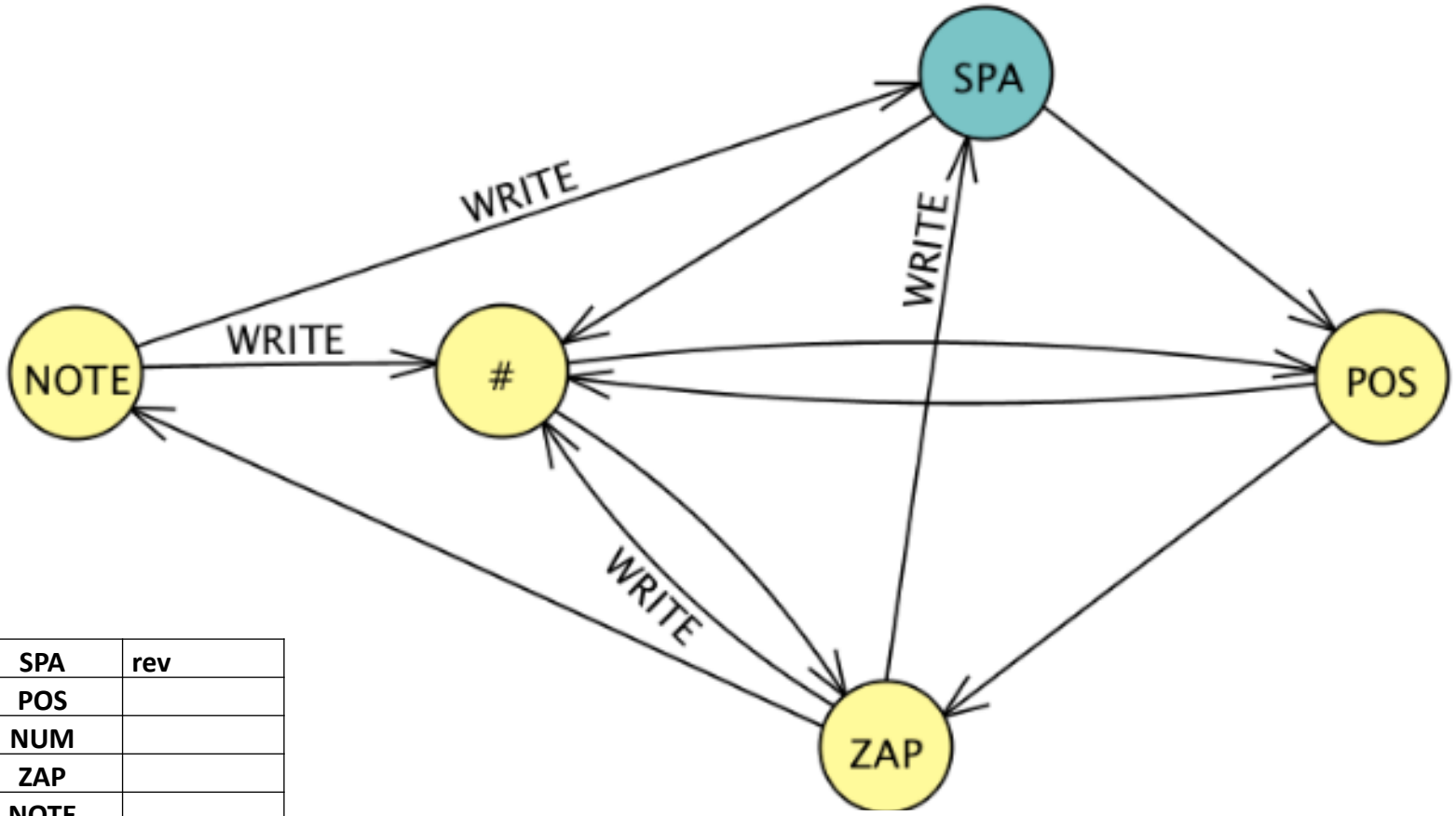
SPA	r
POS	
NUM	
ZAP	
NOTE	

revocar *vt* checa'a **revolcar** …



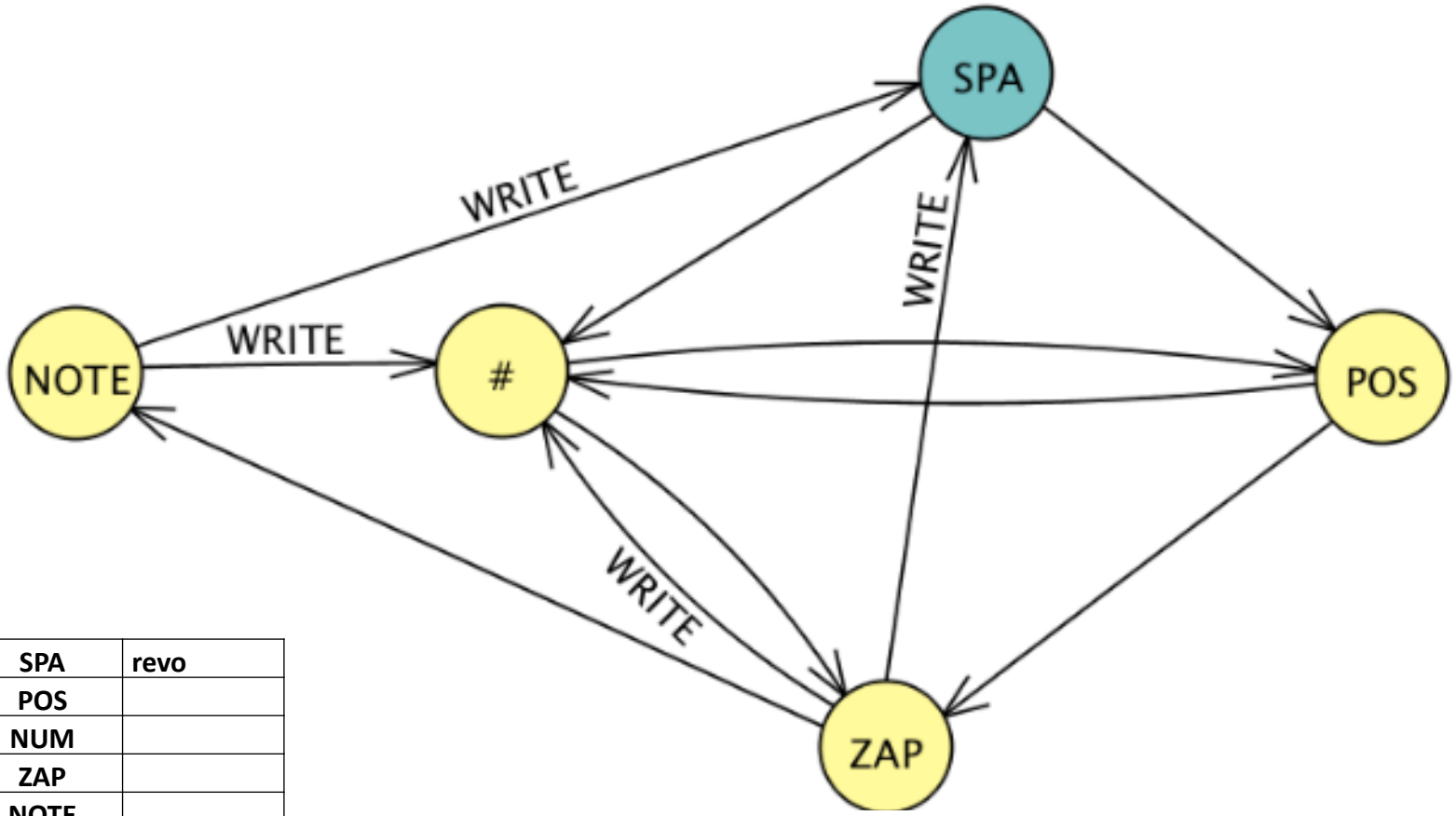
SPA	re
POS	
NUM	
ZAP	
NOTE	

revocar *vt* checa'a **revolcar** …



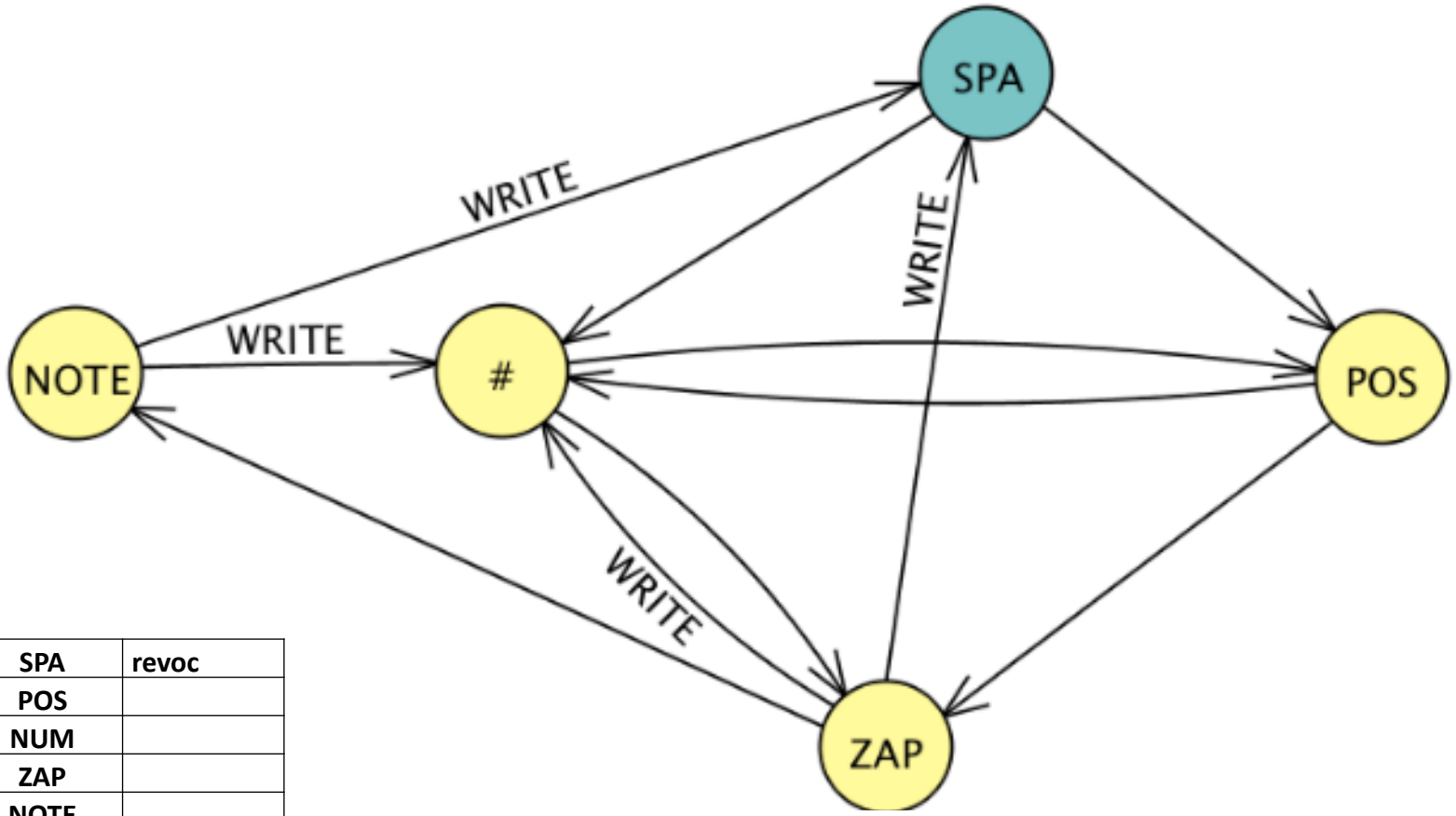
SPA	rev
POS	
NUM	
ZAP	
NOTE	

revocar *vt* checa'a **revolcar** …



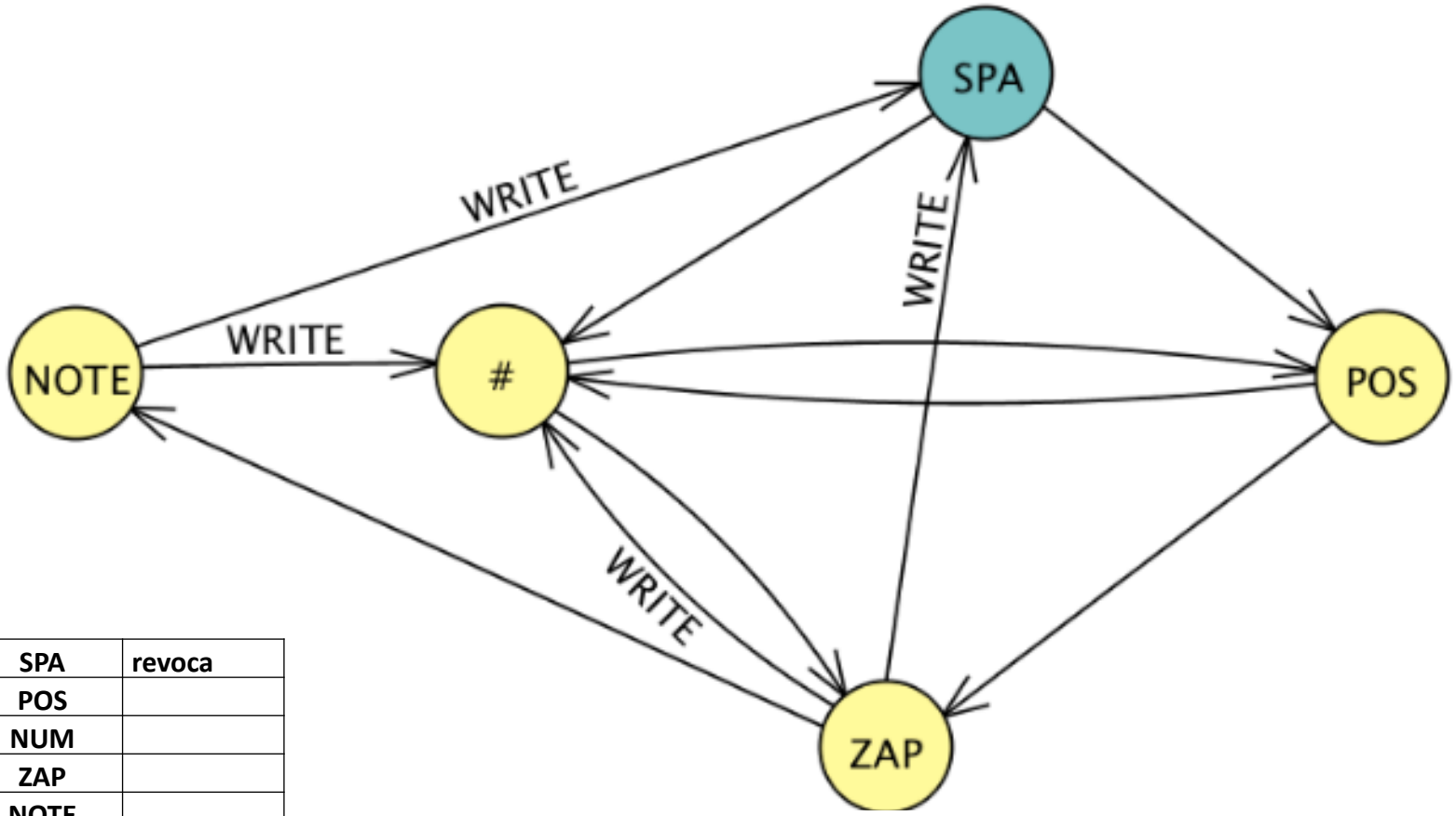
SPA	revo
POS	
NUM	
ZAP	
NOTE	

revo<u>c</u>ar *vt* checa'a
revolcar …



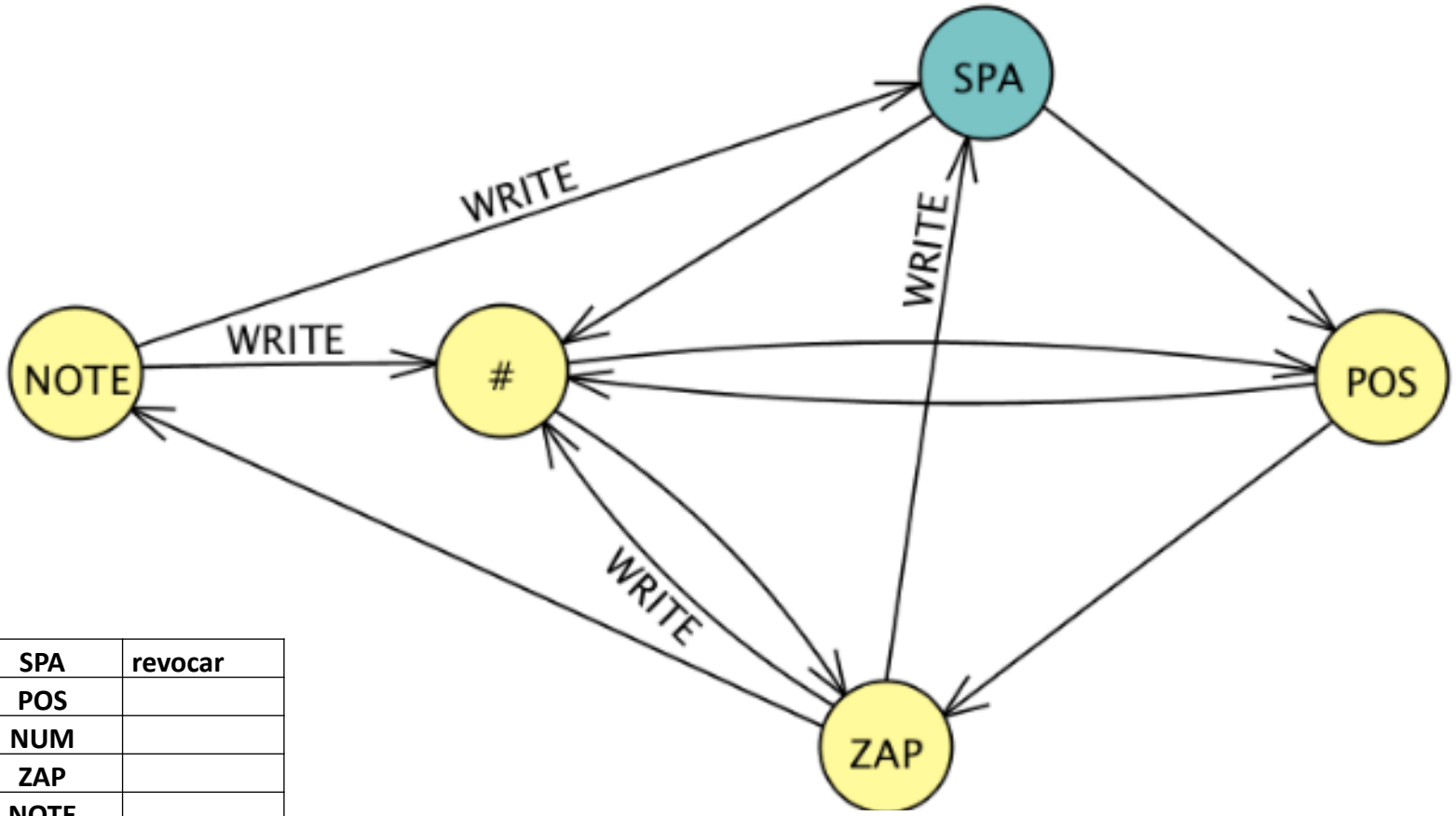
SPA	revoc
POS	
NUM	
ZAP	
NOTE	

revoc<u>a</u>r *vt* checa'a
revolcar …



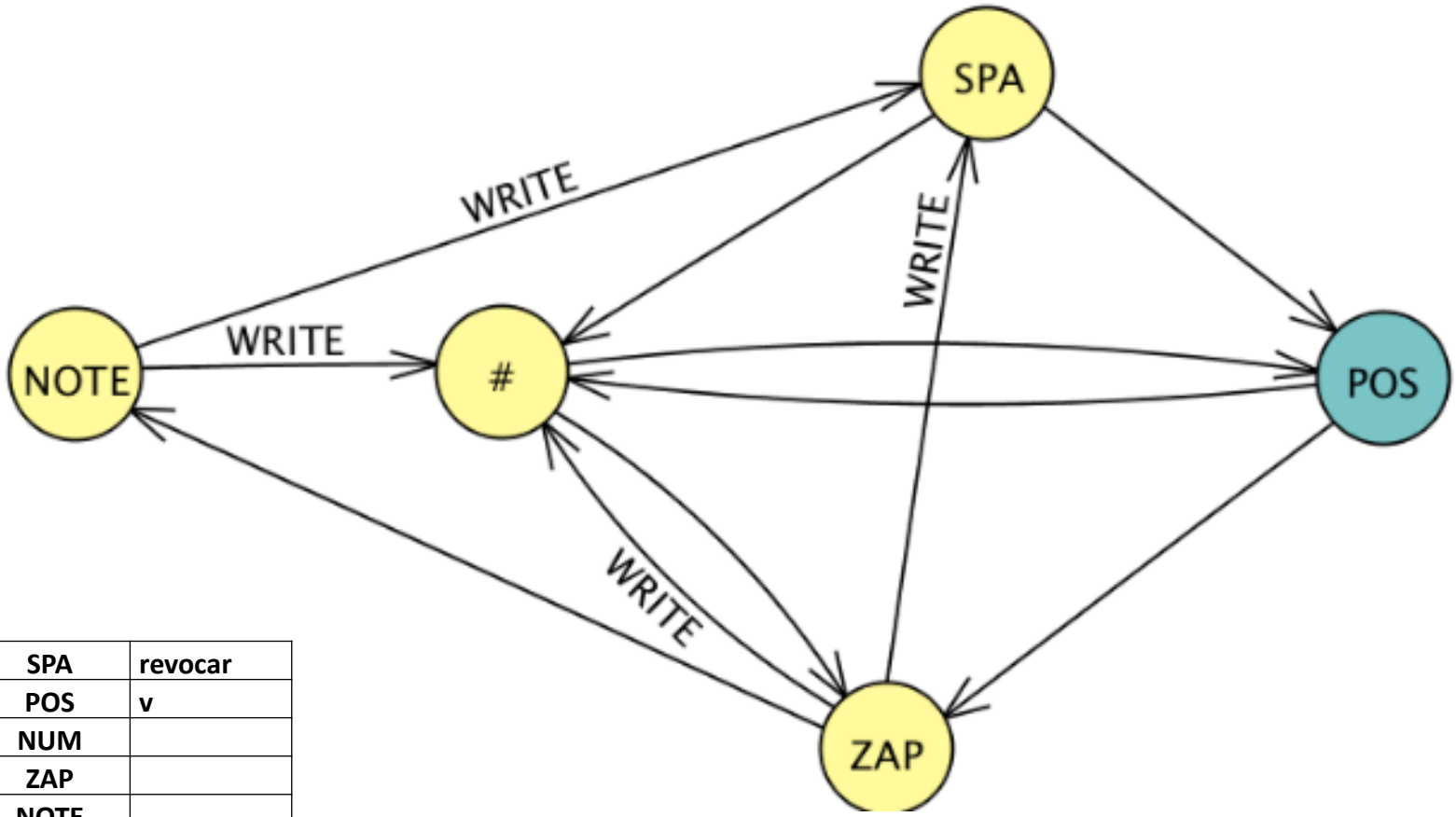
SPA	revoca
POS	
NUM	
ZAP	
NOTE	

revocar *vt* checa'a **revolcar** …



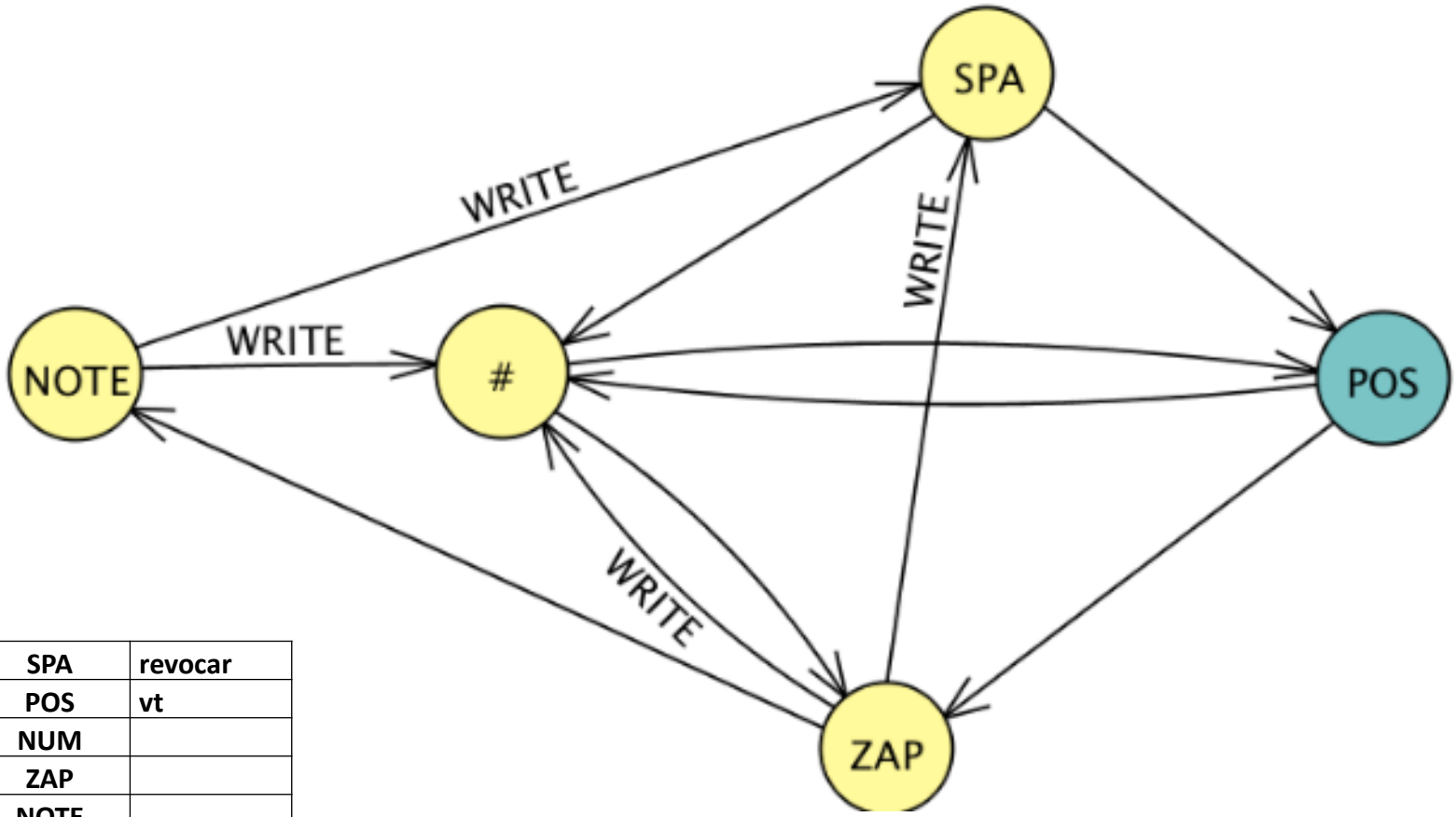
SPA	revocar
POS	
NUM	
ZAP	
NOTE	

revocar *vt* checa'a **revolcar** …



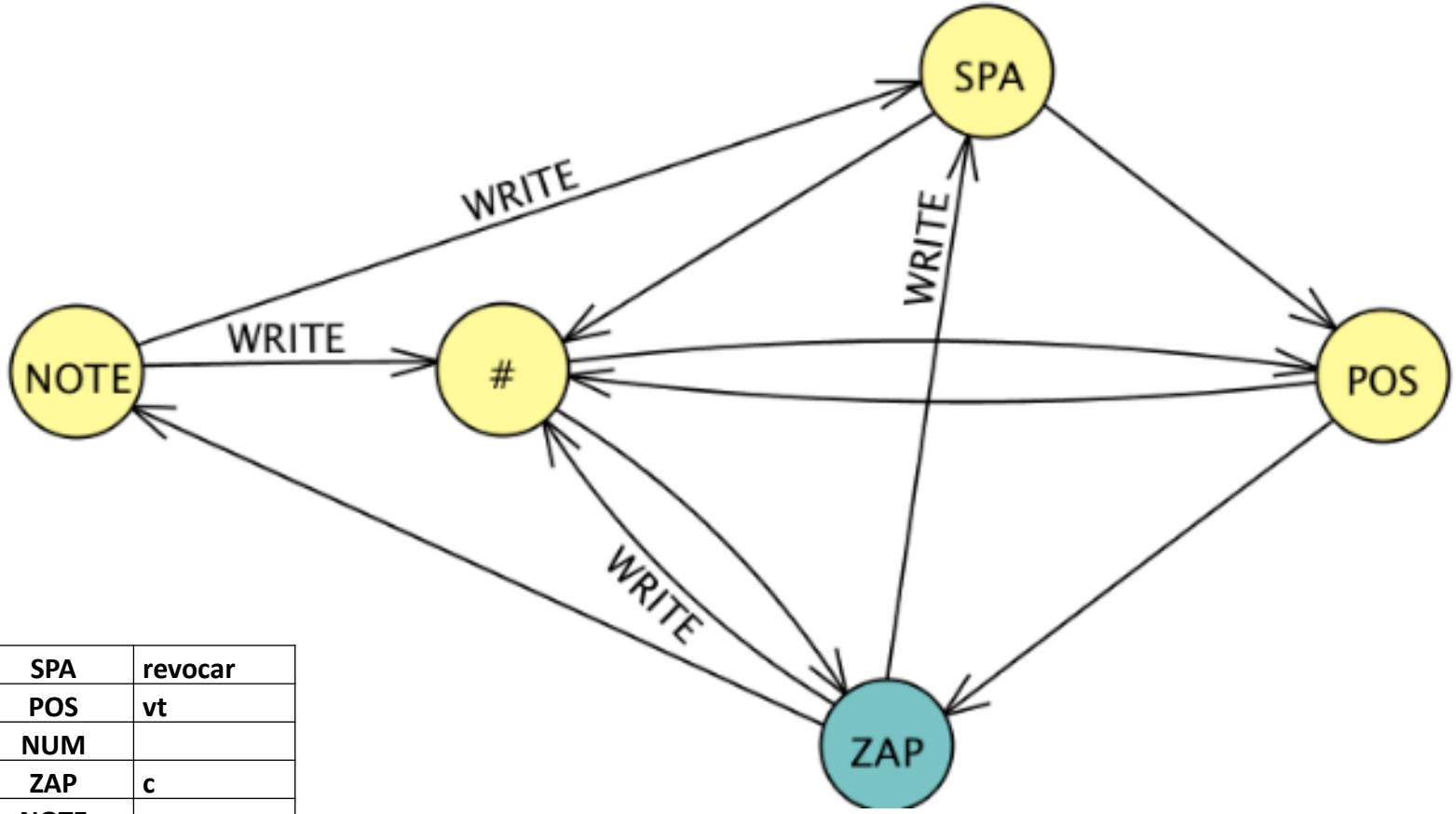
SPA	revocar
POS	v
NUM	
ZAP	
NOTE	

revocar *vt* checa'a **revolcar** …



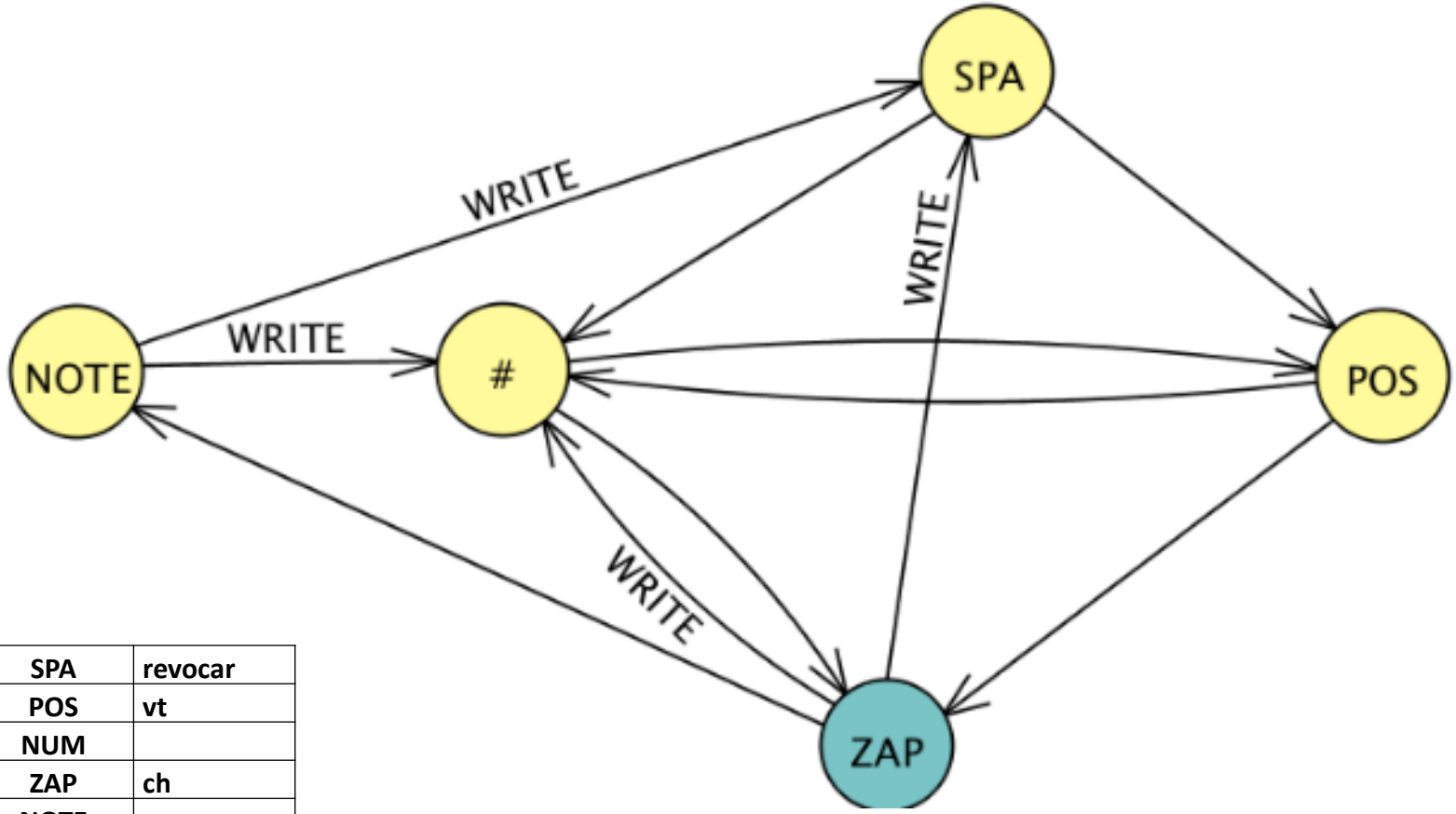
SPA	revocar
POS	vt
NUM	
ZAP	
NOTE	

revocar *vt* checa'a **revolcar** …



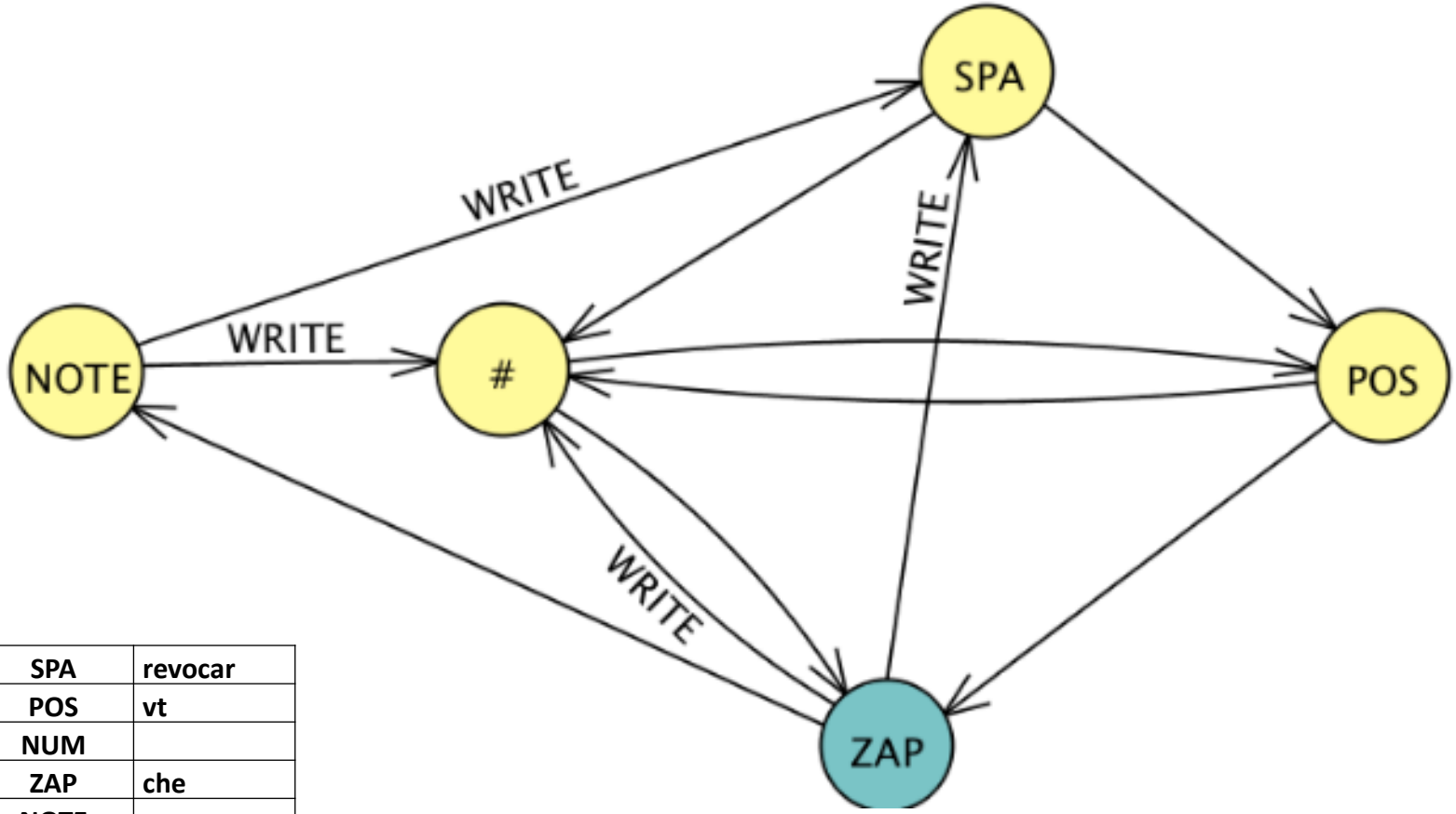
SPA	revocar
POS	vt
NUM	
ZAP	c
NOTE	

revocar *vt* checa'a **revolcar** …



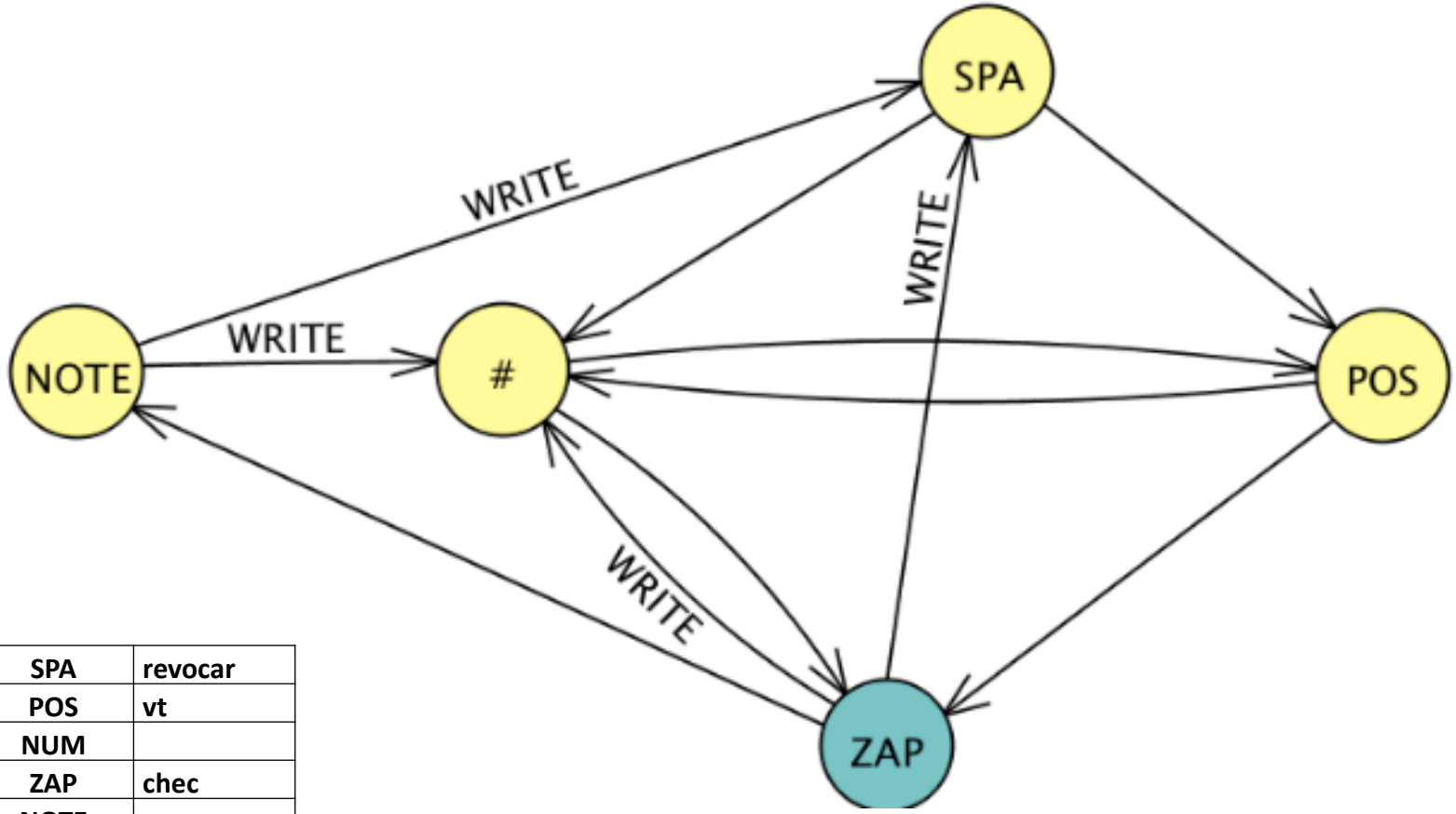
SPA	**revocar**
POS	**vt**
NUM	
ZAP	**ch**
NOTE	

revocar *vt* checa'a **revolcar** …



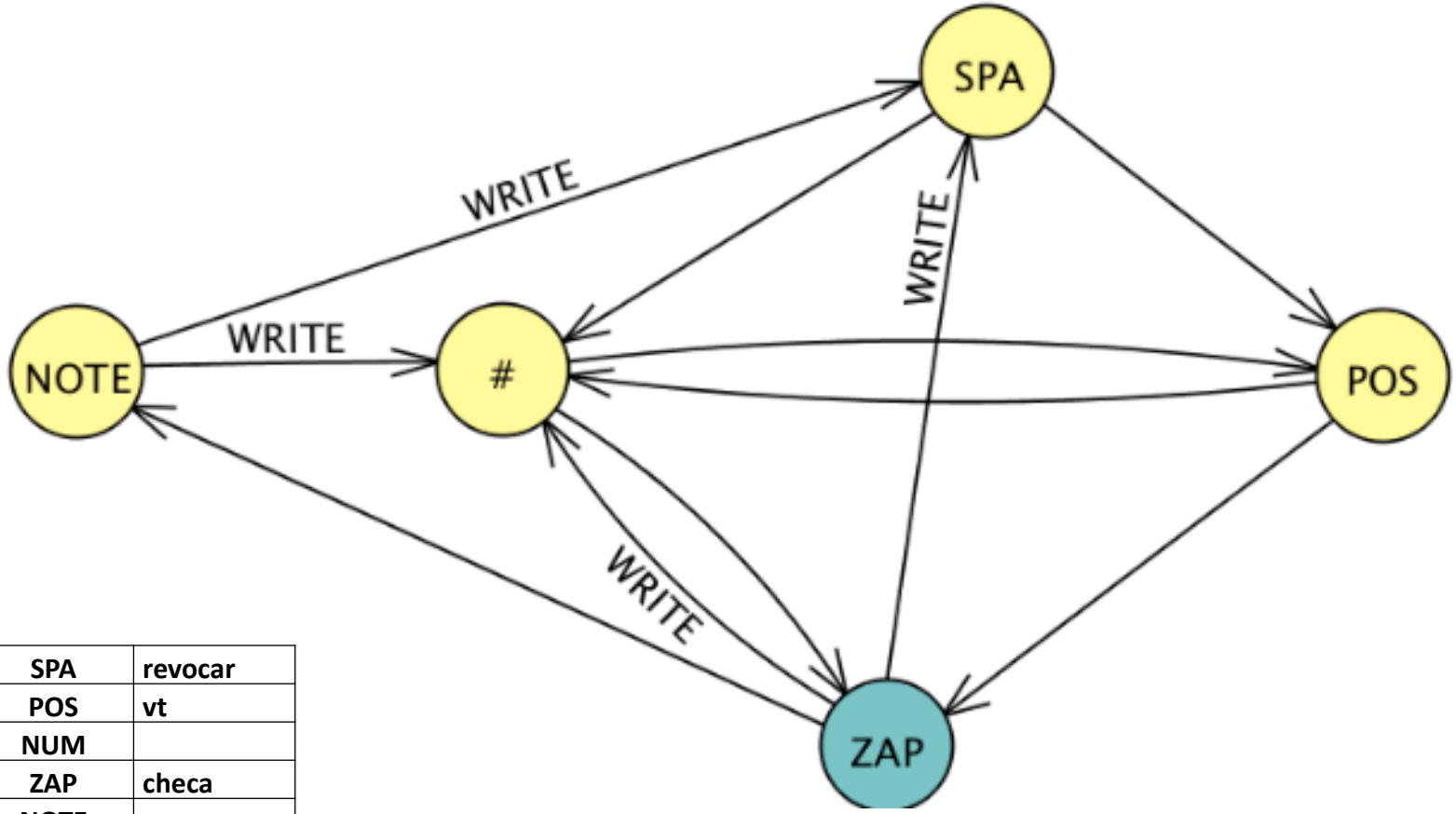
SPA	**revocar**
POS	**vt**
NUM	
ZAP	**che**
NOTE	

revocar *vt* checa'a **revolcar** …



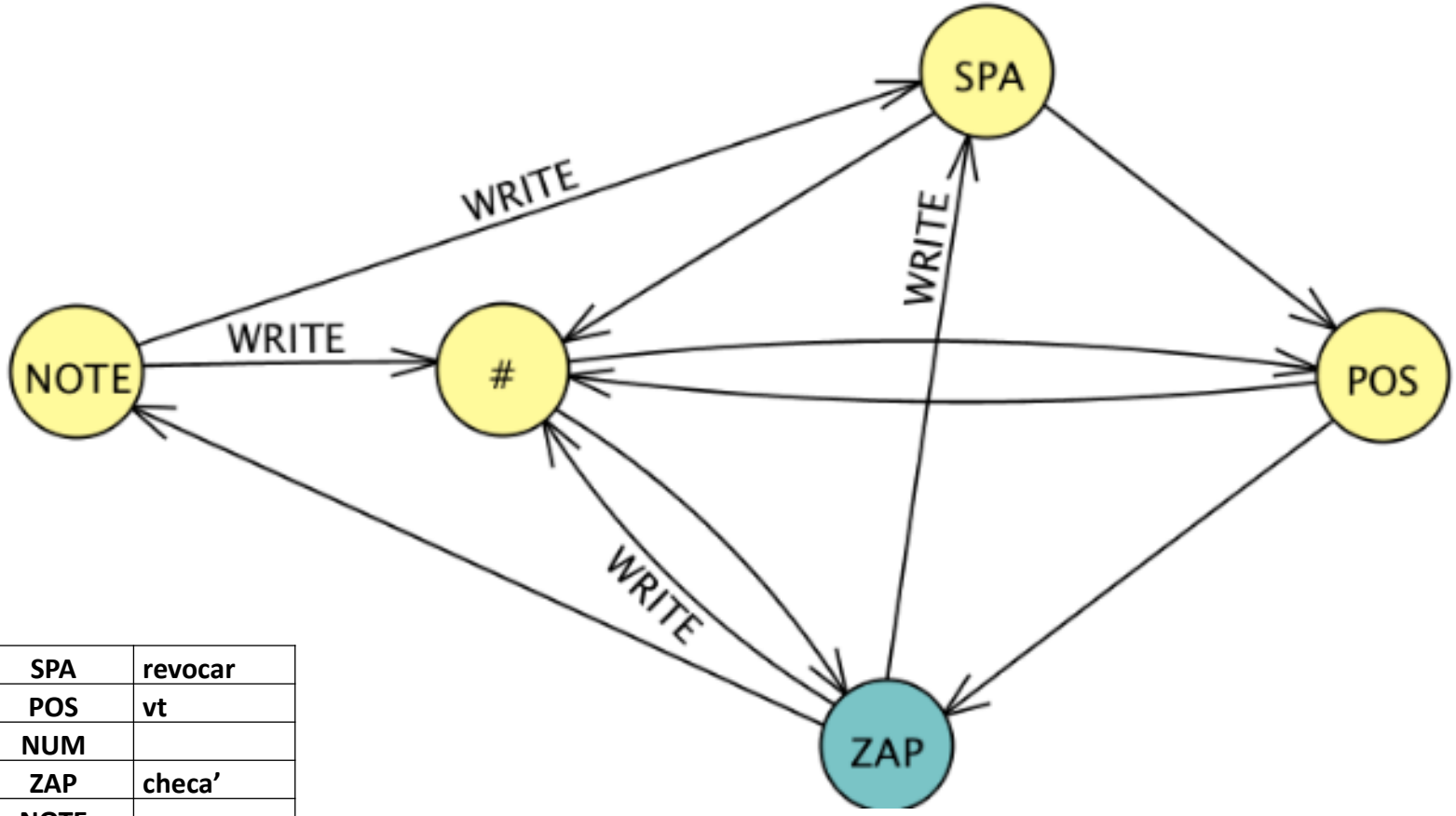
SPA	**revocar**
POS	**vt**
NUM	
ZAP	**chec**
NOTE	

revocar *vt* checa'a **revolcar** …



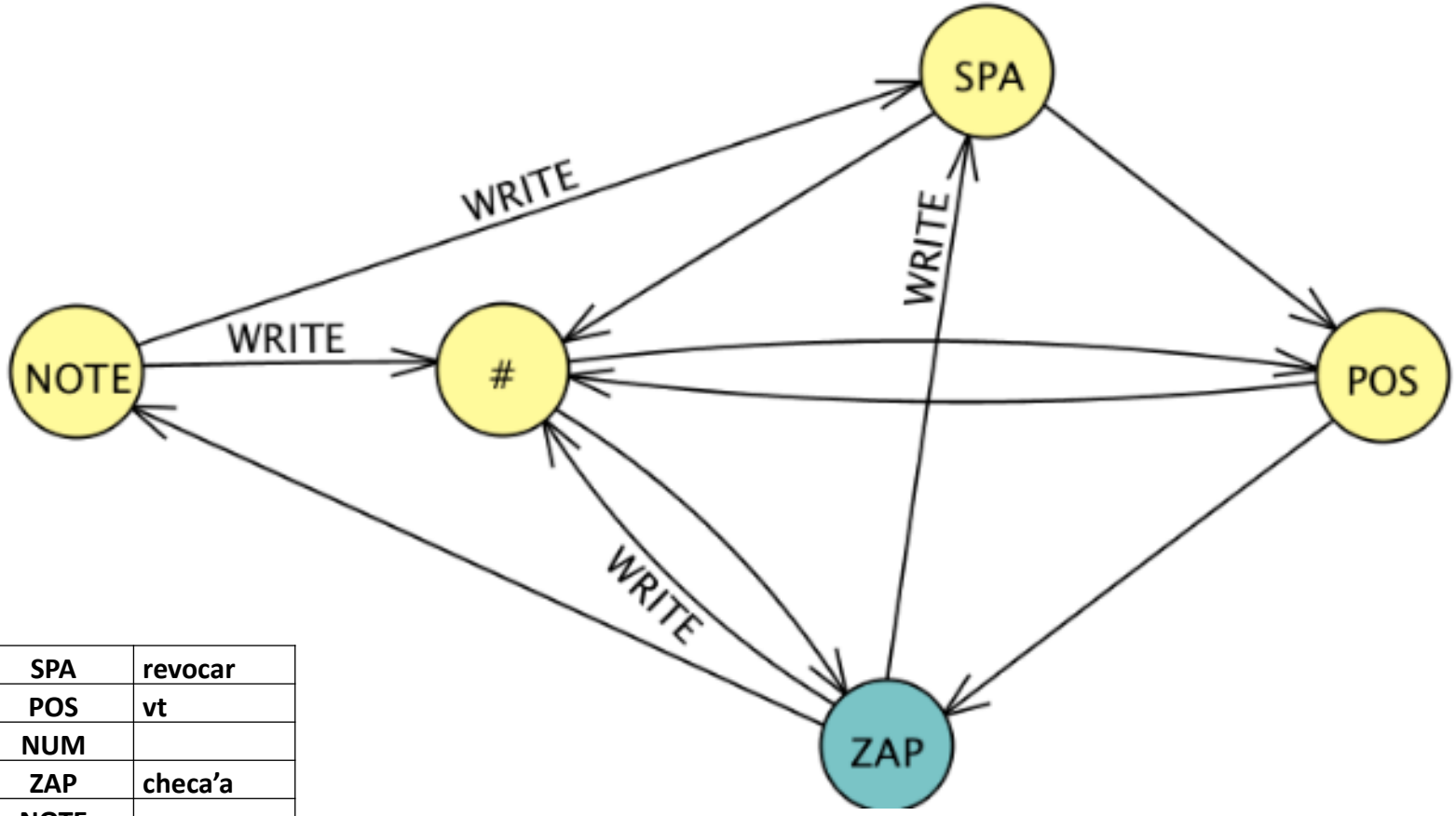
SPA	revocar
POS	vt
NUM	
ZAP	checa
NOTE	

revocar *vt* checa'a
revolcar …



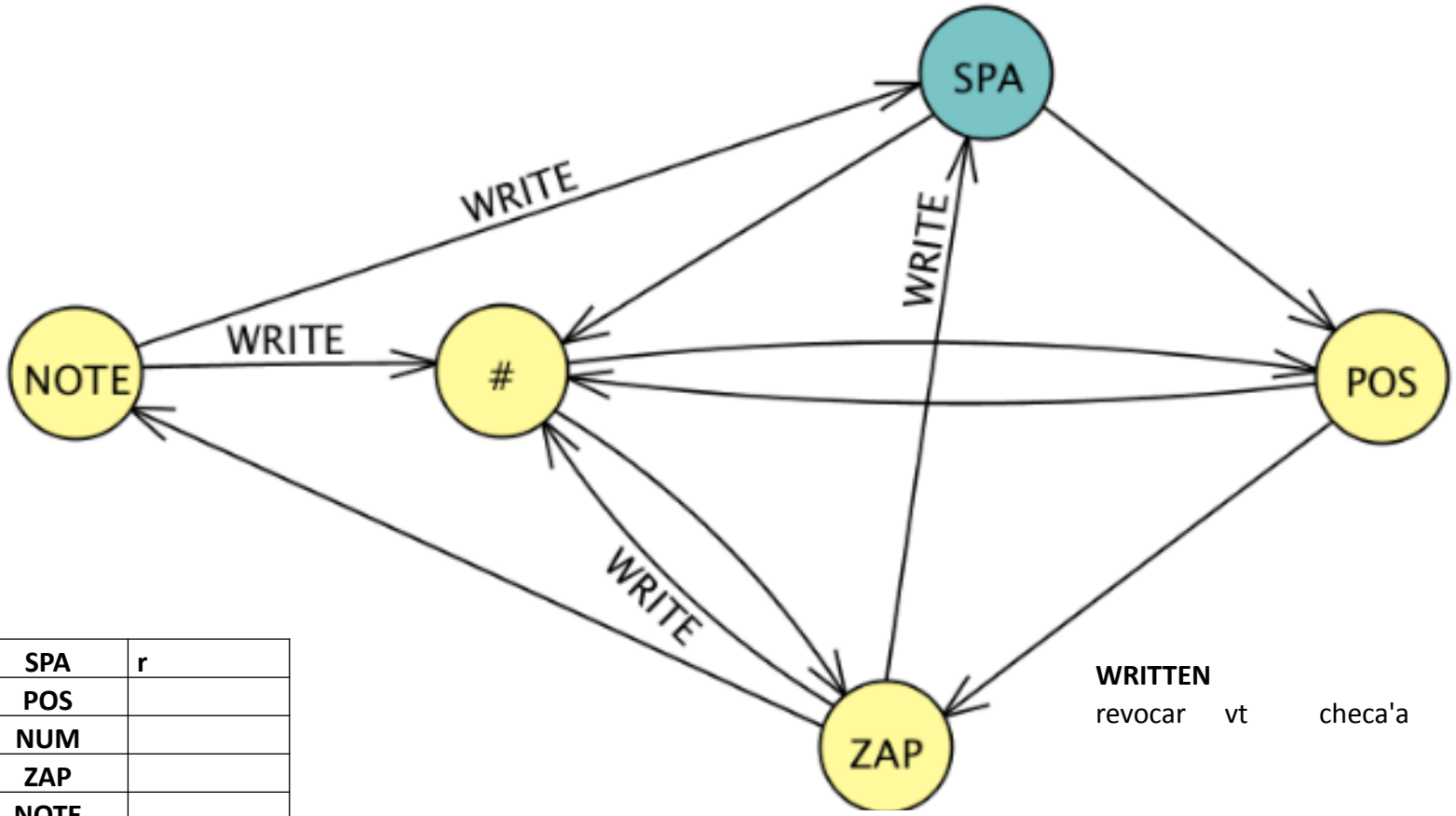
SPA	revocar
POS	vt
NUM	
ZAP	checa'
NOTE	

revocar *vt* checa'a
revolcar …



SPA	**revocar**
POS	**vt**
NUM	
ZAP	**checa'a**
NOTE	

revocar *vt* checa'a **revolcar** …

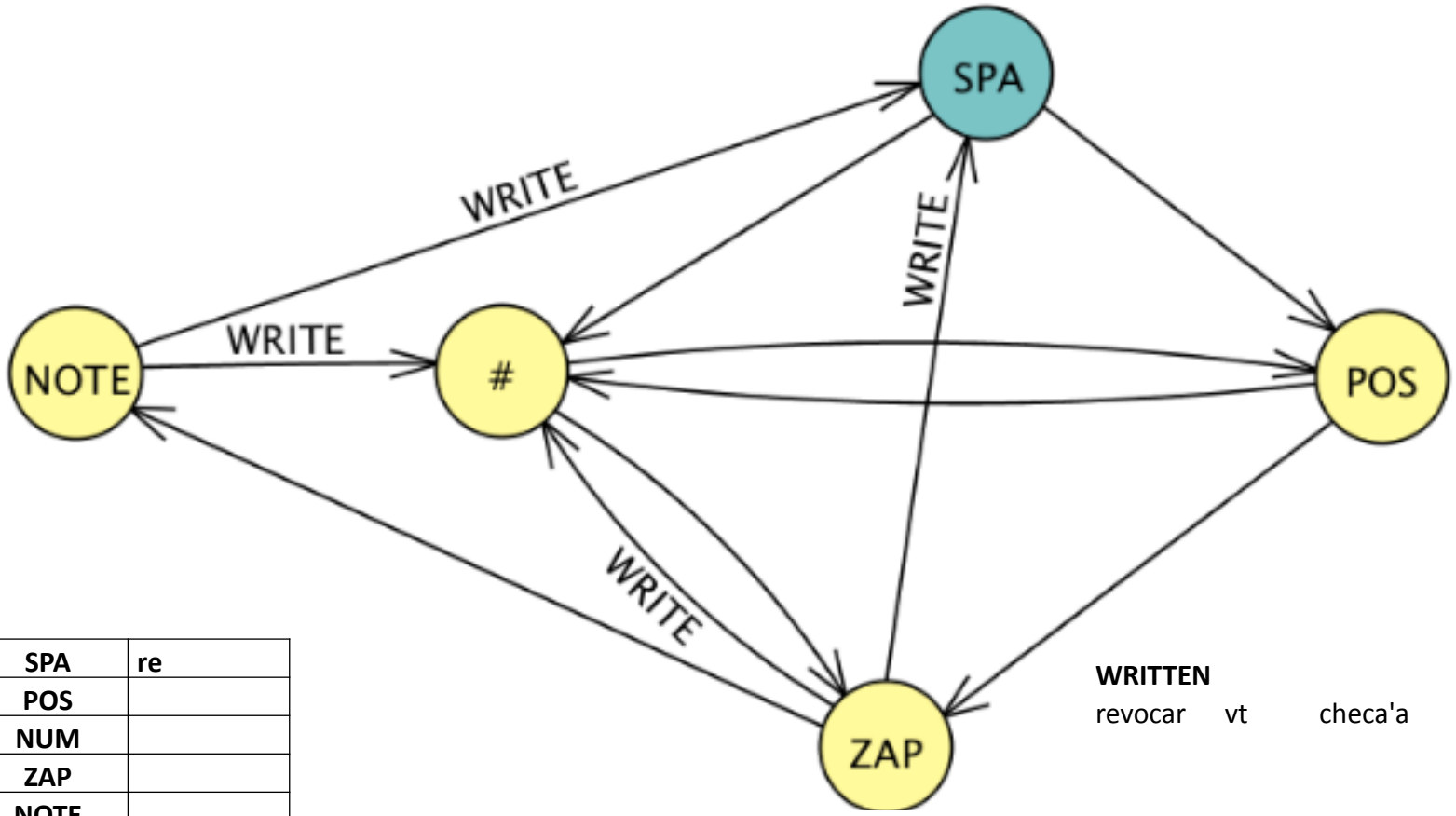


SPA	**r**
POS	
NUM	
ZAP	
NOTE	

WRITTEN

revocar vt checa'a

revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …

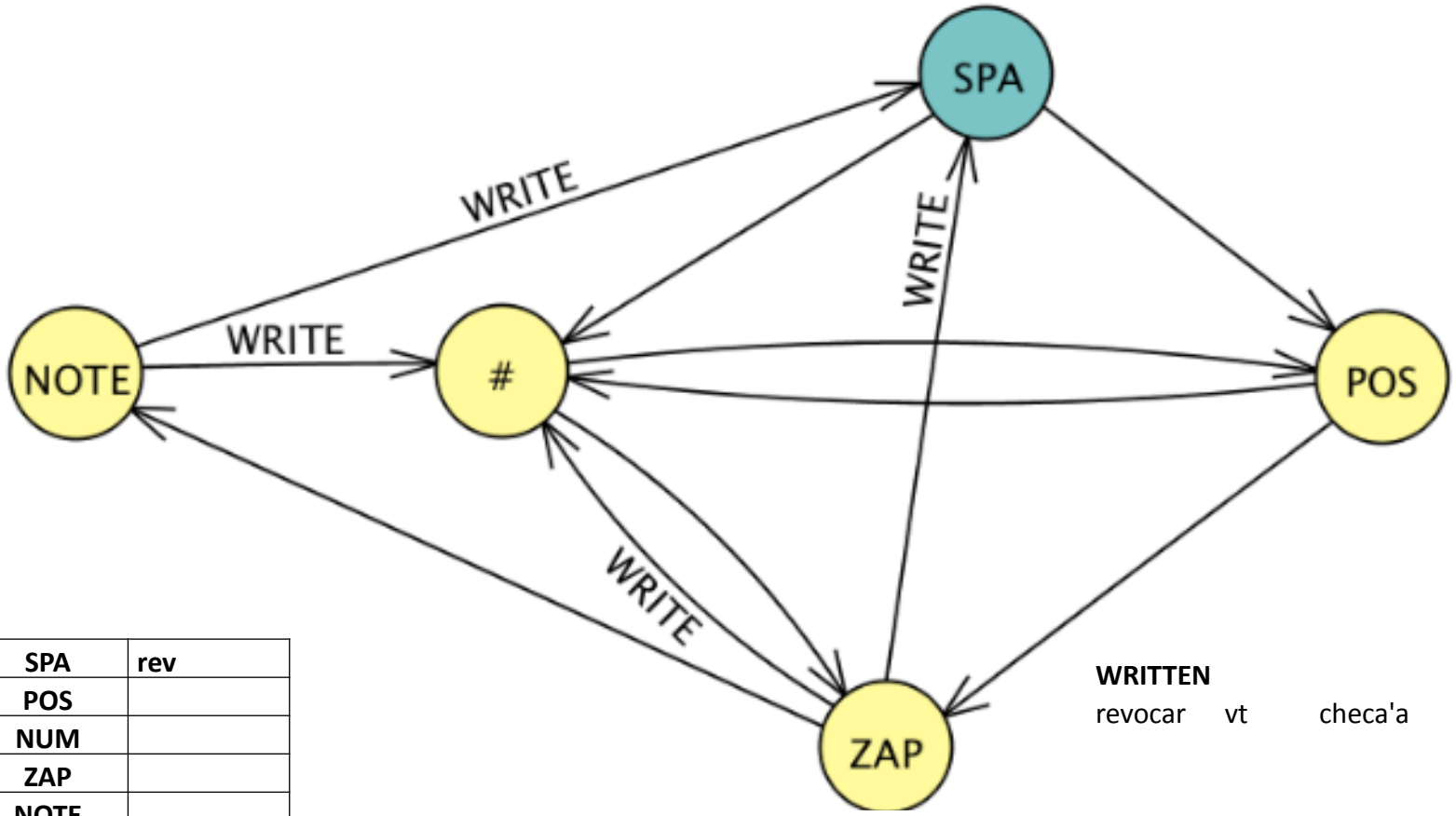


SPA	re
POS	
NUM	
ZAP	
NOTE	

WRITTEN

revocar vt checa'a

revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …

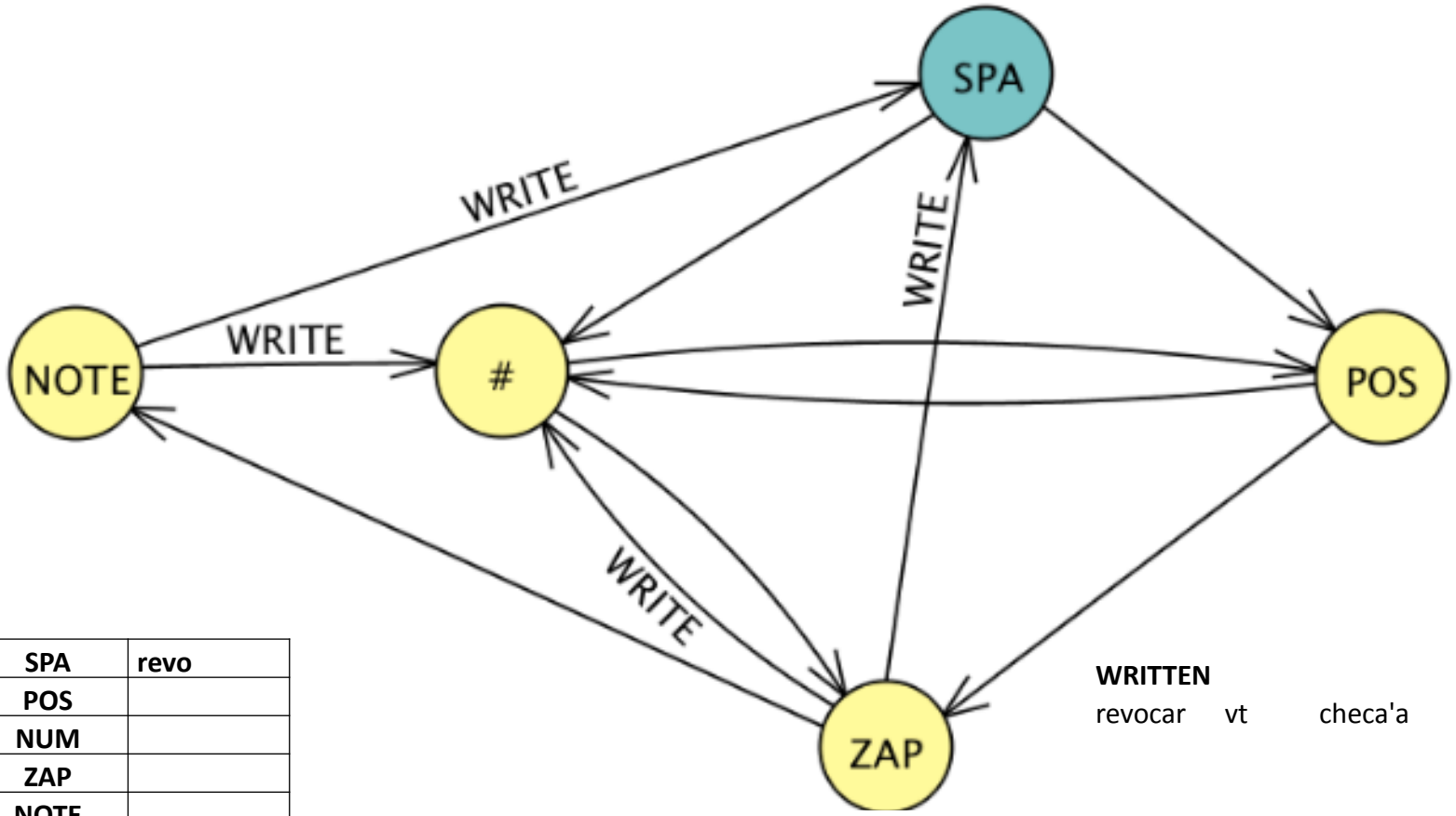


SPA	rev
POS	
NUM	
ZAP	
NOTE	

WRITTEN

revocar vt checa'a

revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …

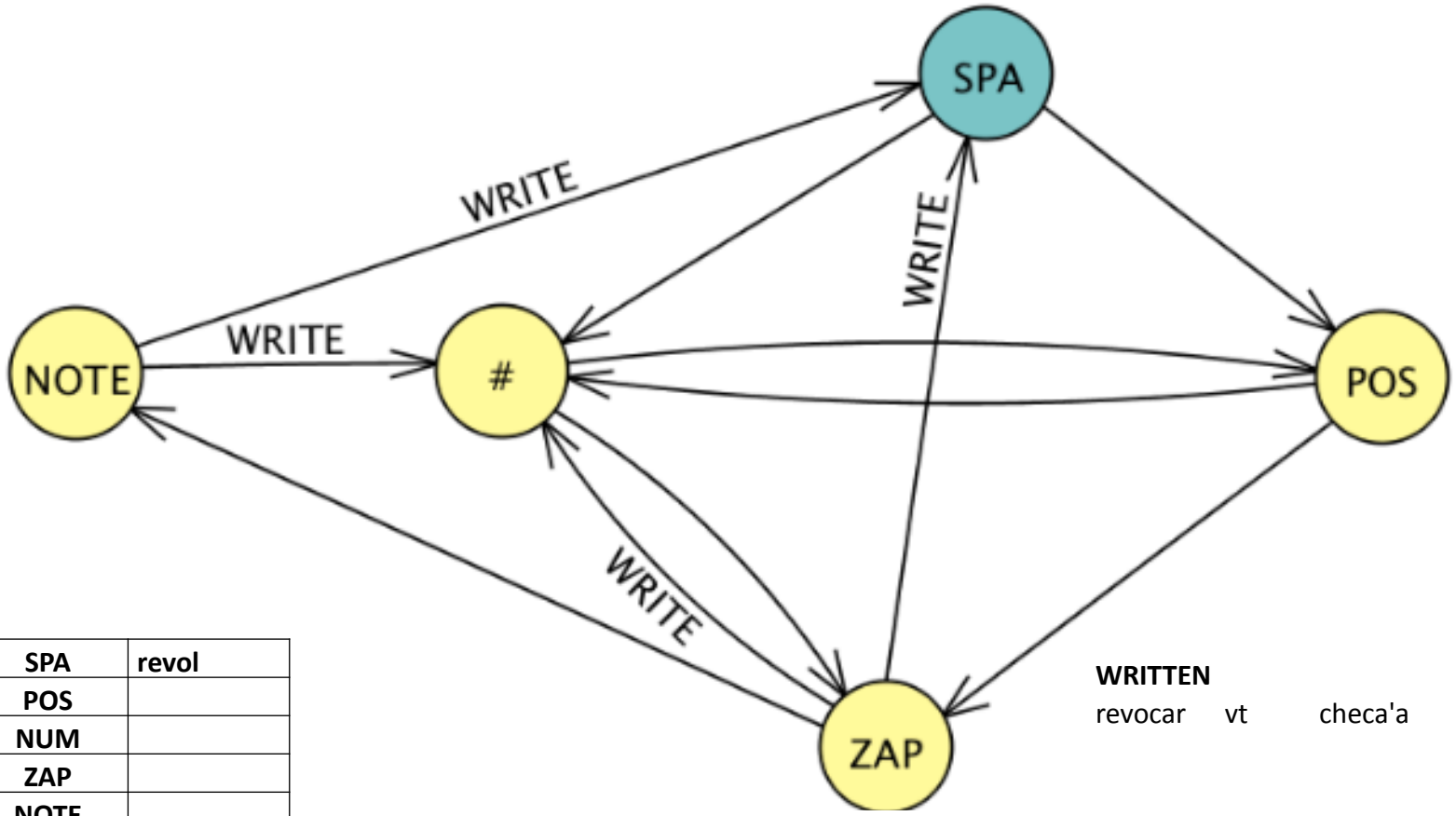


SPA	revo
POS	
NUM	
ZAP	
NOTE	

WRITTEN

revocar vt checa'a

revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …

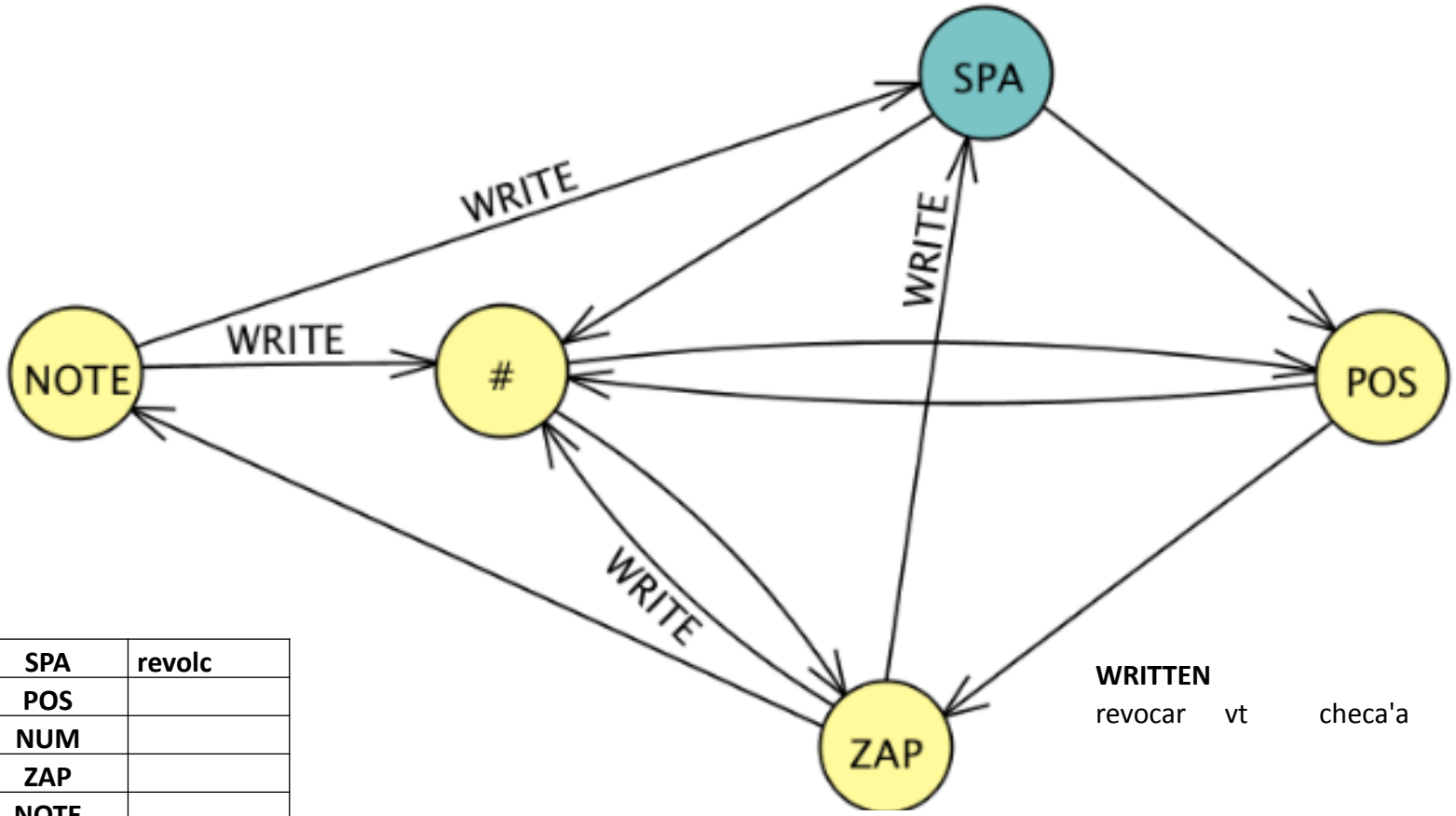


SPA	revol
POS	
NUM	
ZAP	
NOTE	

WRITTEN

revocar vt checa'a

revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …

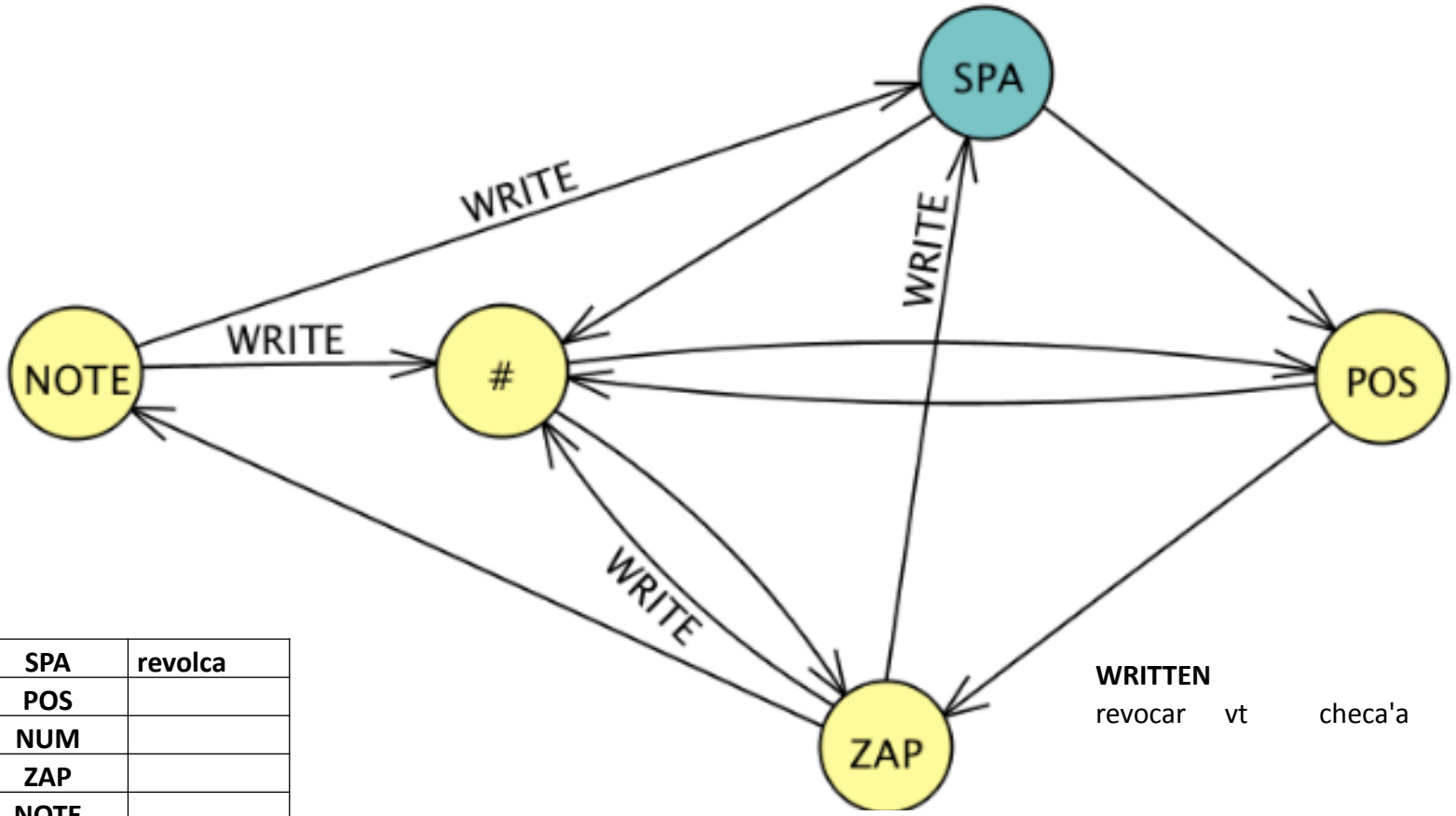


SPA	revolc
POS	
NUM	
ZAP	
NOTE	

WRITTEN

revocar vt checa'a

revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …

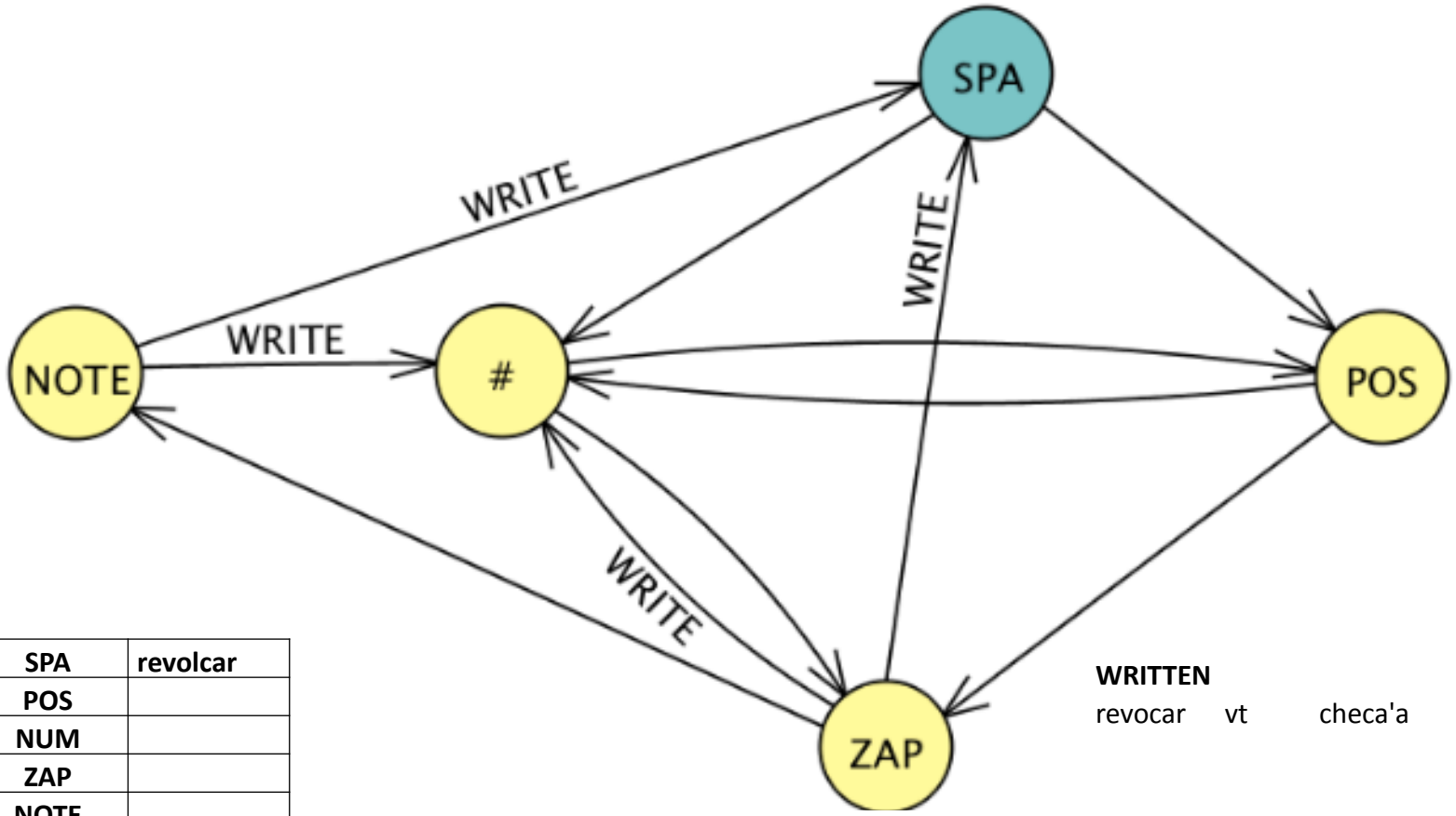


SPA	revolca
POS	
NUM	
ZAP	
NOTE	

WRITTEN

revocar vt checa'a

revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …

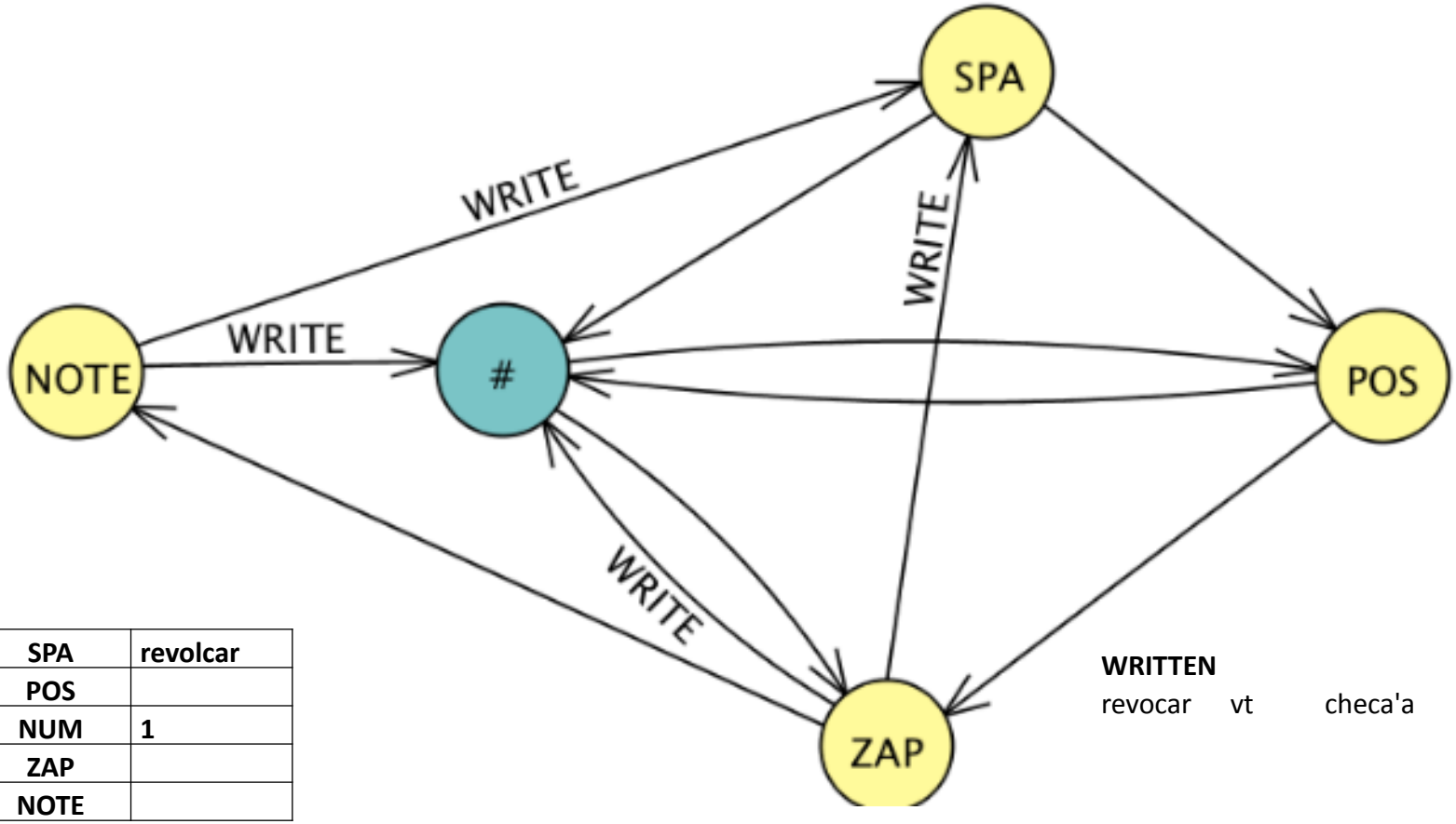


SPA	revolcar
POS	
NUM	
ZAP	
NOTE	

WRITTEN

revocar vt checa'a

revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …

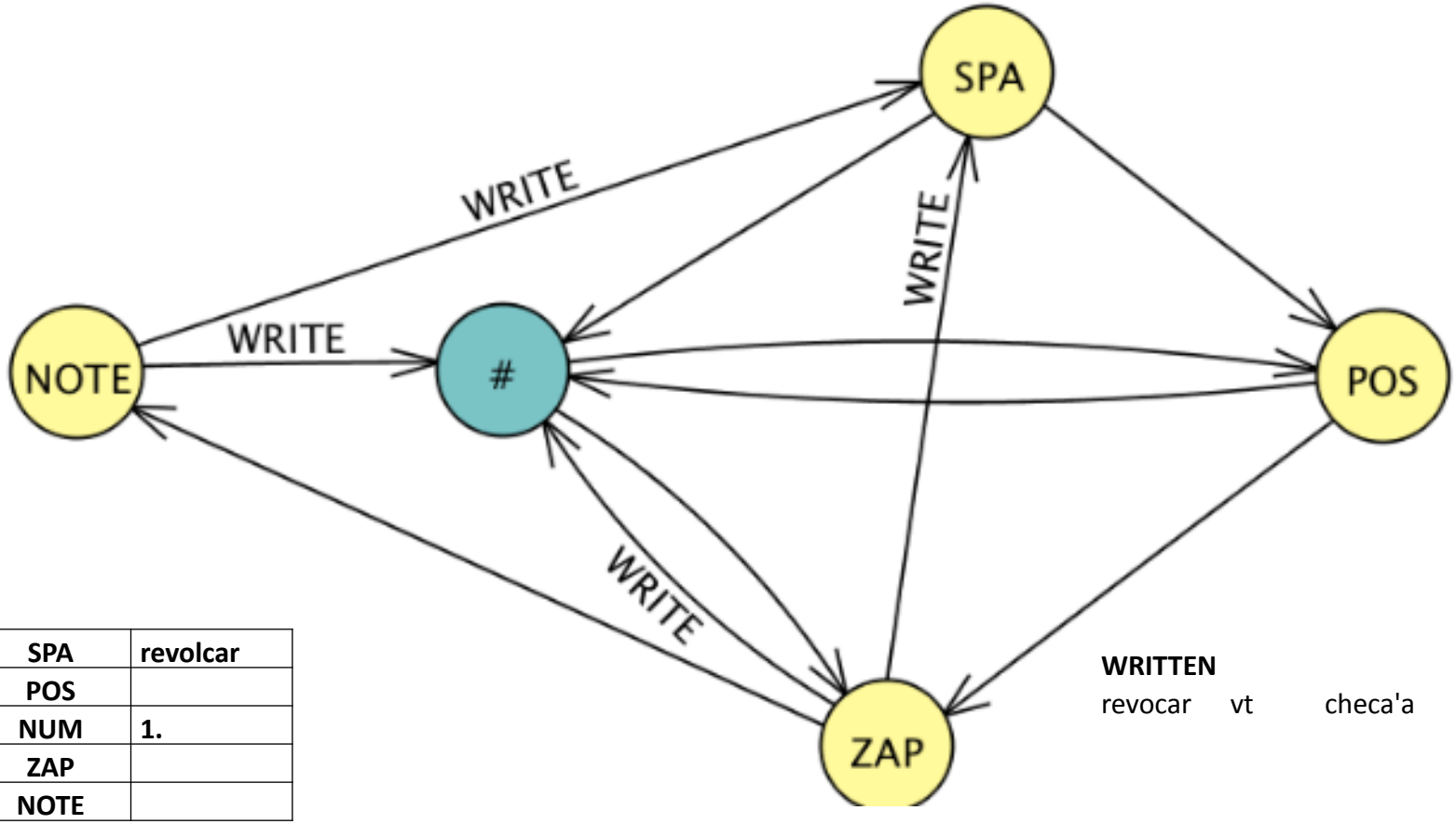


SPA	revolcar
POS	
NUM	1
ZAP	
NOTE	

WRITTEN

revocar vt checa'a

revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …

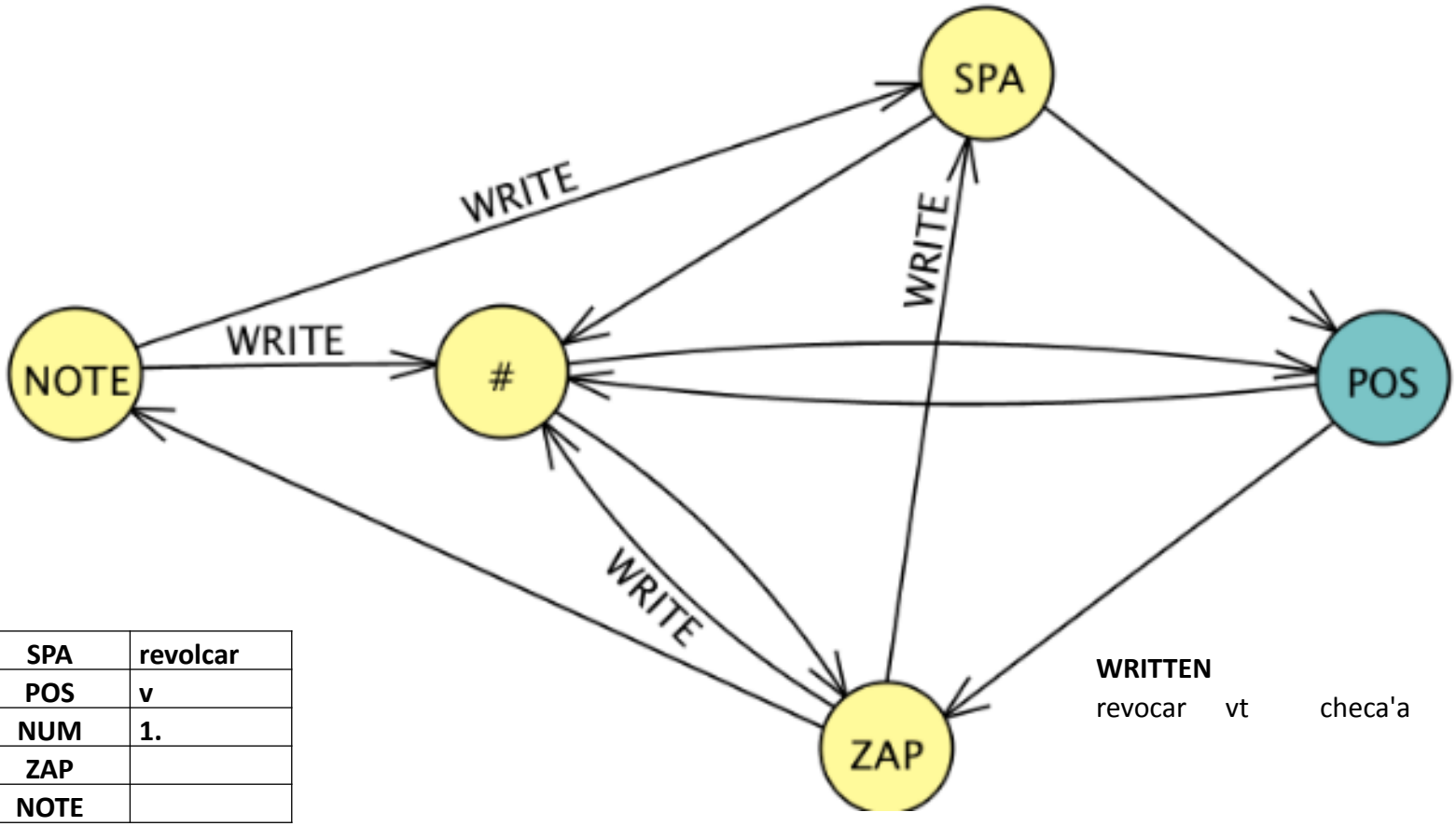


SPA	revolcar
POS	
NUM	1.
ZAP	
NOTE	

WRITTEN

revocar vt checa'a

revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …

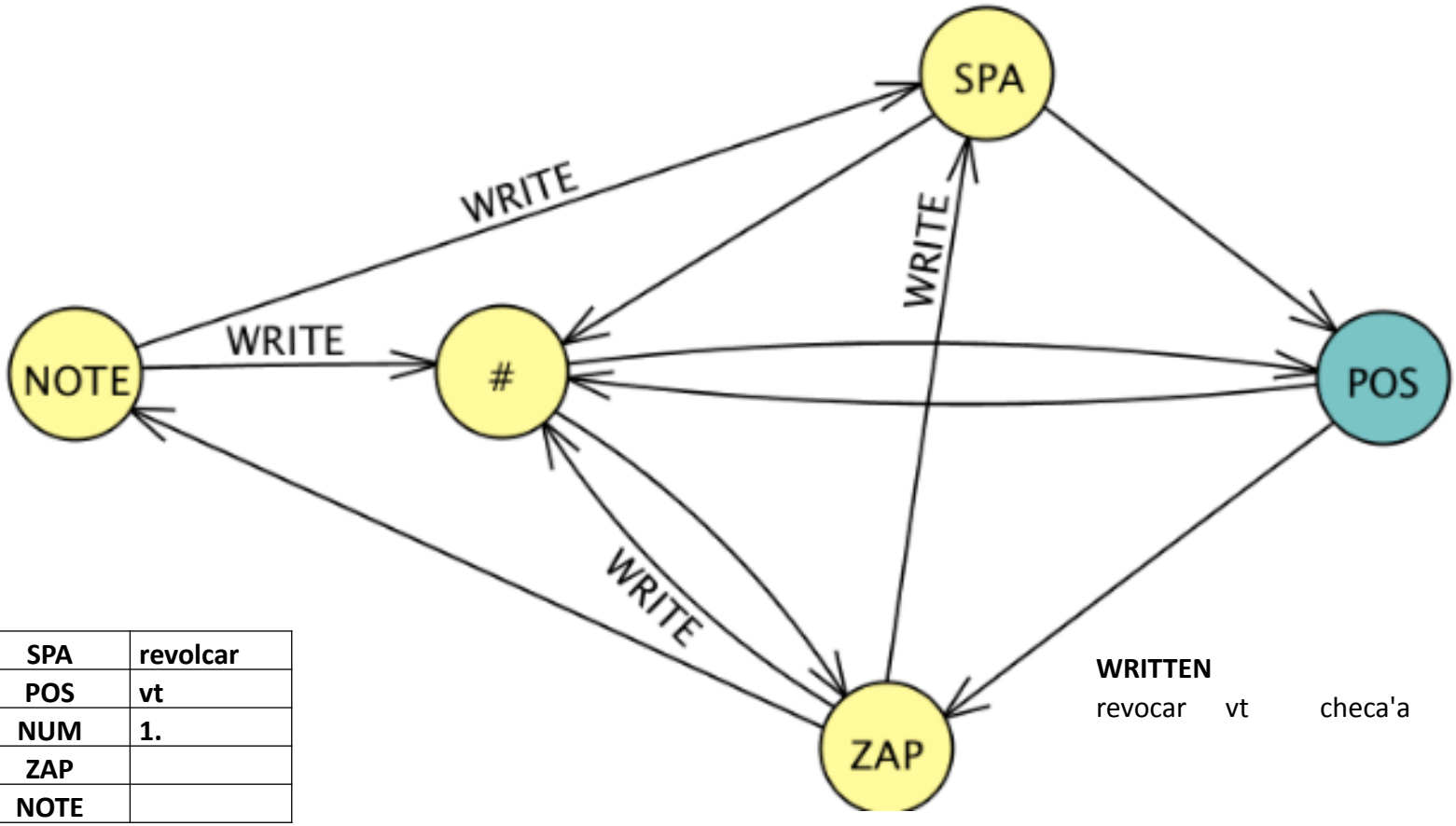


SPA	revolcar
POS	v
NUM	1.
ZAP	
NOTE	

WRITTEN

revocar vt checa'a

revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …

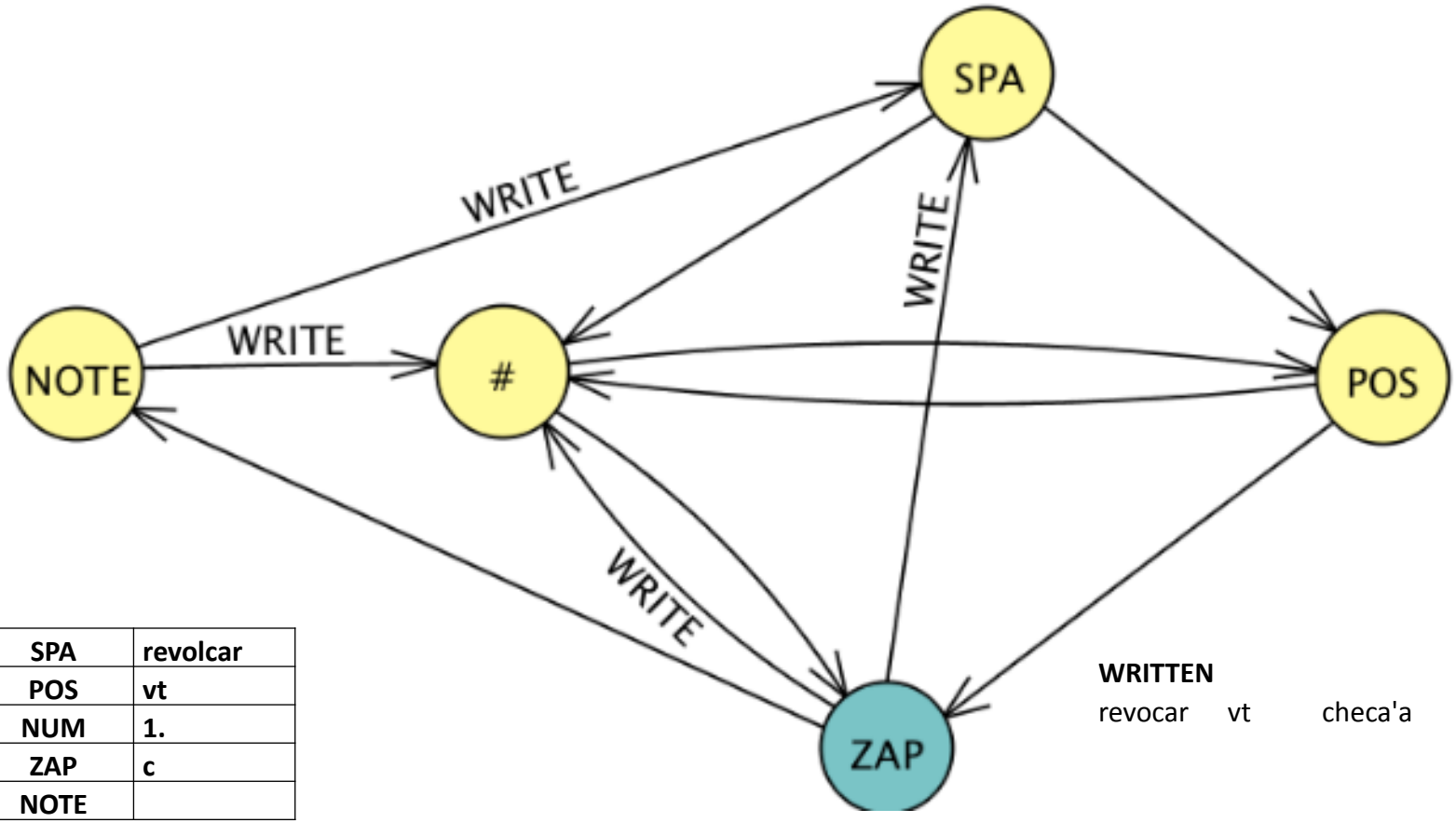


SPA	revolcar
POS	vt
NUM	1.
ZAP	
NOTE	

WRITTEN

revocar vt checa'a

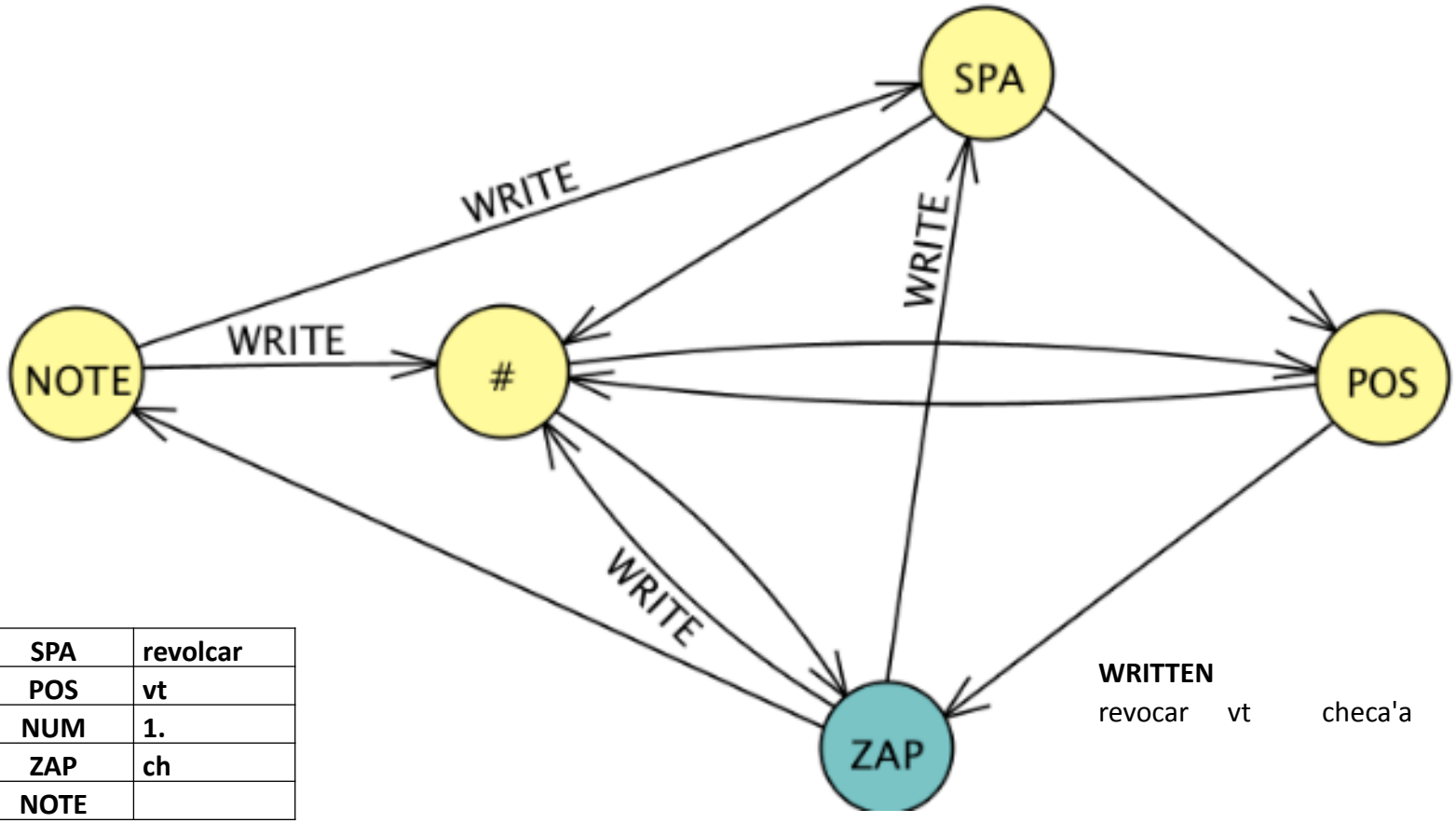
revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …



SPA	**revolcar**
POS	**vt**
NUM	**1.**
ZAP	**c**
NOTE	

WRITTEN
revocar vt checa'a

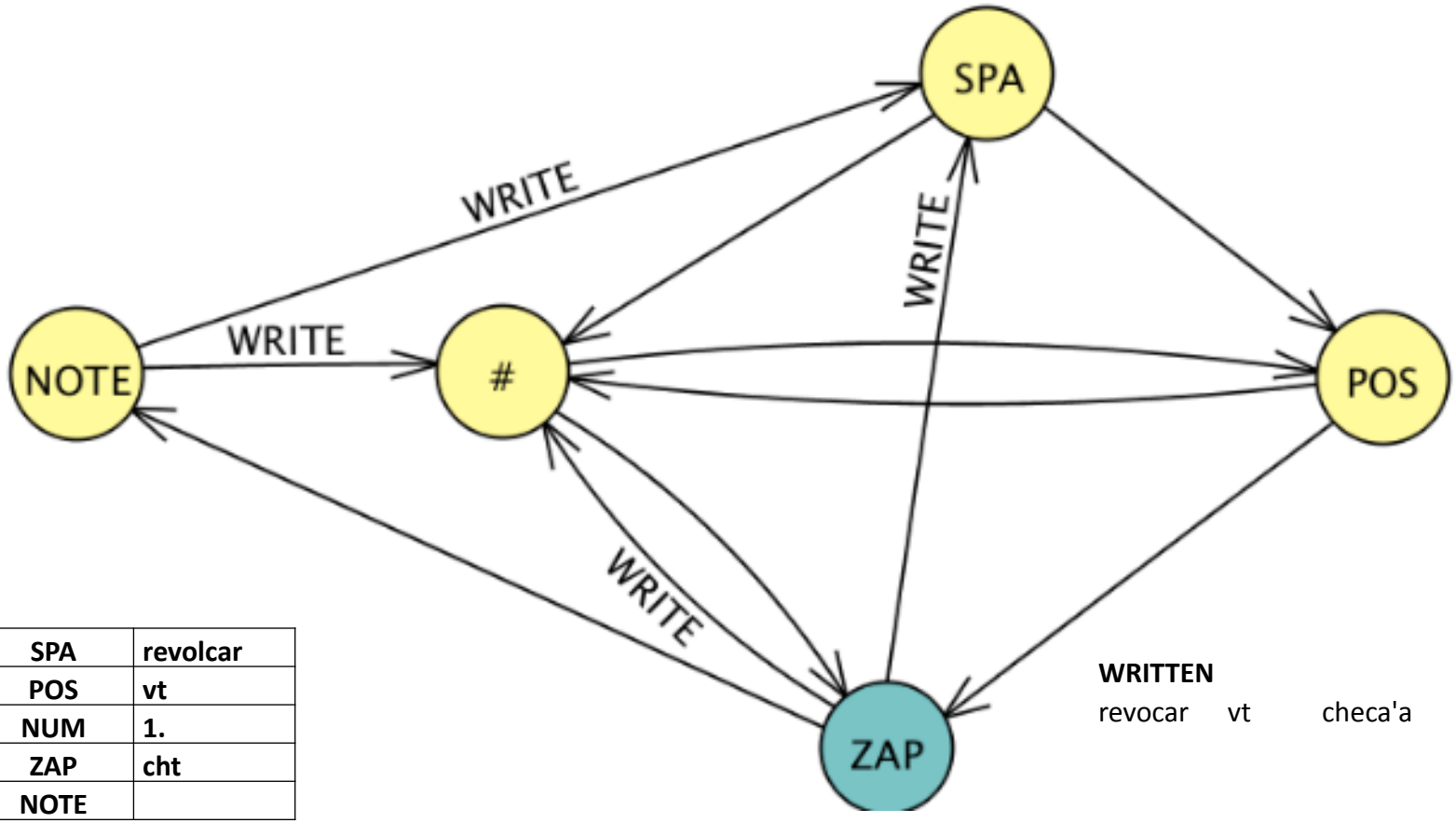
revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …



SPA	**revolcar**
POS	**vt**
NUM	**1.**
ZAP	**ch**
NOTE	

WRITTEN
revocar vt checa'a

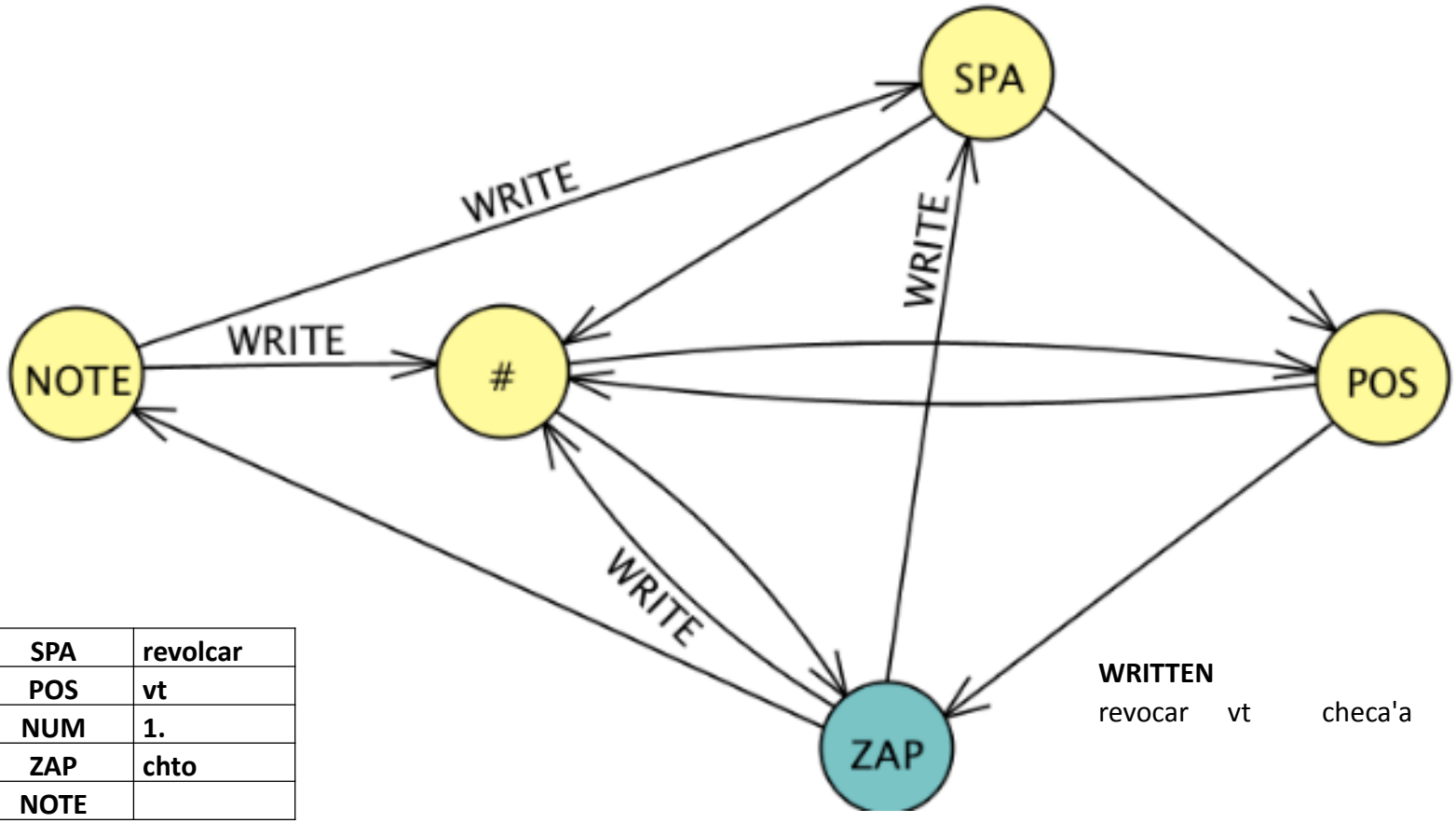
revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …



SPA	**revolcar**
POS	**vt**
NUM	**1.**
ZAP	**cht**
NOTE	

WRITTEN
revocar vt checa'a

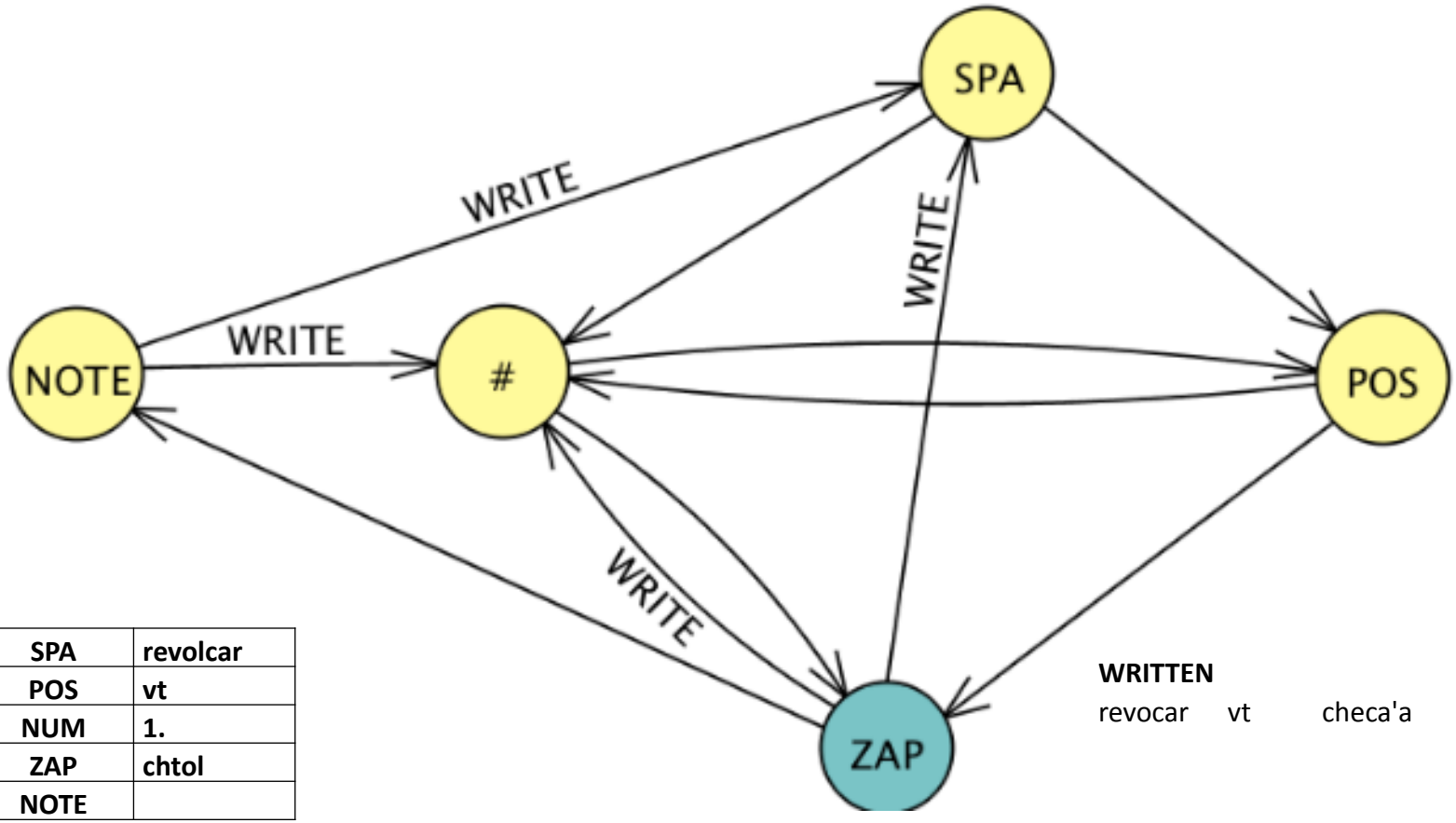
revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …



SPA	**revolcar**
POS	**vt**
NUM	**1.**
ZAP	**chto**
NOTE	

WRITTEN
revocar vt checa'a

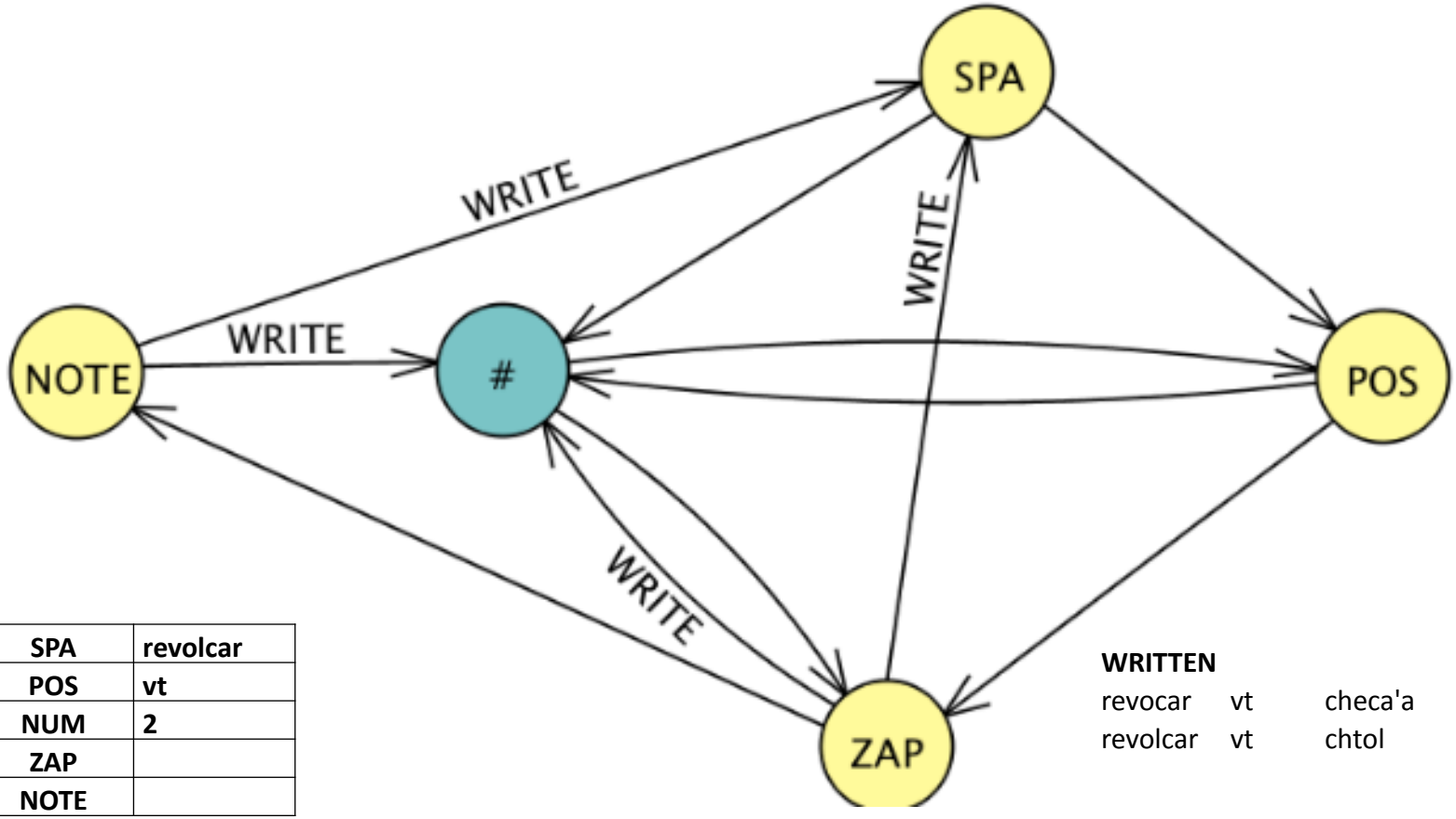
revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …



SPA	**revolcar**
POS	**vt**
NUM	**1.**
ZAP	**chtol**
NOTE	

WRITTEN
revocar vt checa'a

revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …

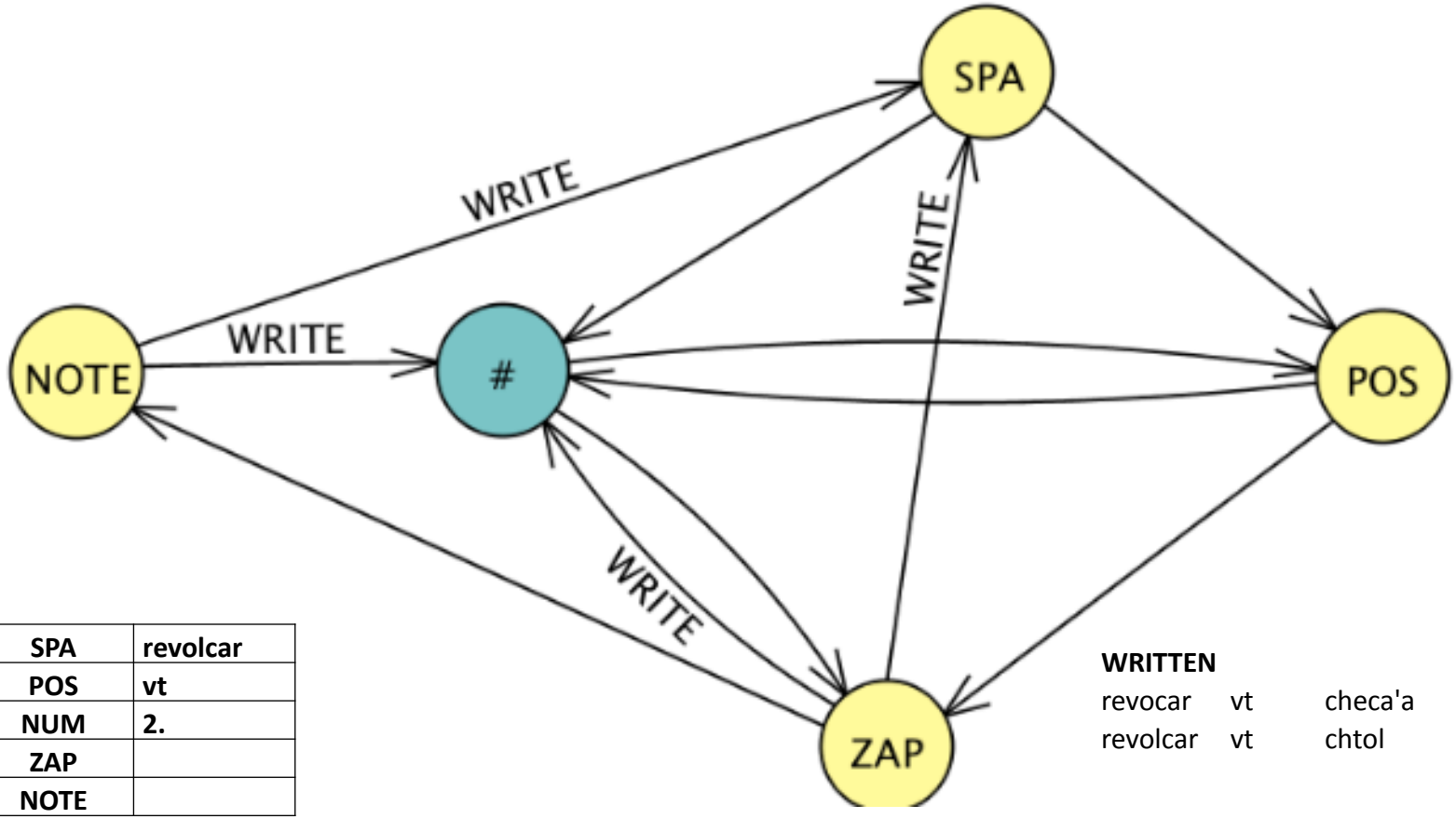


SPA	**revolcar**
POS	**vt**
NUM	**2**
ZAP	
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol

revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …

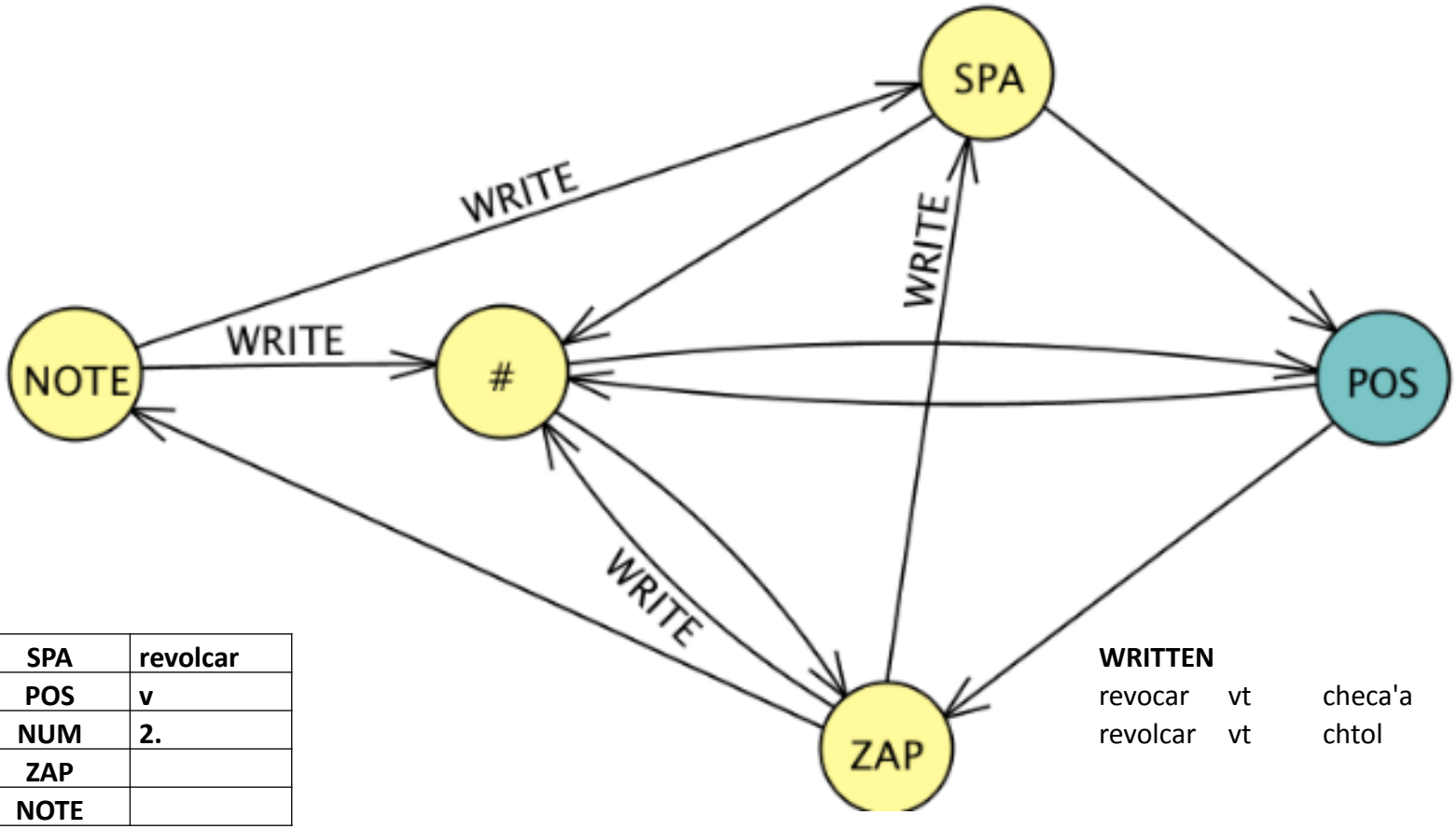


SPA	**revolcar**
POS	**vt**
NUM	**2.**
ZAP	
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol

revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …

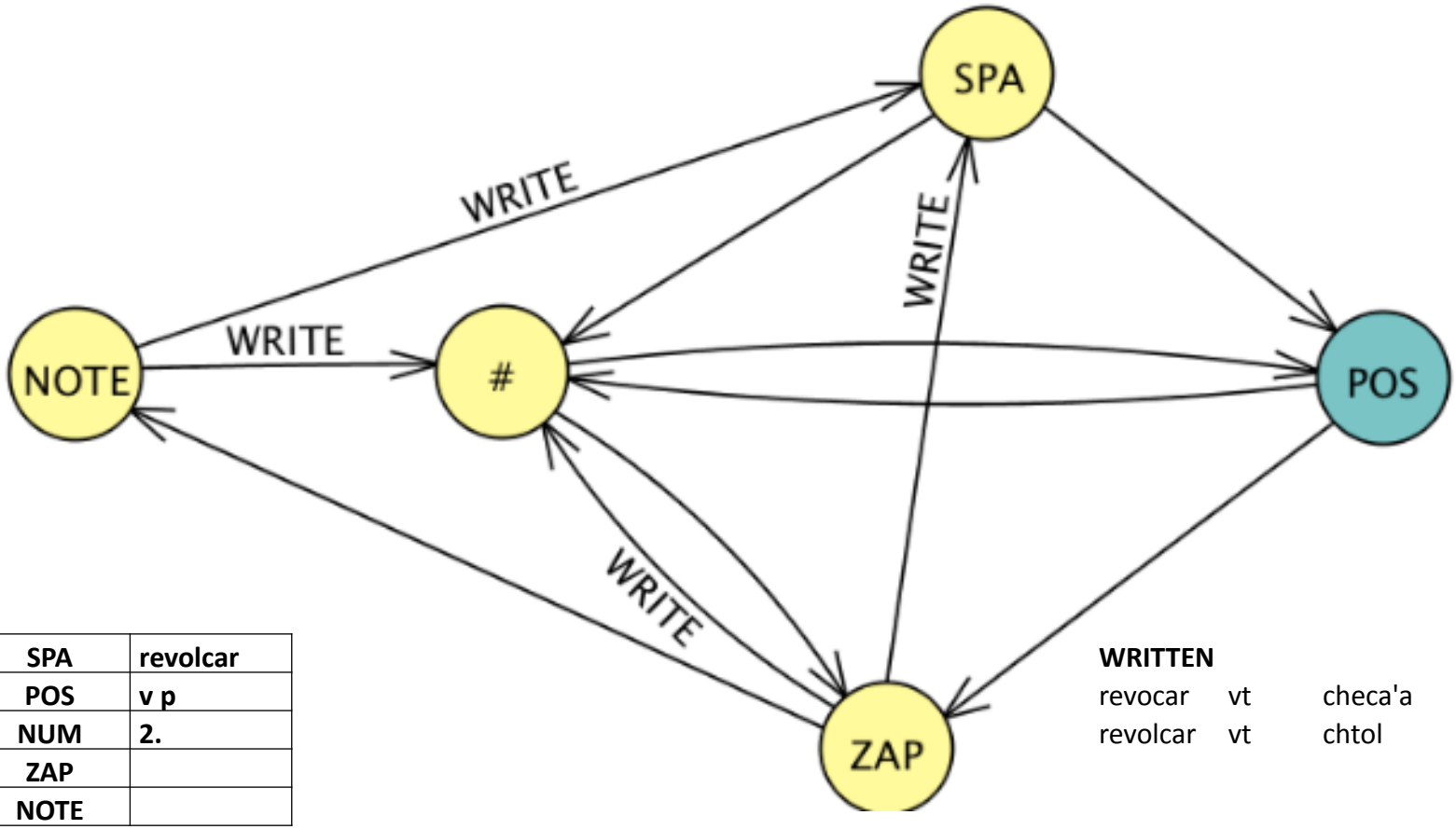


SPA	**revolcar**
POS	**v**
NUM	**2.**
ZAP	
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol

revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …

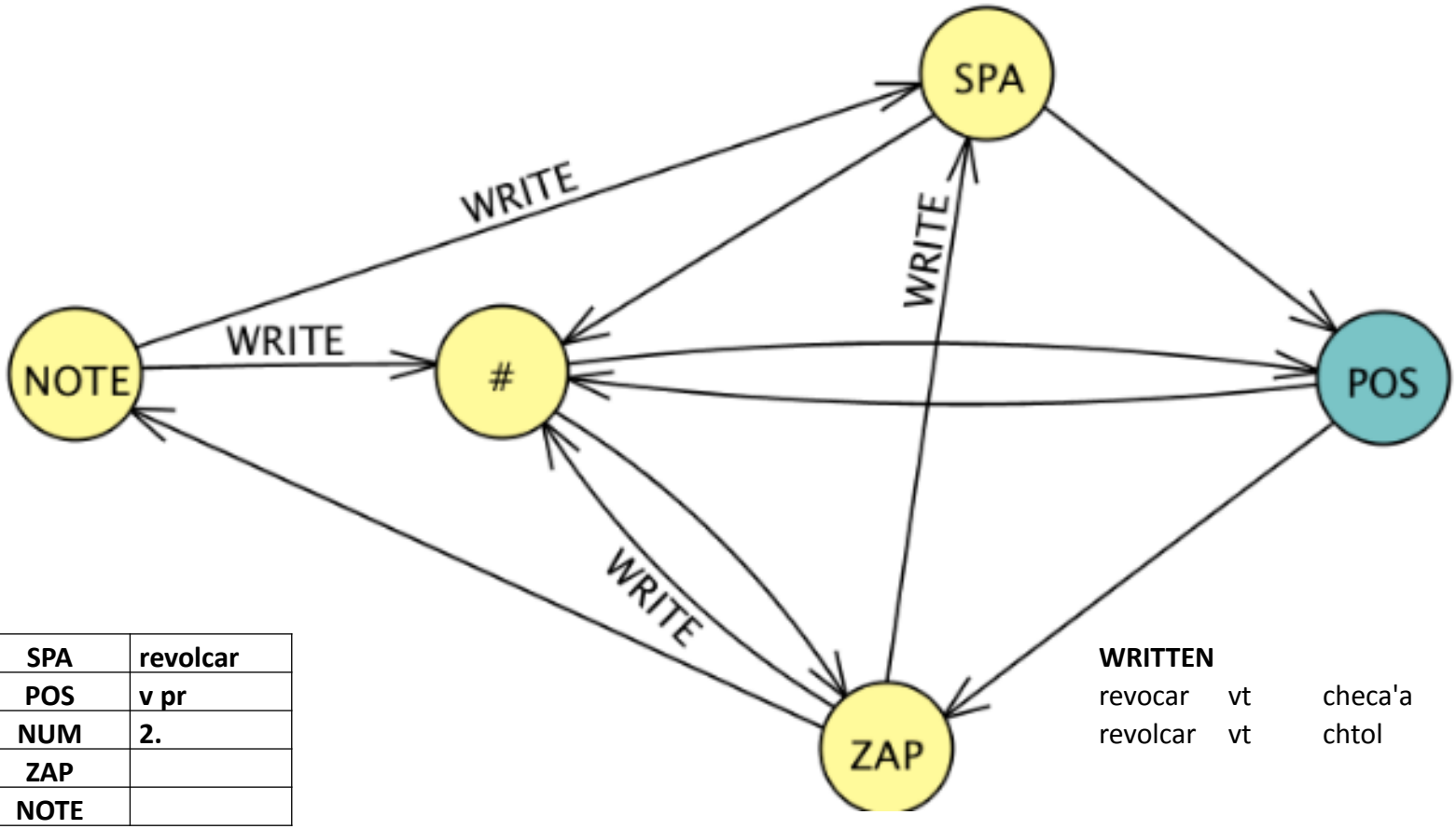


SPA	**revolcar**
POS	**v p**
NUM	**2.**
ZAP	
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol

revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …

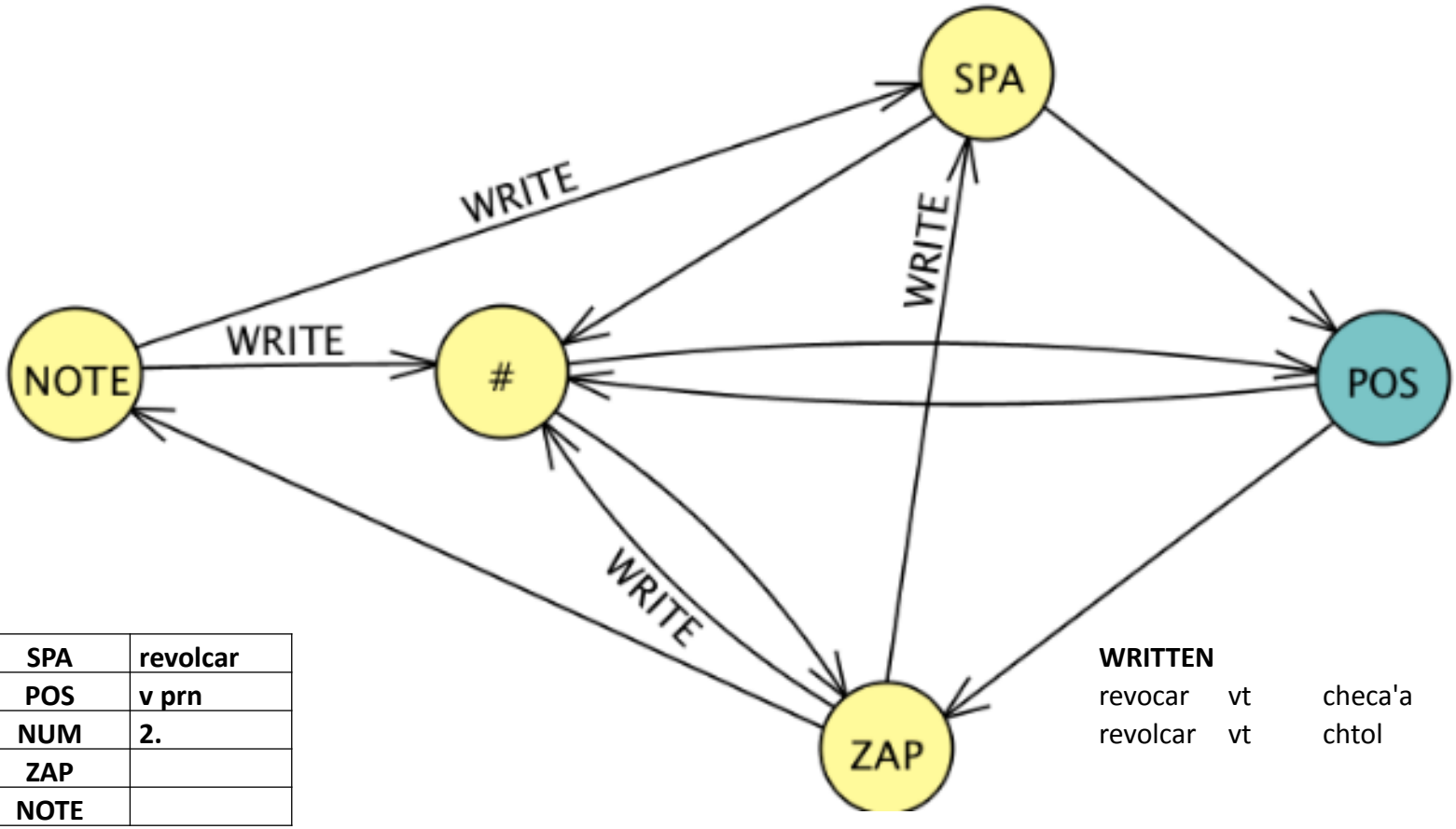


SPA	**revolcar**
POS	**v pr**
NUM	**2.**
ZAP	
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol

revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …

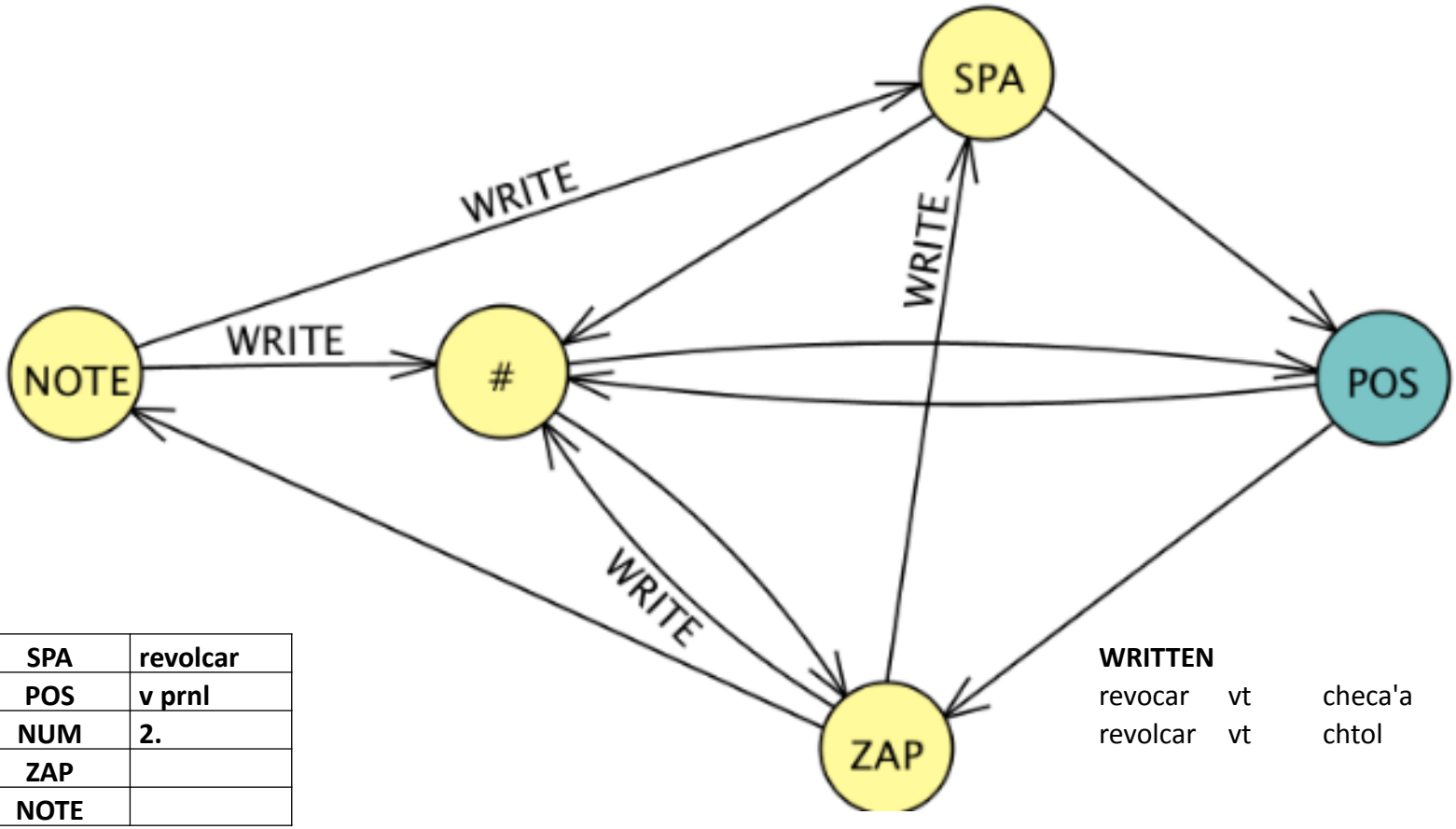


SPA	**revolcar**
POS	**v prn**
NUM	**2.**
ZAP	
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol

revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …

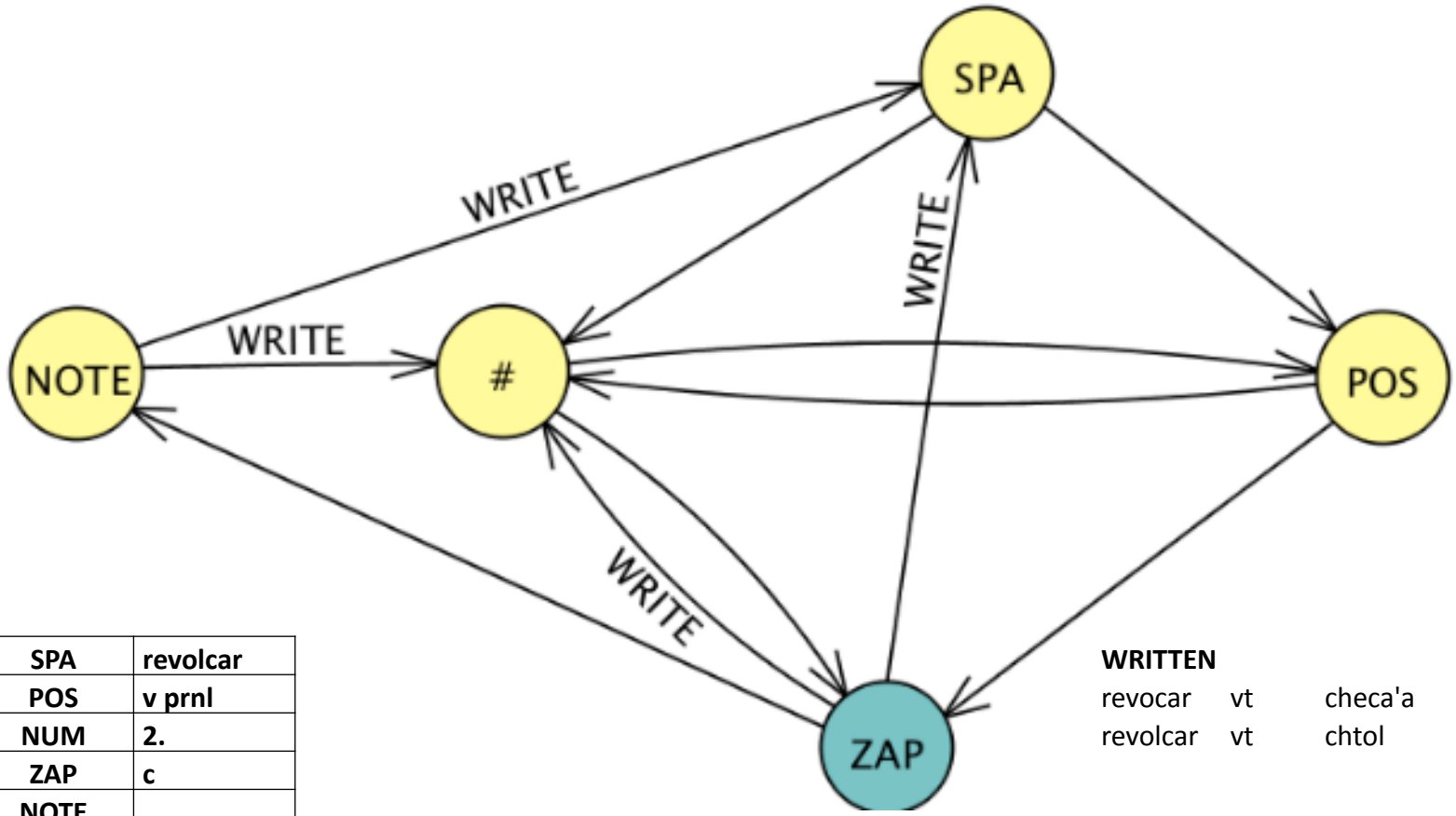


SPA	**revolcar**
POS	**v prnl**
NUM	**2.**
ZAP	
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol

revolcar 1. *vt* chtol 2. *v prnl* chata' **revolver** …

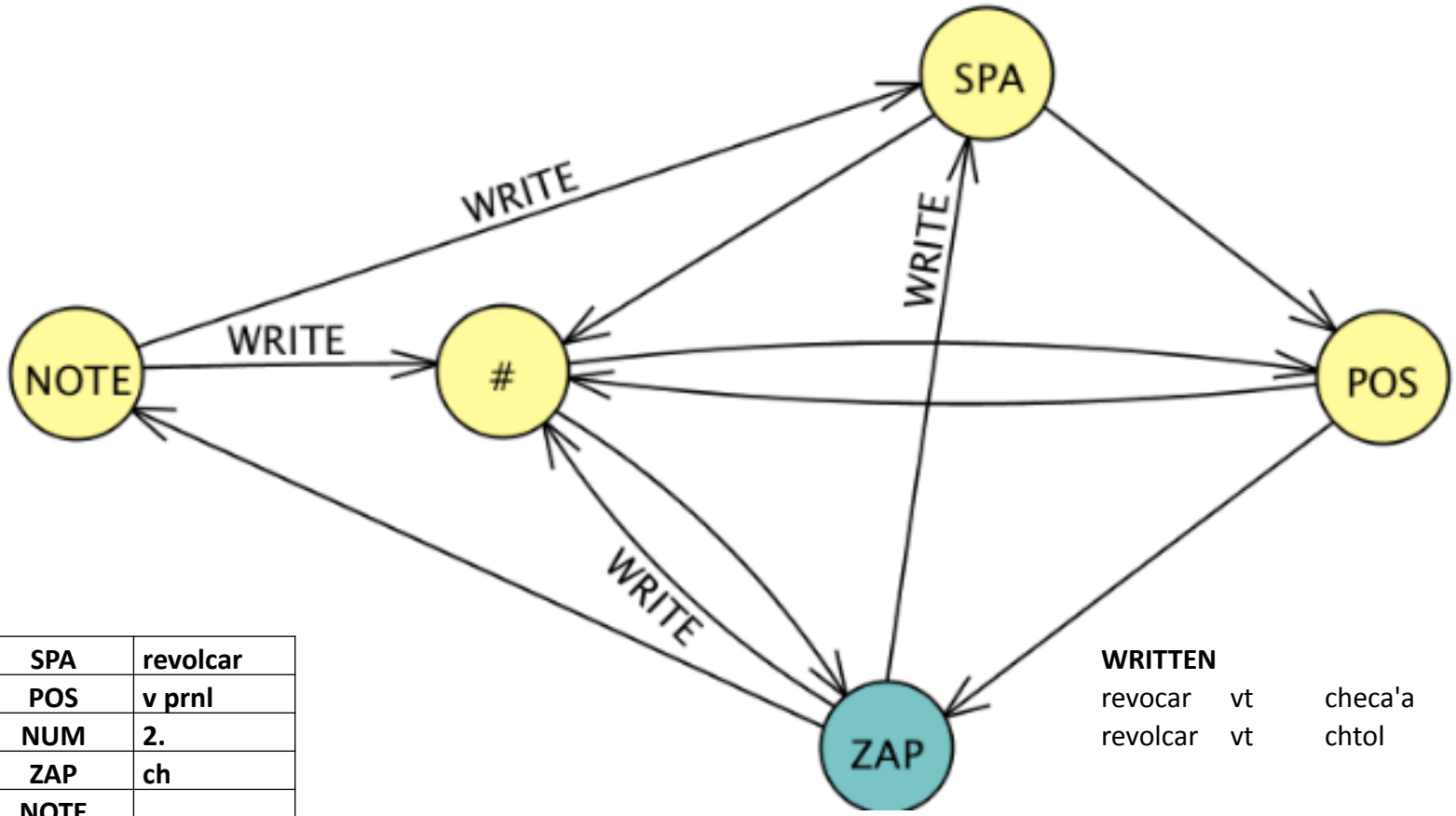


SPA	**revolcar**
POS	**v prnl**
NUM	**2.**
ZAP	**c**
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol

revolcar 1. *vt* chtol 2. *v prnl* c<u>h</u>ata' **revolver** …

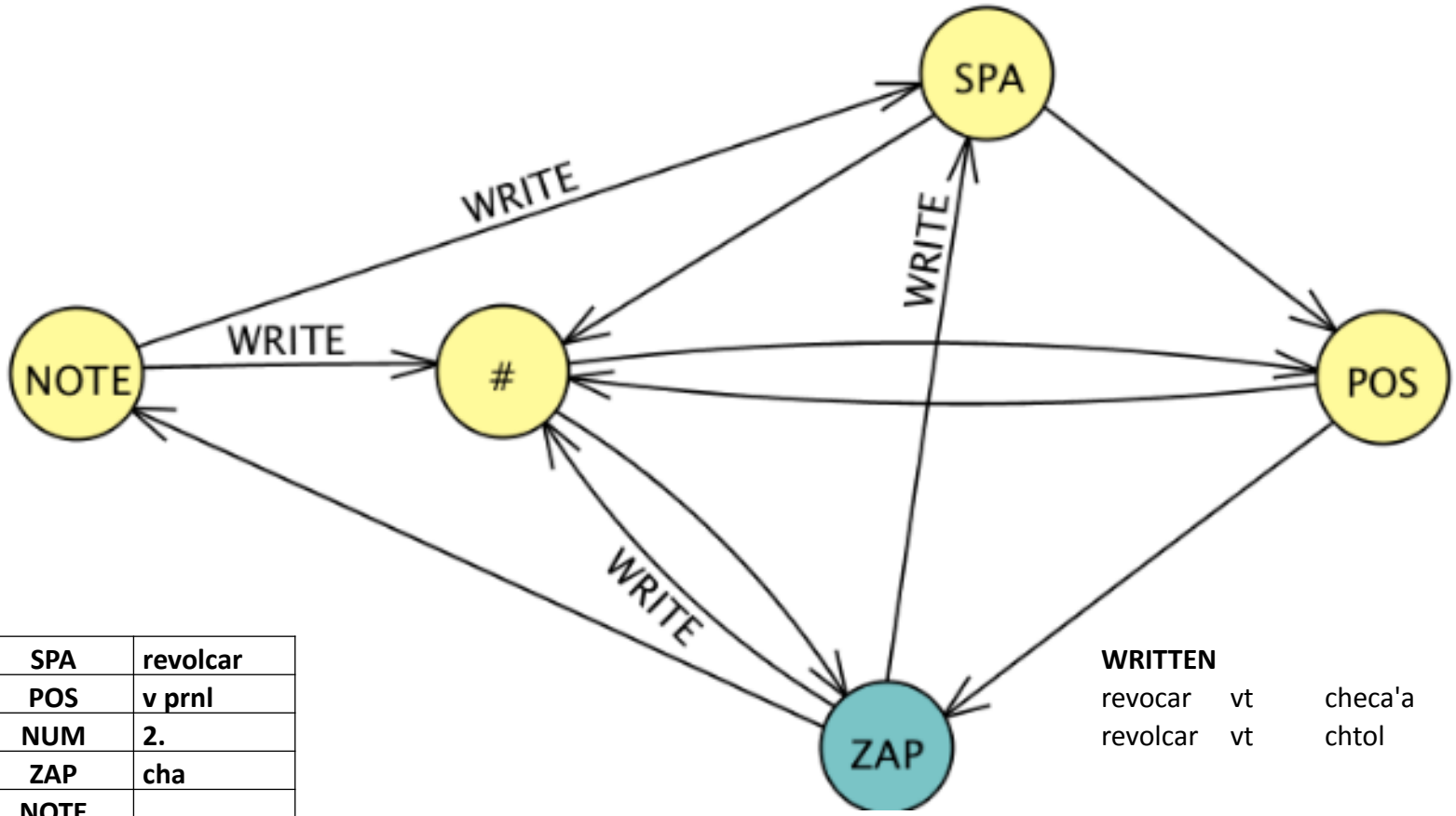


SPA	**revolcar**
POS	**v prnl**
NUM	**2.**
ZAP	**ch**
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol

revolcar 1. *vt* chtol 2. *v prnl* cha̲ta' **revolver** …

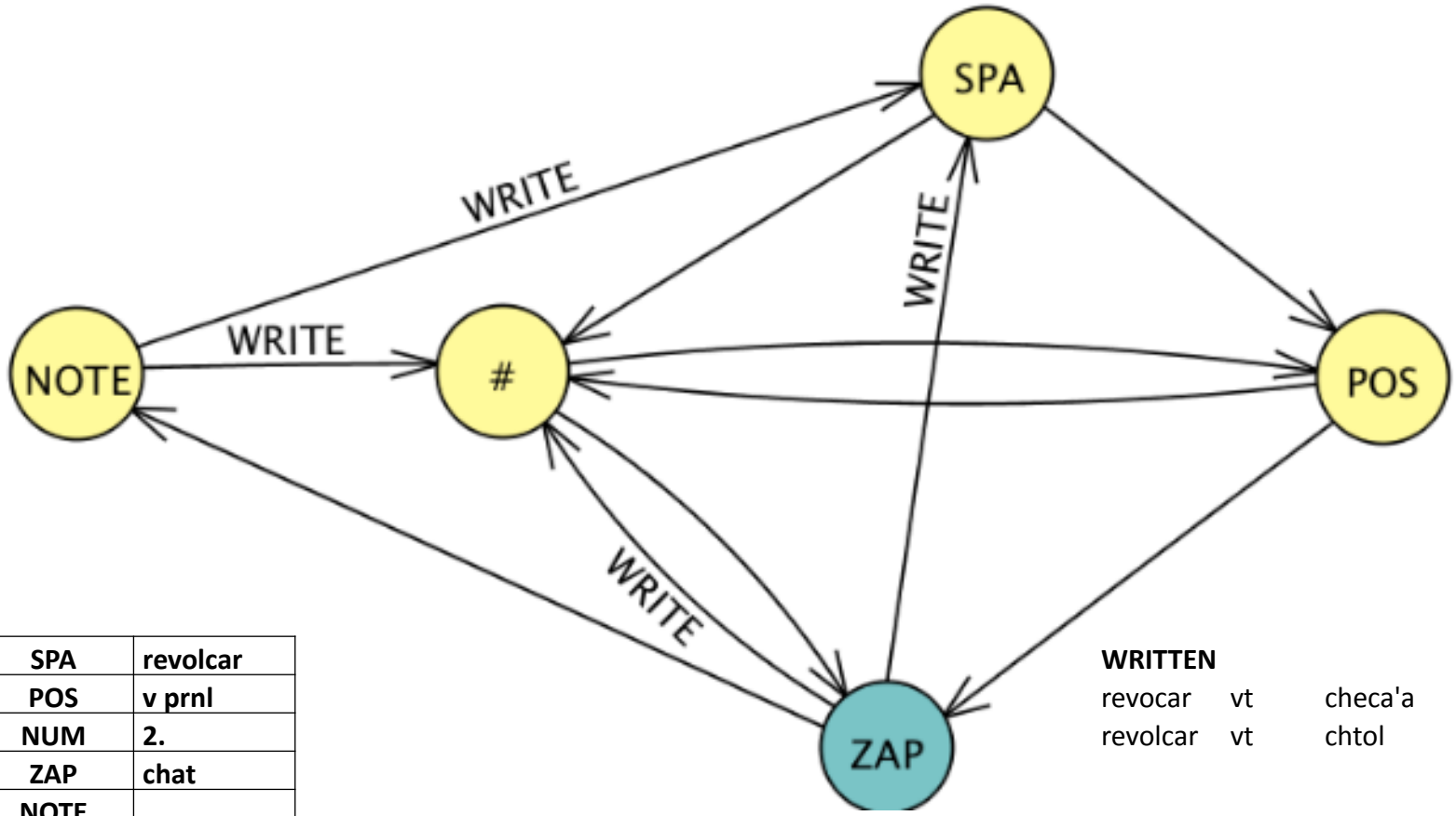


SPA	**revolcar**
POS	**v prnl**
NUM	**2.**
ZAP	**cha**
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol

revolcar 1. *vt* chtol 2. *v prnl* chaṭa' **revolver** …

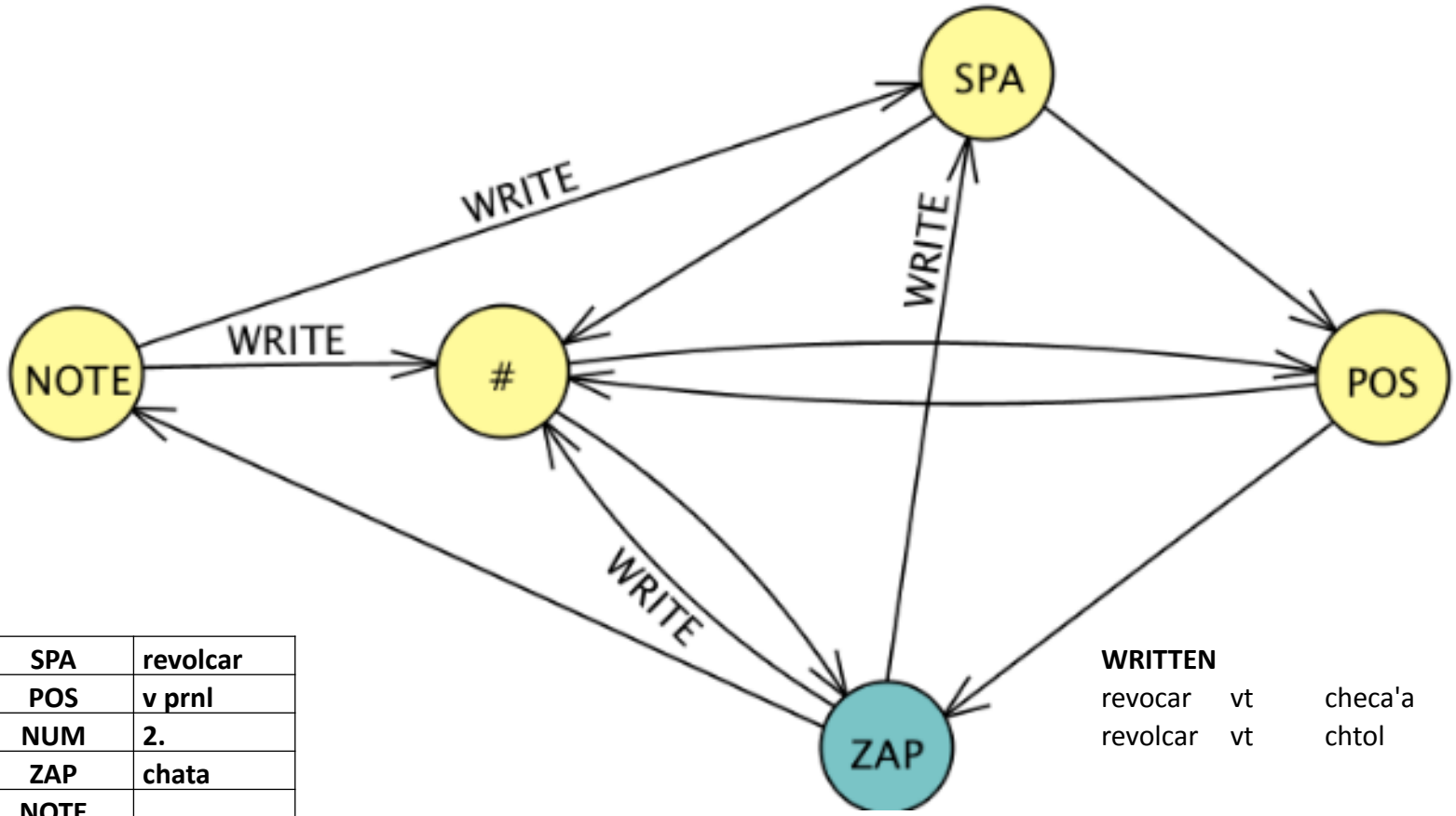


SPA	**revolcar**
POS	**v prnl**
NUM	**2.**
ZAP	**chat**
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol

revolcar 1. *vt* chtol 2. *v prnl* chata’ **revolver** …

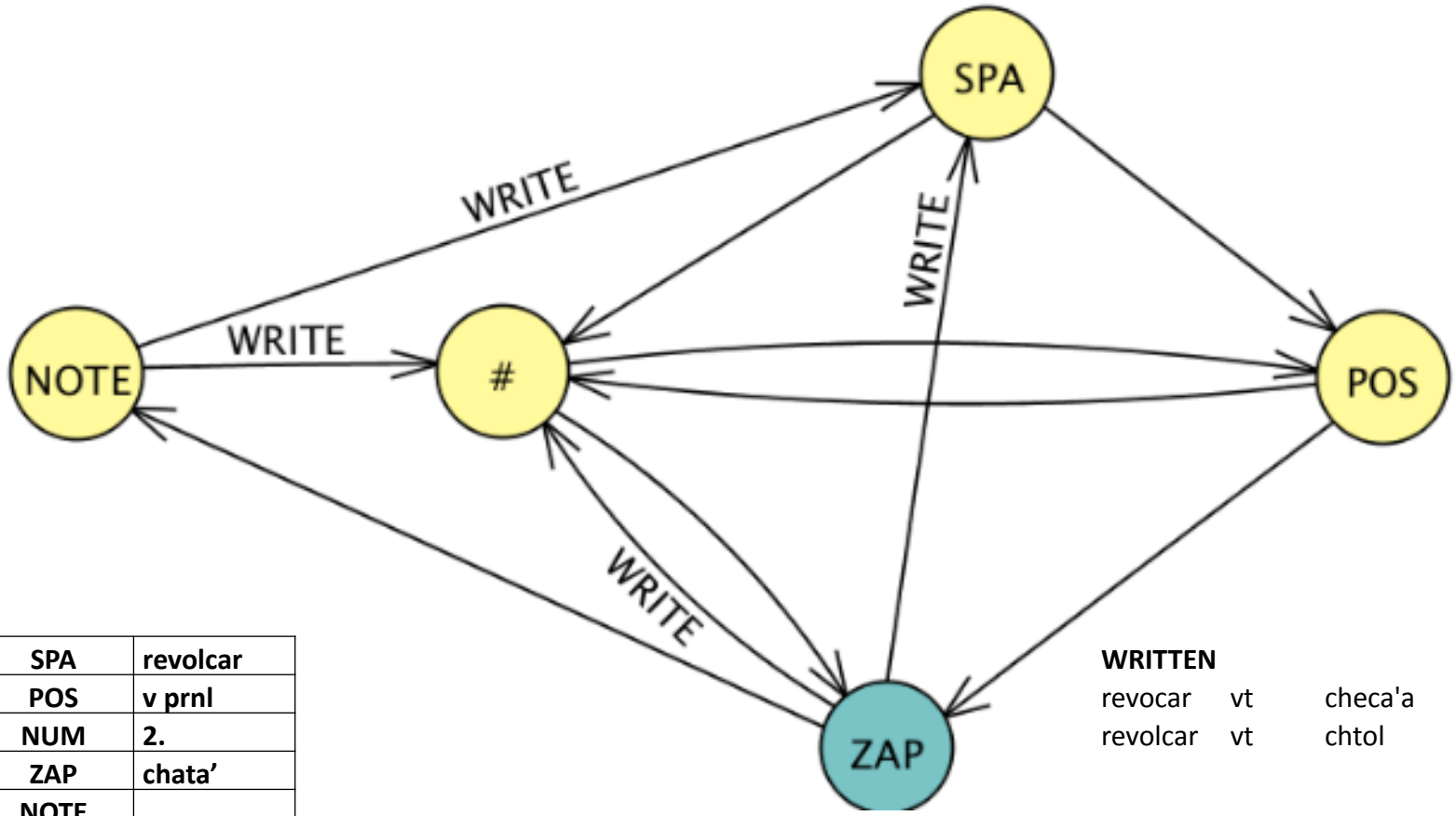


SPA	**revolcar**
POS	**v prnl**
NUM	**2.**
ZAP	**chata**
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol

revolcar 1. *vt* chtol 2. *v prnl* chata’ **revolver** …

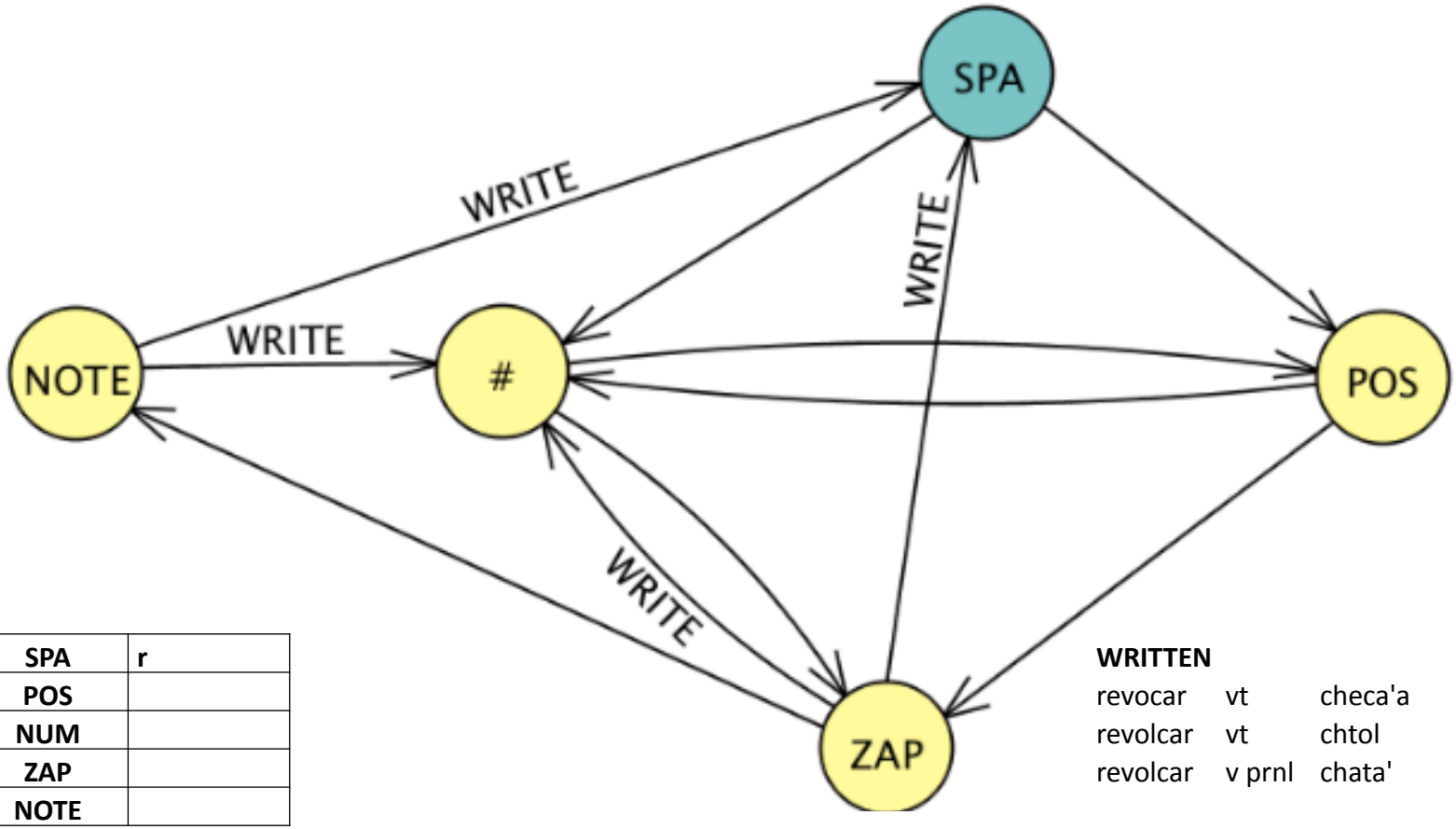


SPA	**revolcar**
POS	**v prnl**
NUM	**2.**
ZAP	**chata’**
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol

revolcar 1. *vt* chtol 2. *v prnl* chata’ **revolver** …



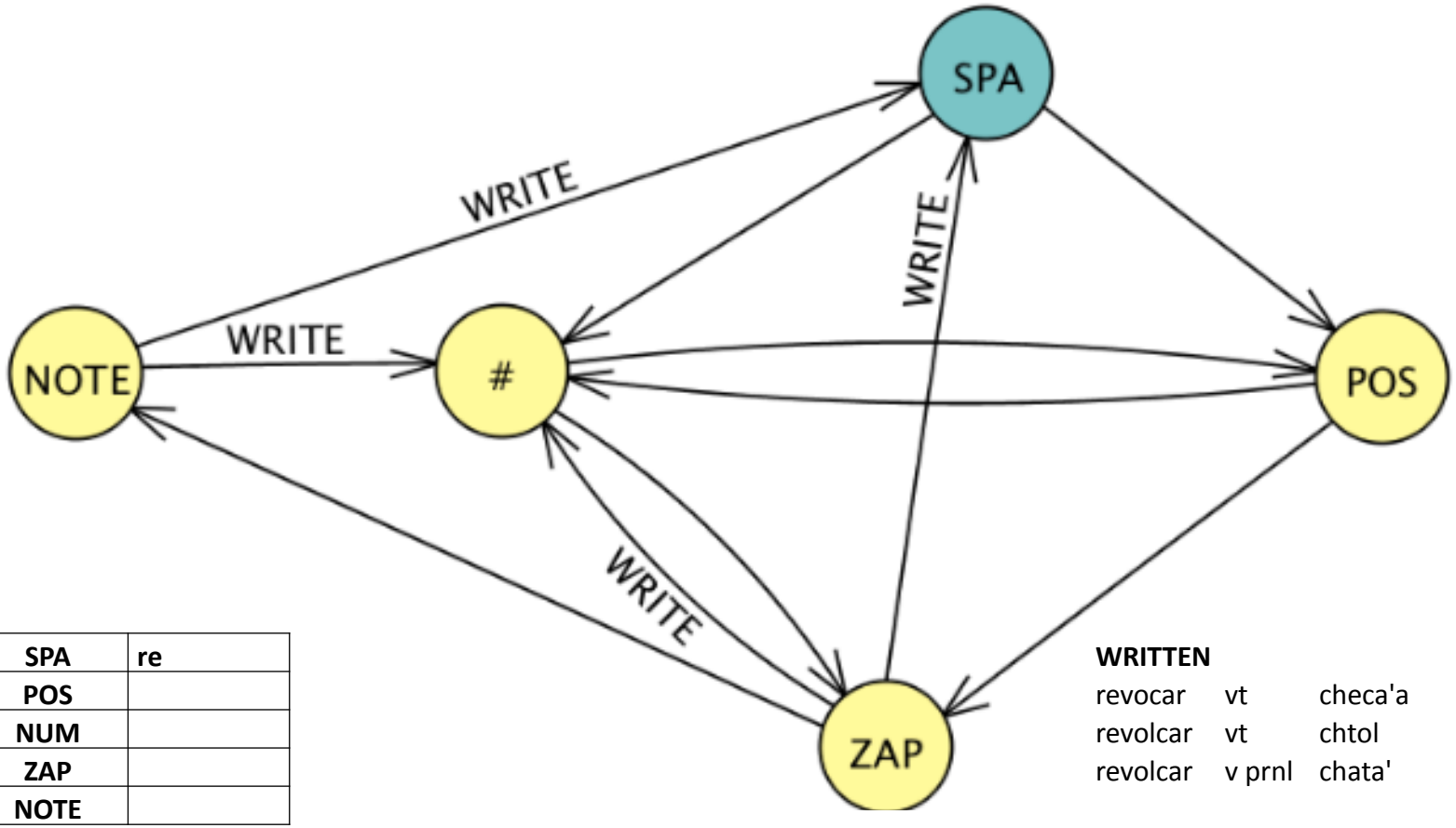
SPA	**r**
POS	
NUM	
ZAP	
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'

revolver *vt* 1. chix 2. chta (líquido)

rey …

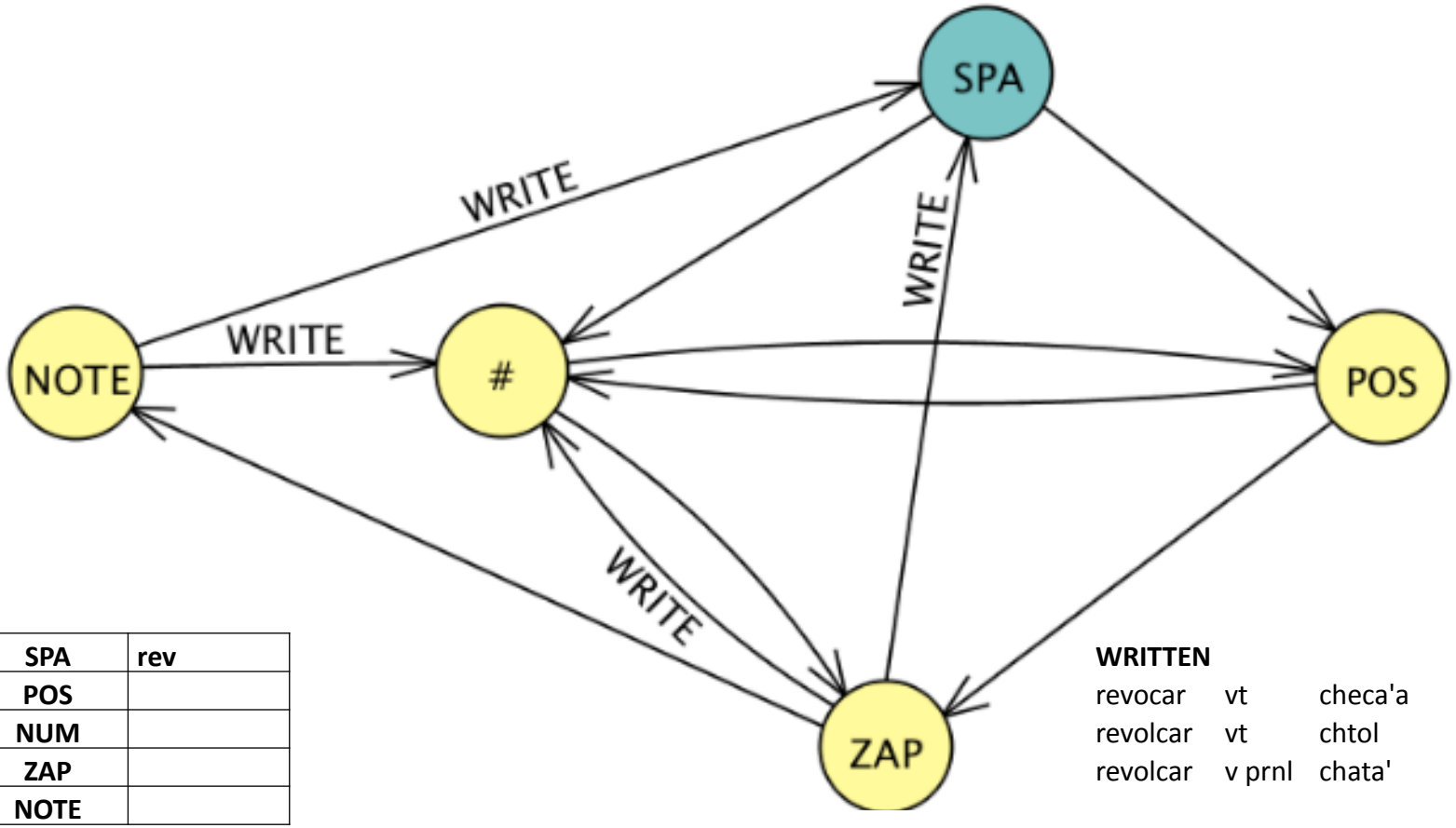


SPA	re
POS	
NUM	
ZAP	
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'

revolver *vt* 1. chix 2. chta (líquido)
rey …

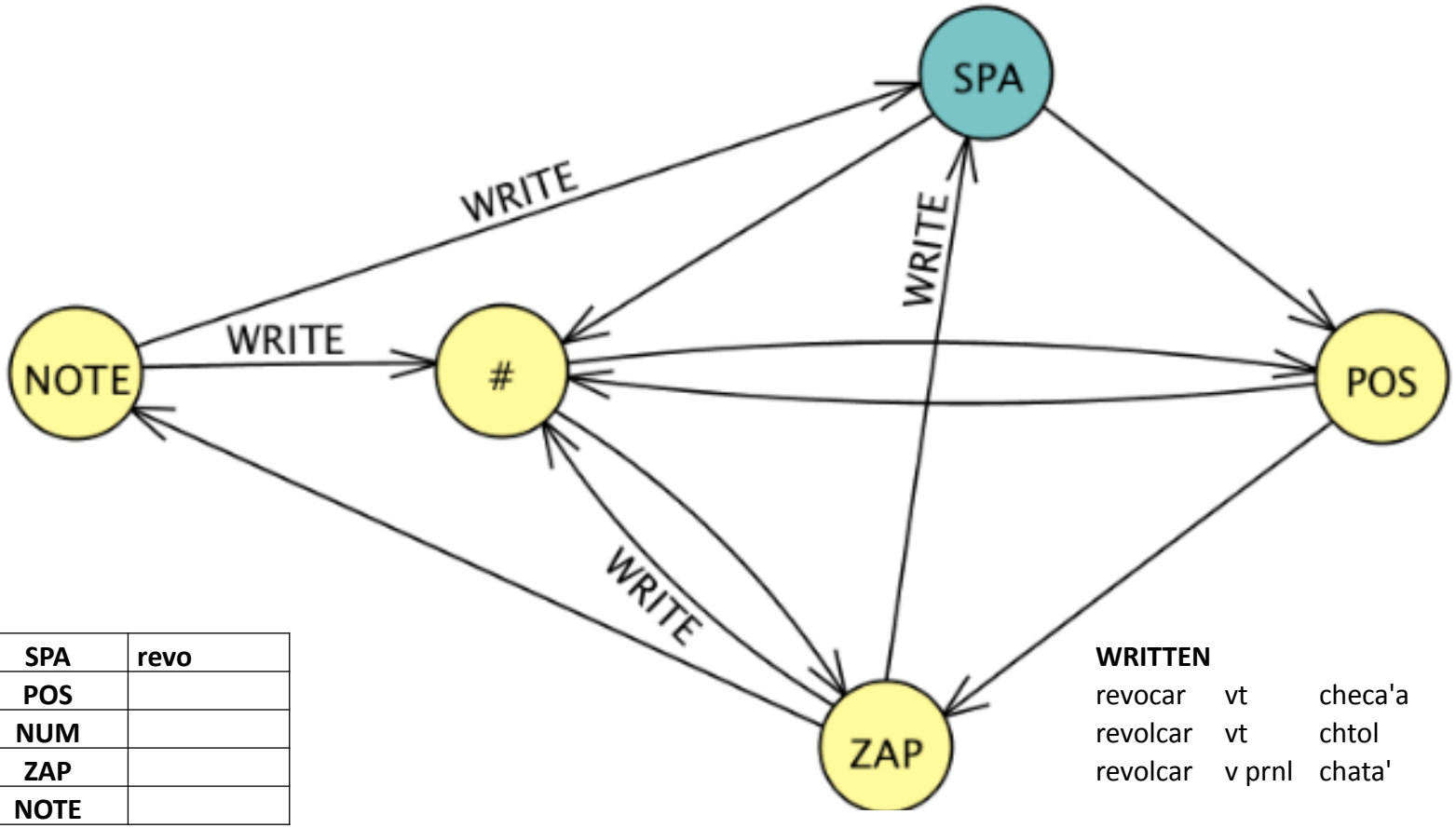


SPA	rev
POS	
NUM	
ZAP	
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'

revolver *vt* 1. chix 2. chta (líquido)
rey …

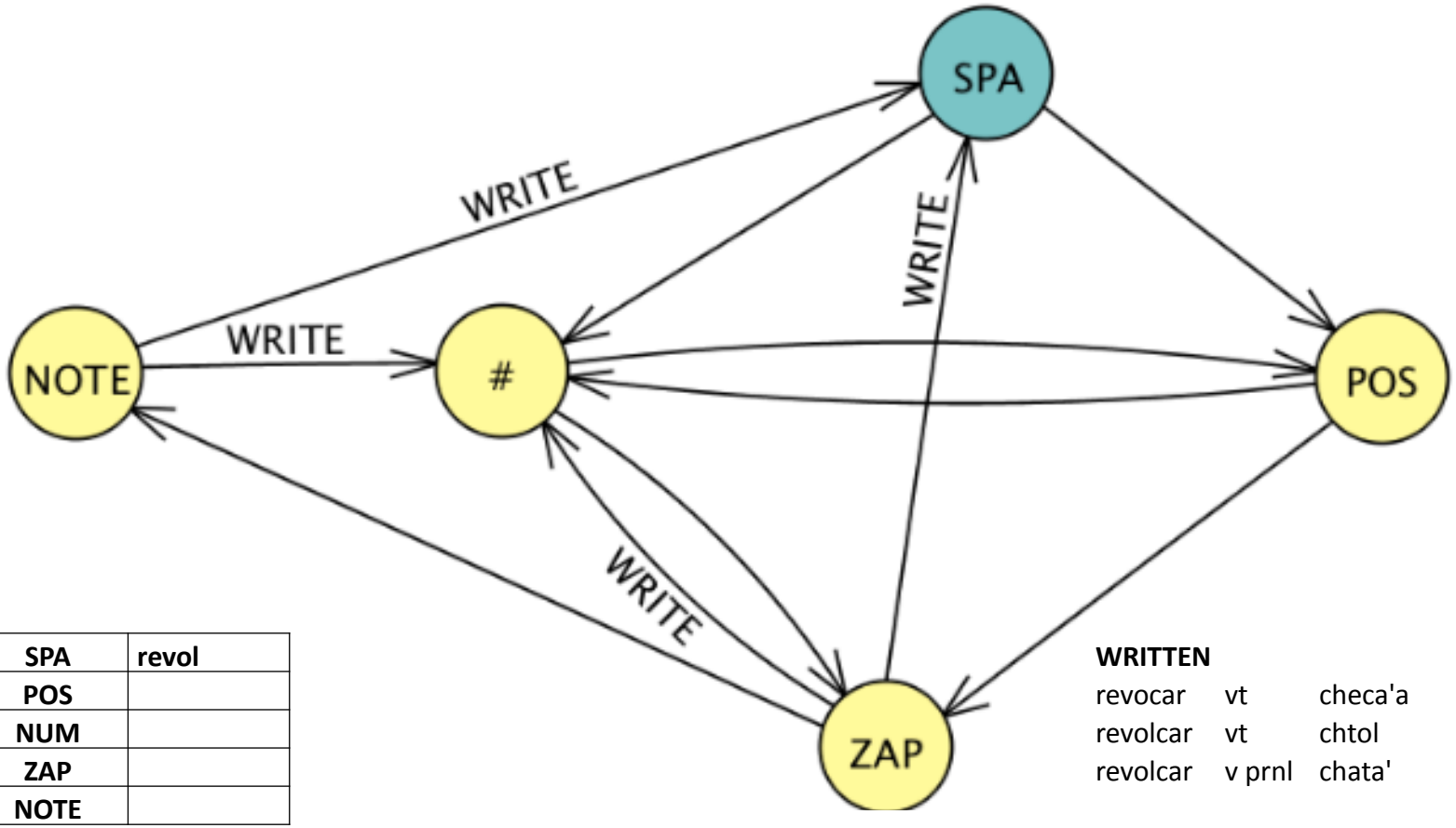


SPA	revo
POS	
NUM	
ZAP	
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'

revolver *vt* 1. chix 2. chta (líquido) **rey** …

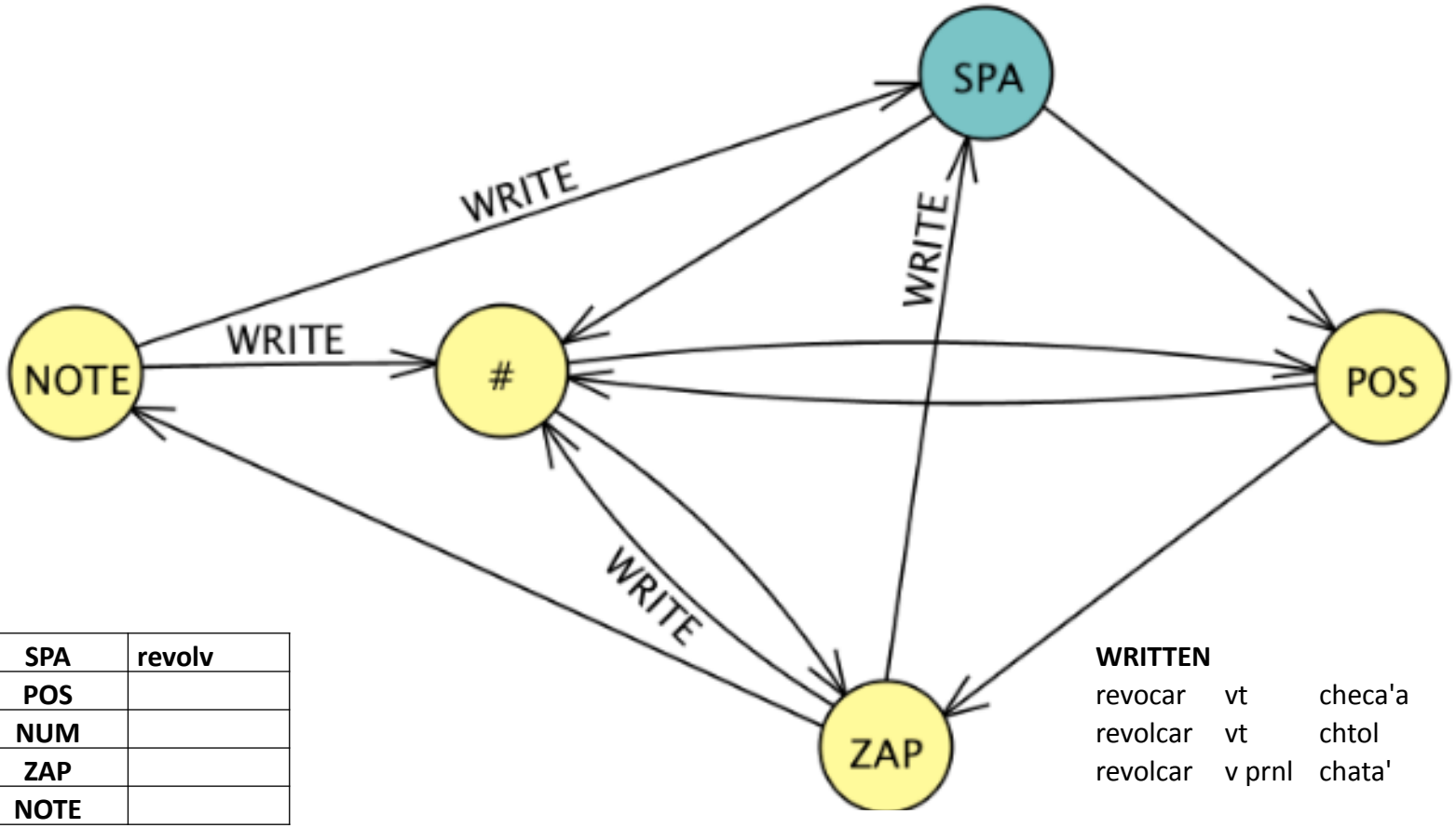


SPA	revol
POS	
NUM	
ZAP	
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'

revolver *vt* 1. chix 2. chta (líquido) **rey** …

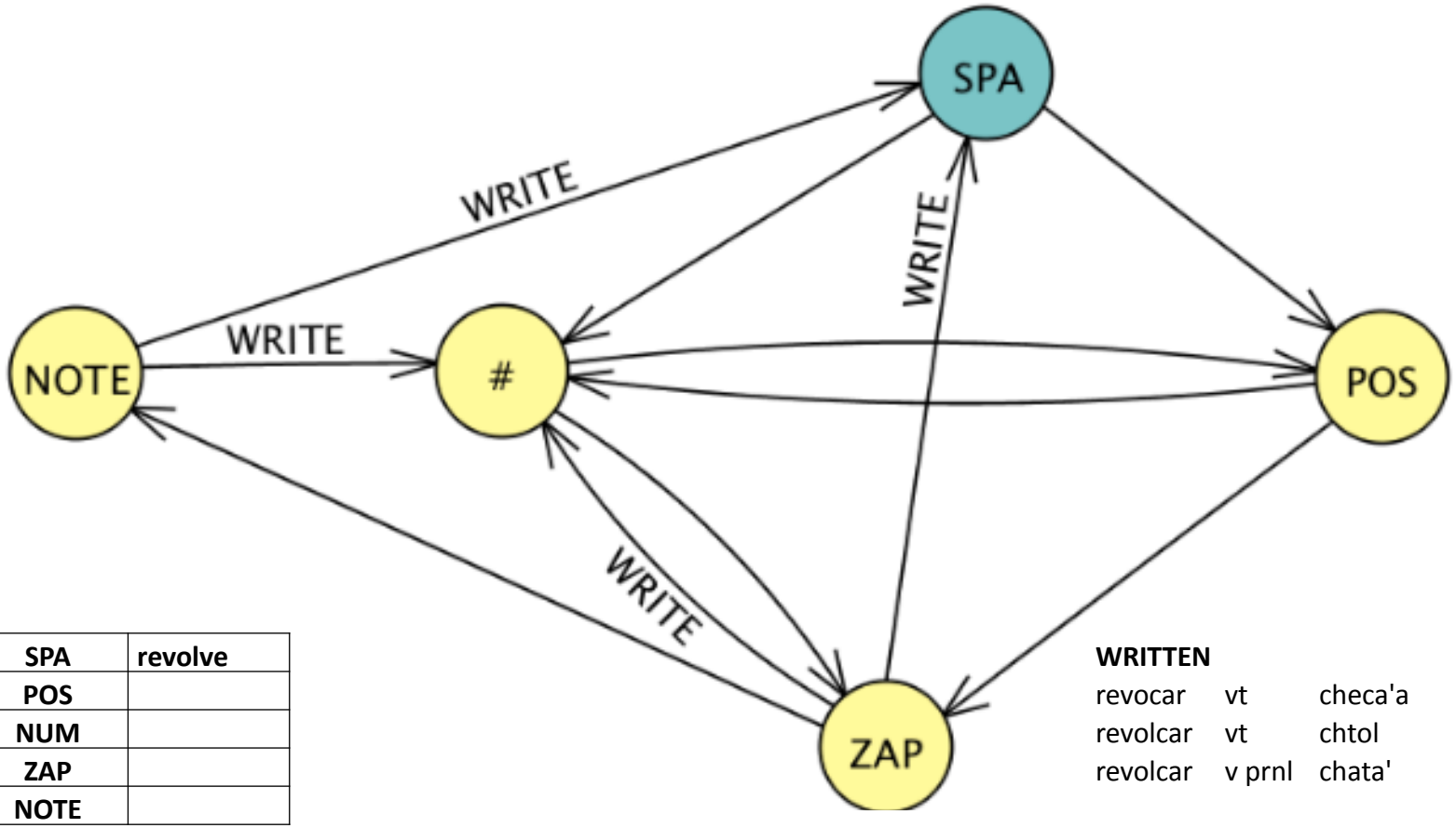


SPA	revolv
POS	
NUM	
ZAP	
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'

revolver *vt* 1. chix 2. chta (líquido) **rey** …



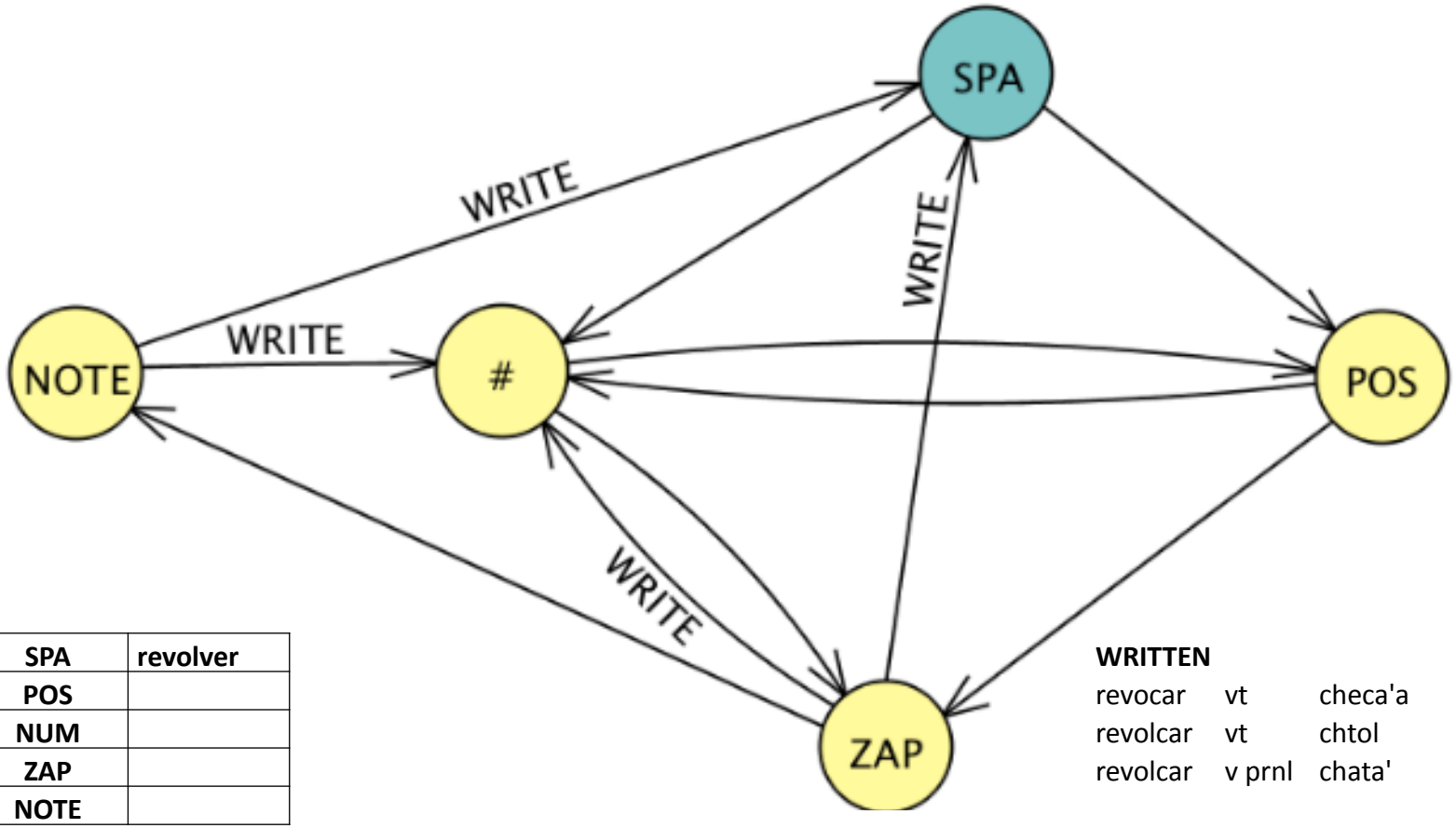
SPA	revolve
POS	
NUM	
ZAP	
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'

revolver *vt* 1. chix 2. chta (líquido)

rey …

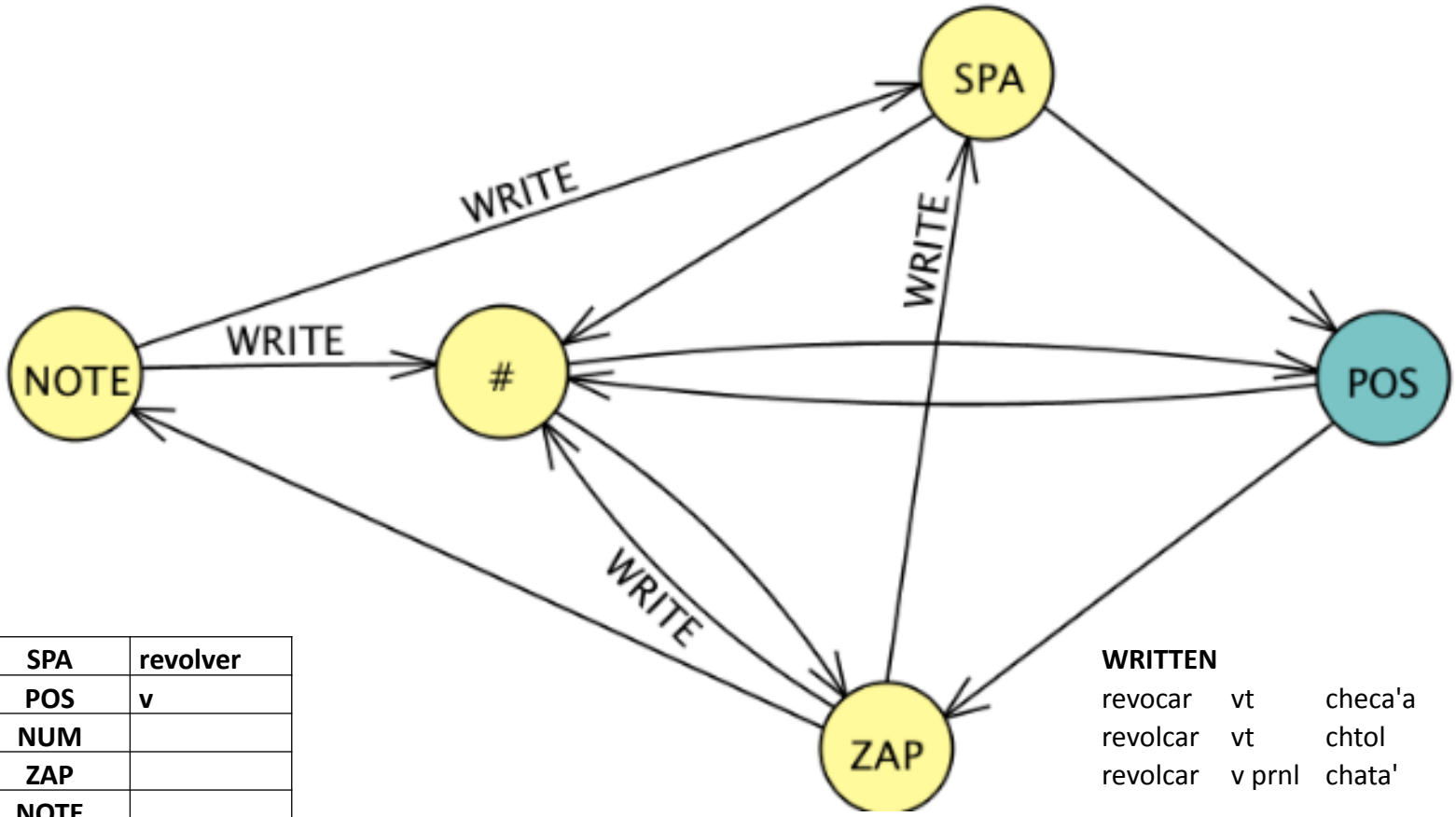


SPA	revolver
POS	
NUM	
ZAP	
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'

revolver *vt* 1. chix 2. chta (líquido) **rey** …

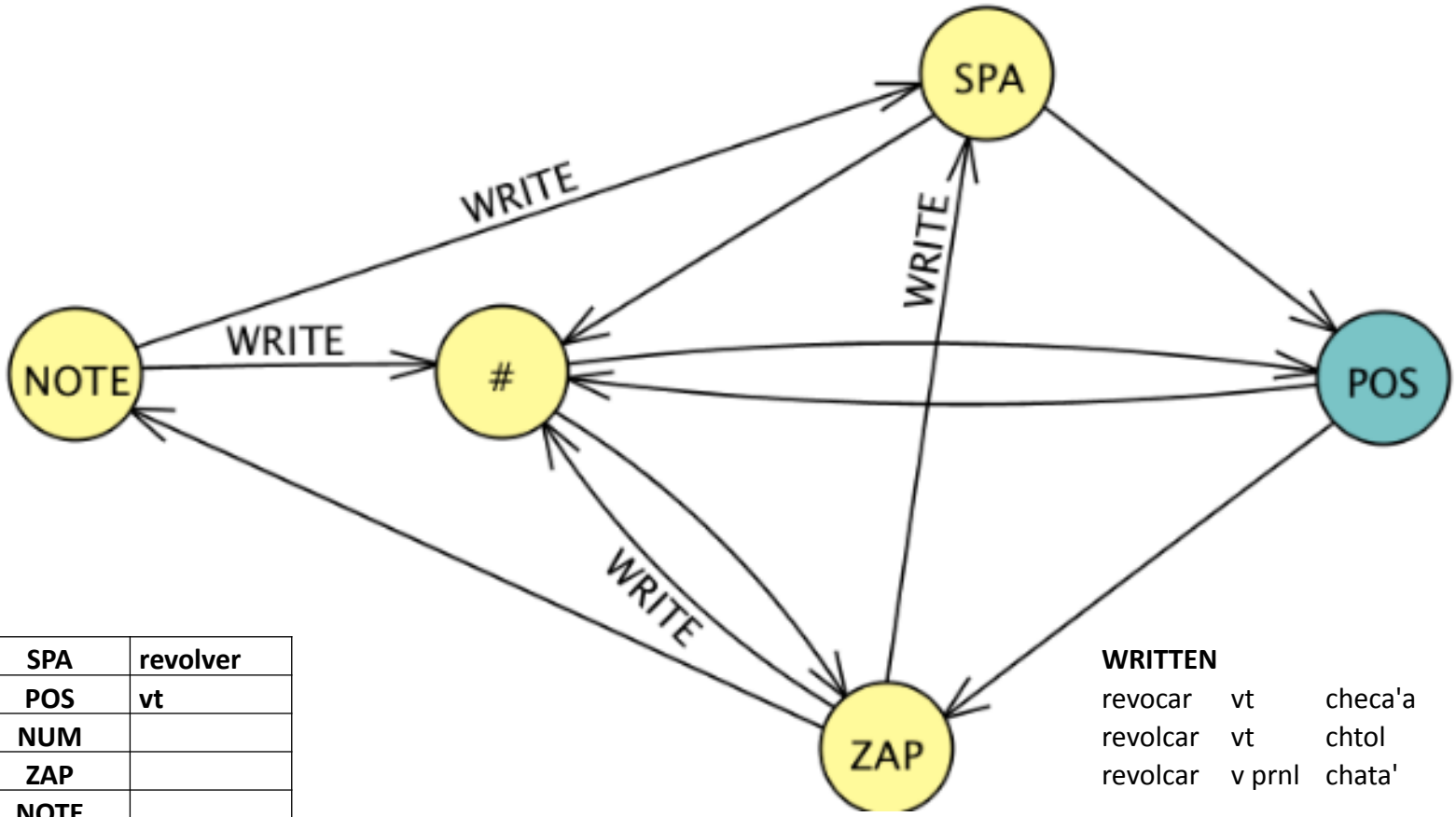


SPA	revolver
POS	v
NUM	
ZAP	
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'

revolver *vt* 1. chix 2. chta (líquido) **rey** …

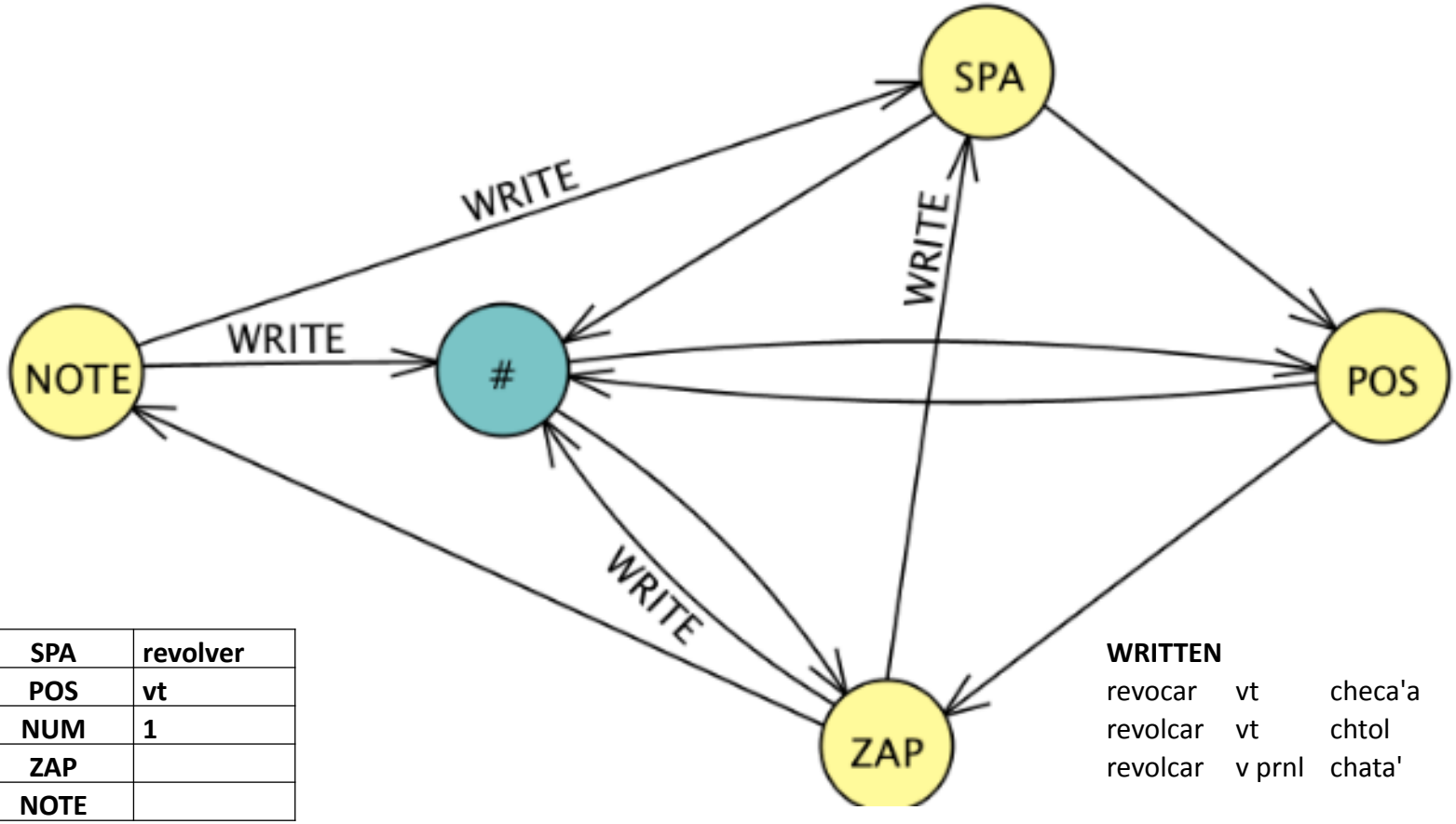


SPA	**revolver**
POS	**vt**
NUM	
ZAP	
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'

revolver *vt* 1. chix 2. chta (líquido) **rey** …

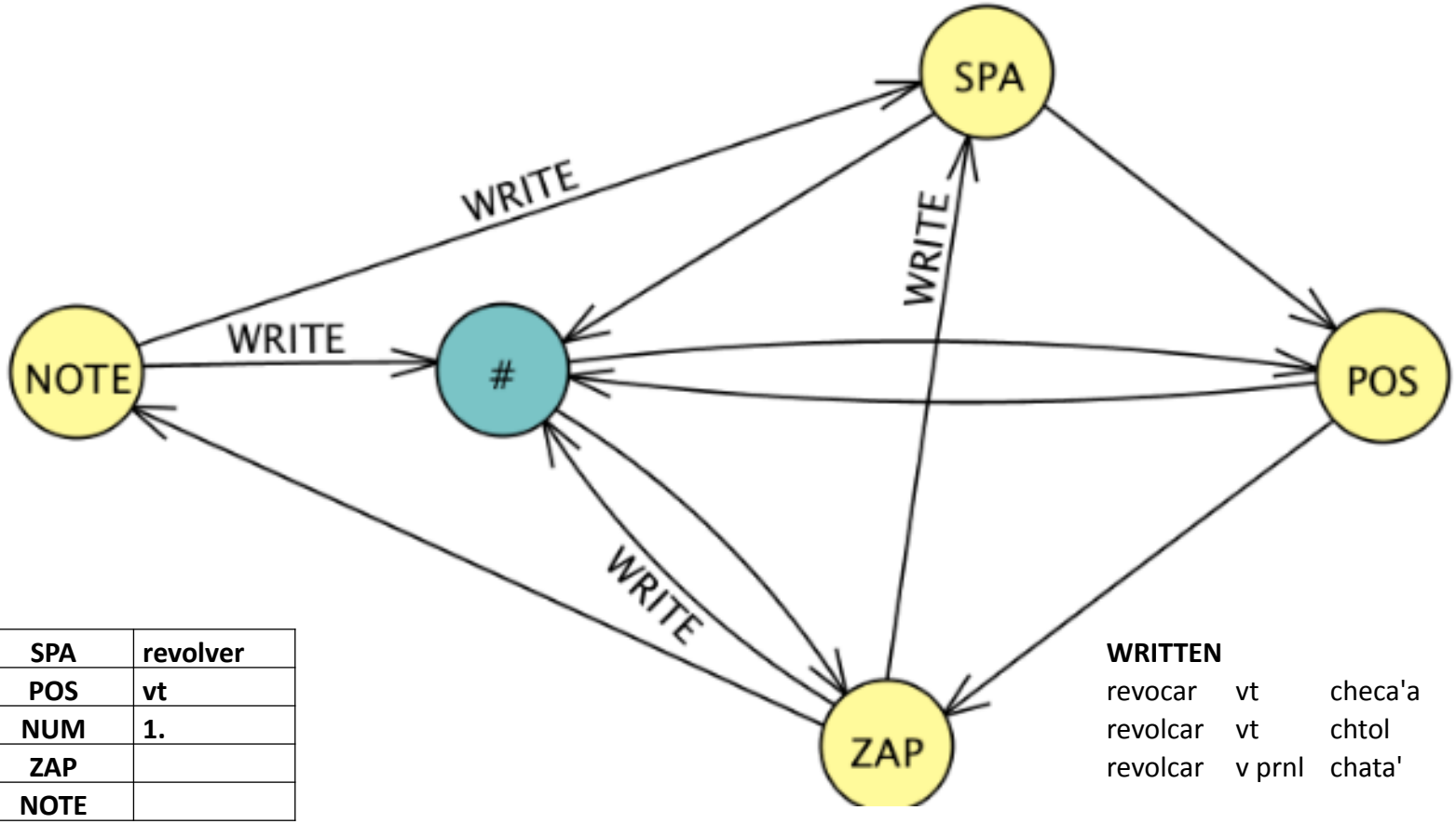


SPA	**revolver**
POS	**vt**
NUM	**1**
ZAP	
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'

revolver *vt* 1. chix 2. chta (líquido) **rey** …

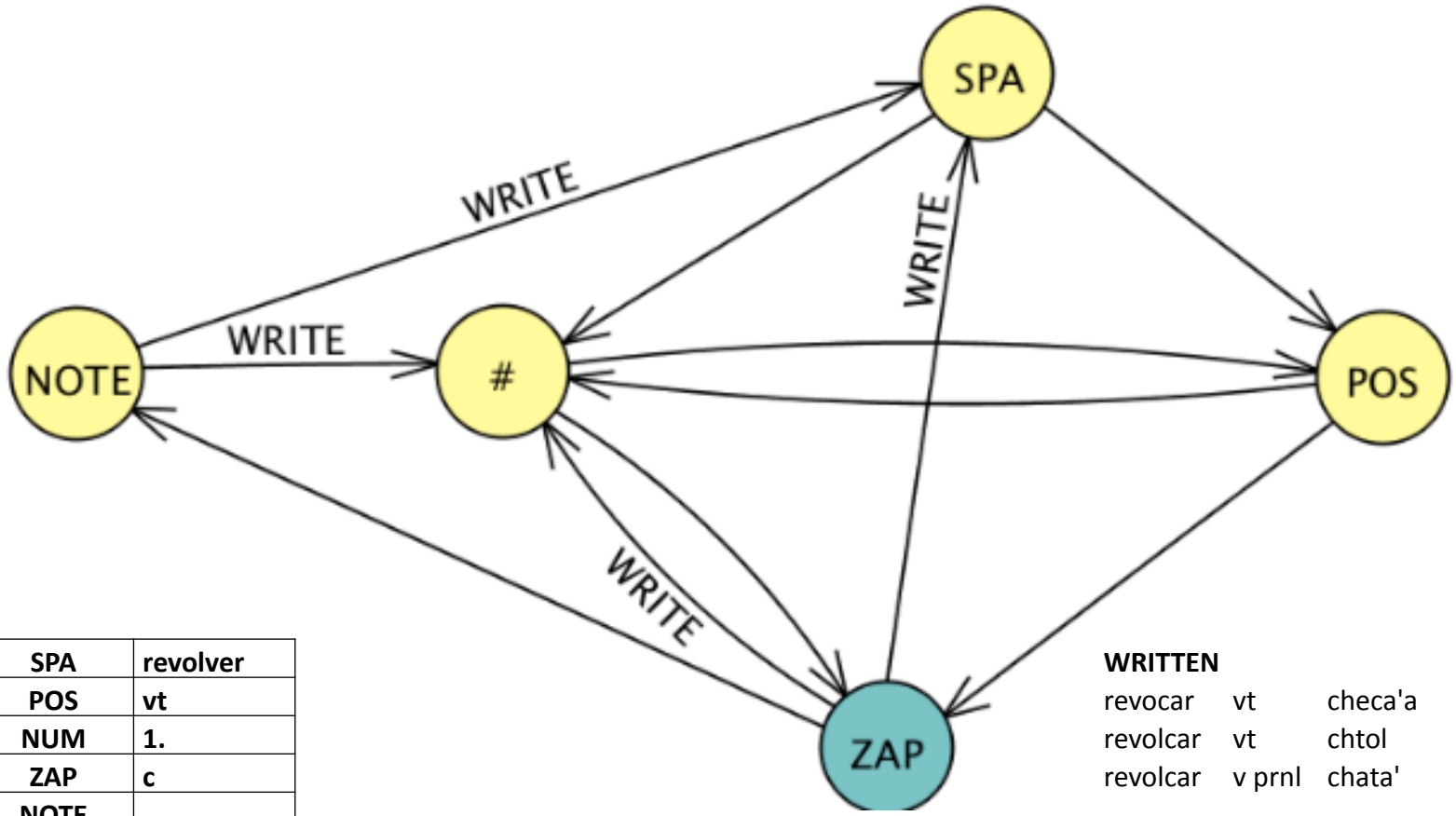


SPA	**revolver**
POS	**vt**
NUM	**1.**
ZAP	
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'

revolver *vt* 1. chix 2. chta (líquido) **rey** …

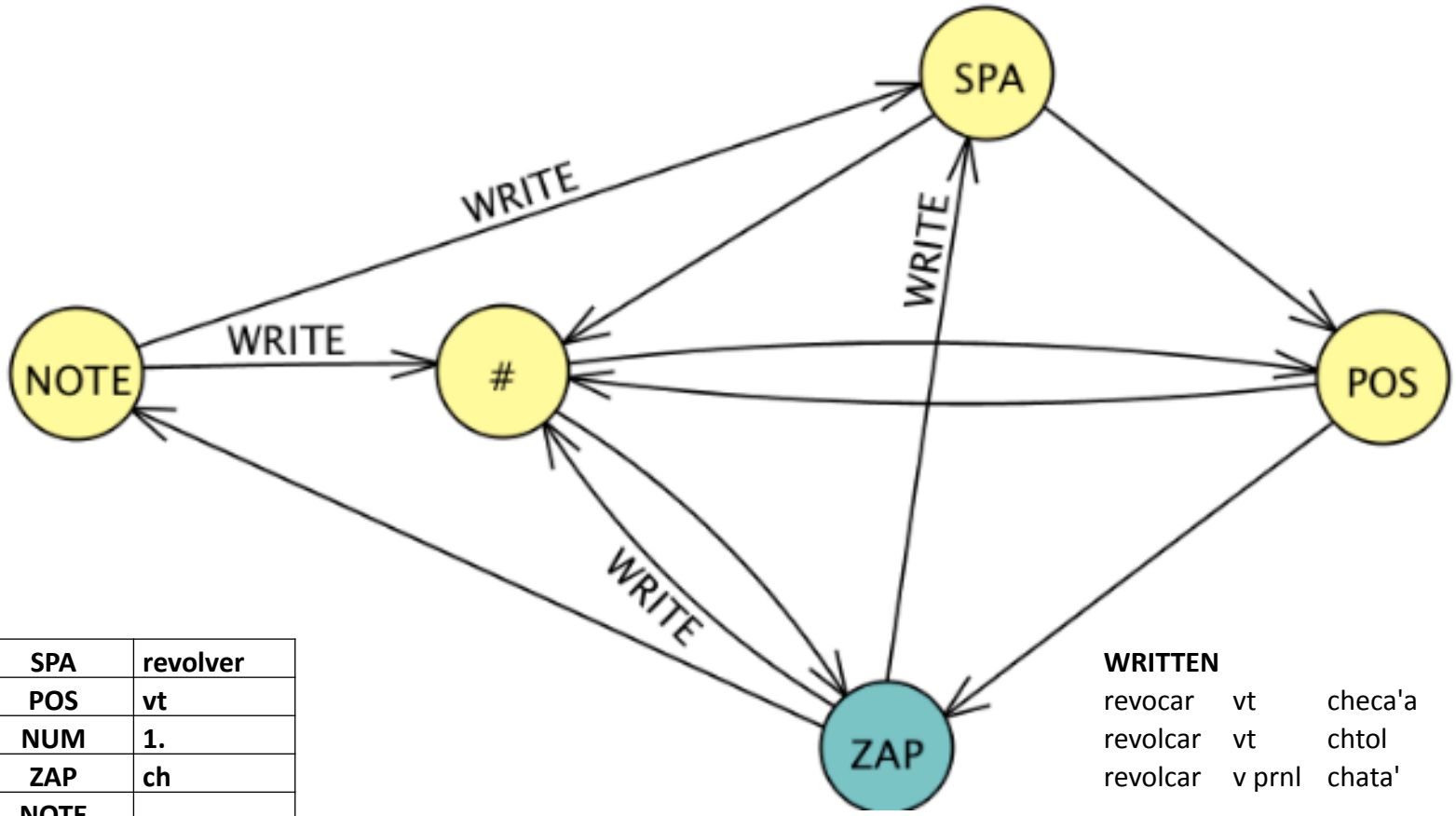


SPA	**revolver**
POS	**vt**
NUM	**1.**
ZAP	**c**
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'

revolver *vt* 1. chix 2. chta (líquido) **rey** …

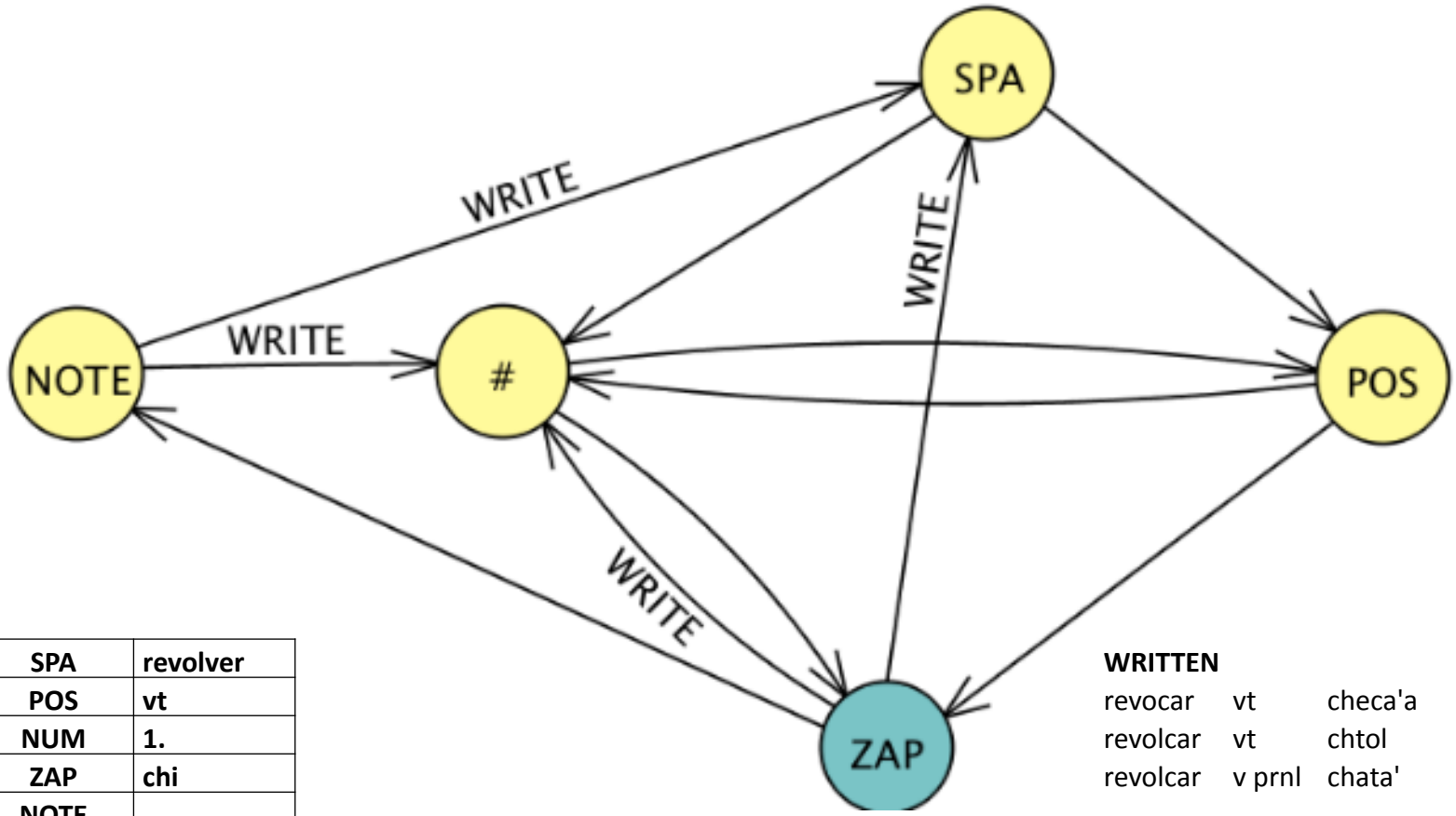


SPA	**revolver**
POS	**vt**
NUM	**1.**
ZAP	**ch**
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'

revolver *vt* 1. chix 2. chta (líquido) **rey** …

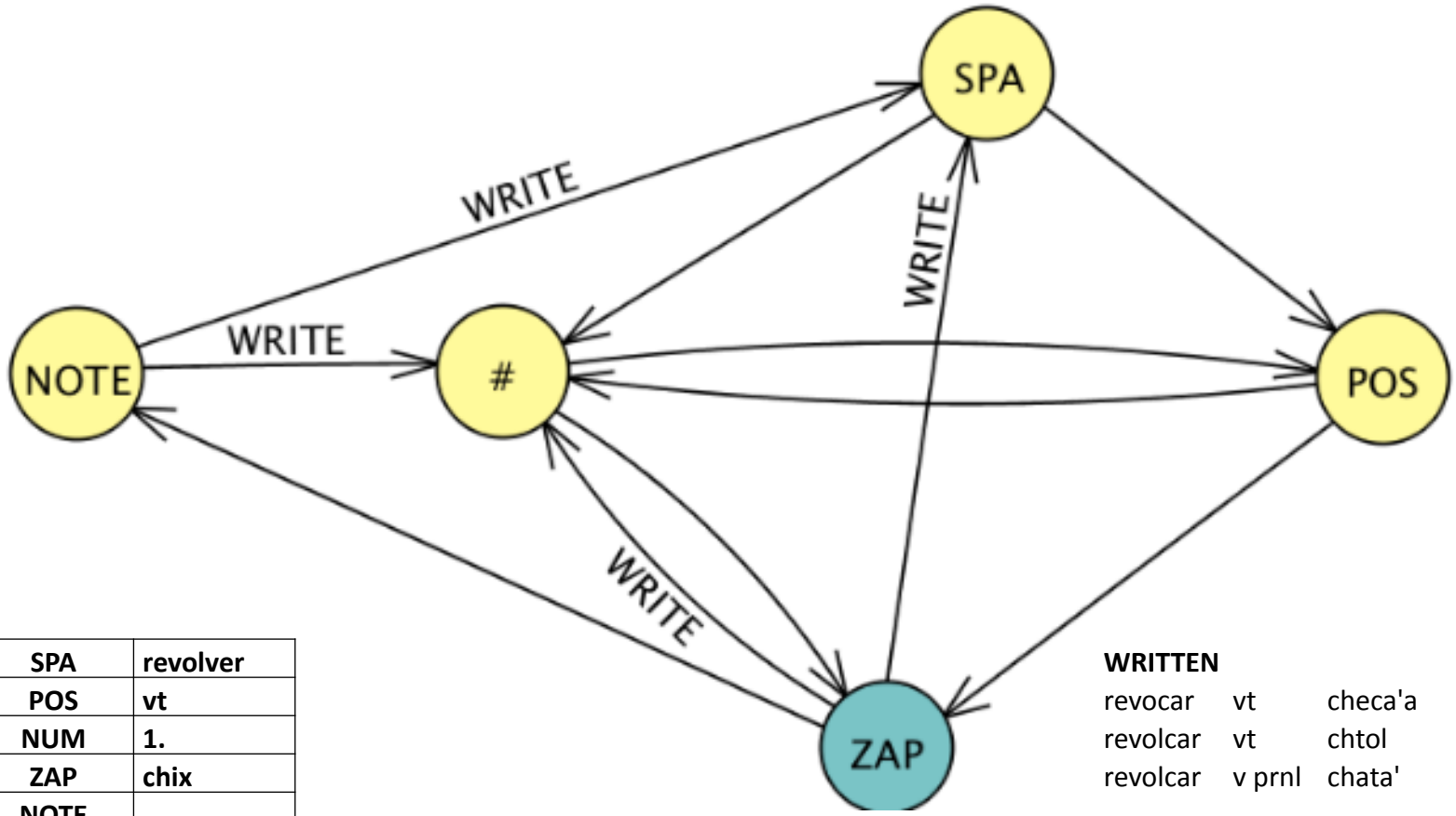


SPA	**revolver**
POS	**vt**
NUM	**1.**
ZAP	**chi**
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'

revolver *vt* 1. chix 2. chta (líquido) **rey** …

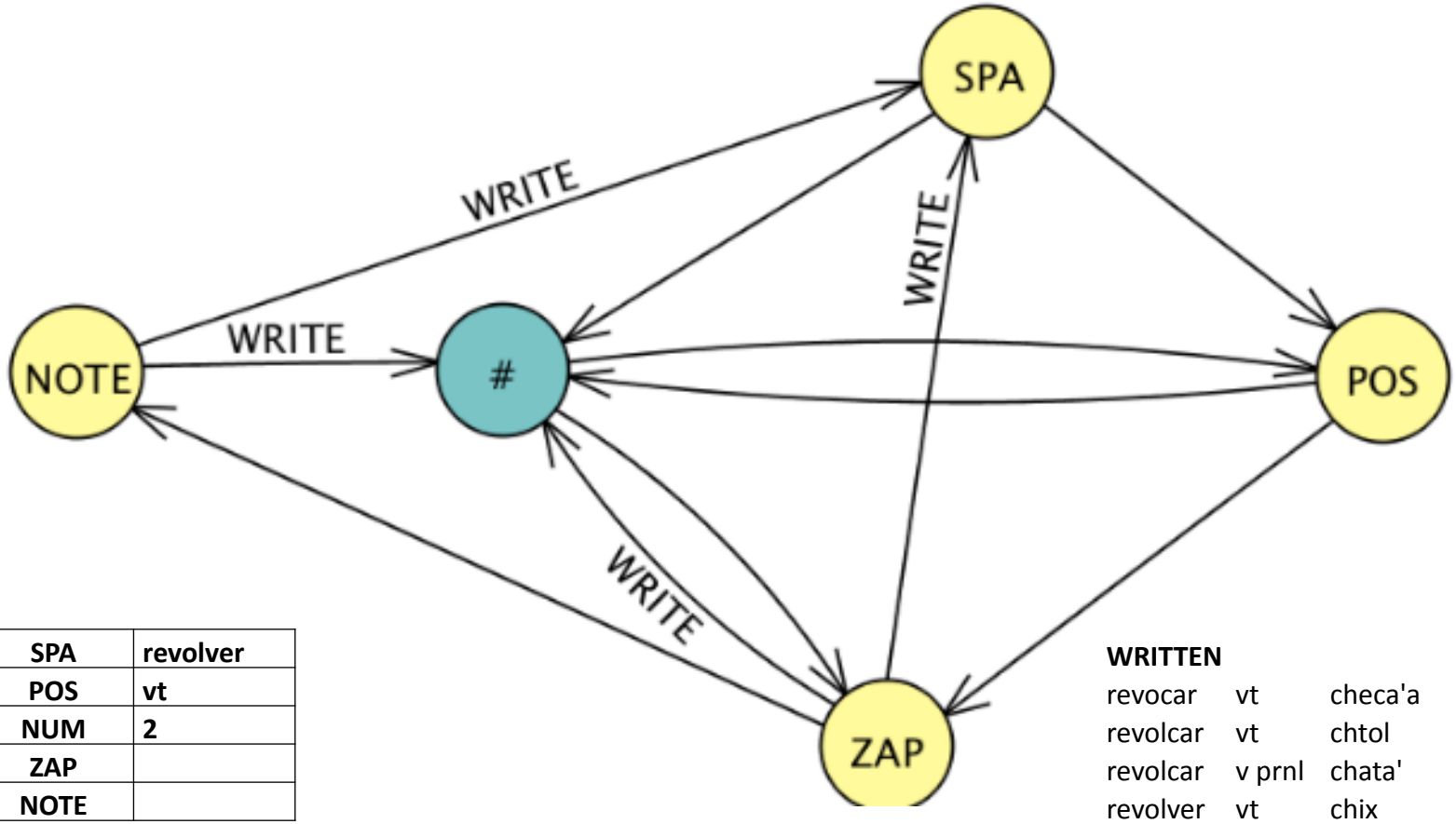


SPA	**revolver**
POS	**vt**
NUM	**1.**
ZAP	**chix**
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'

revolver *vt* 1. chix 2. chta (líquido) **rey** …

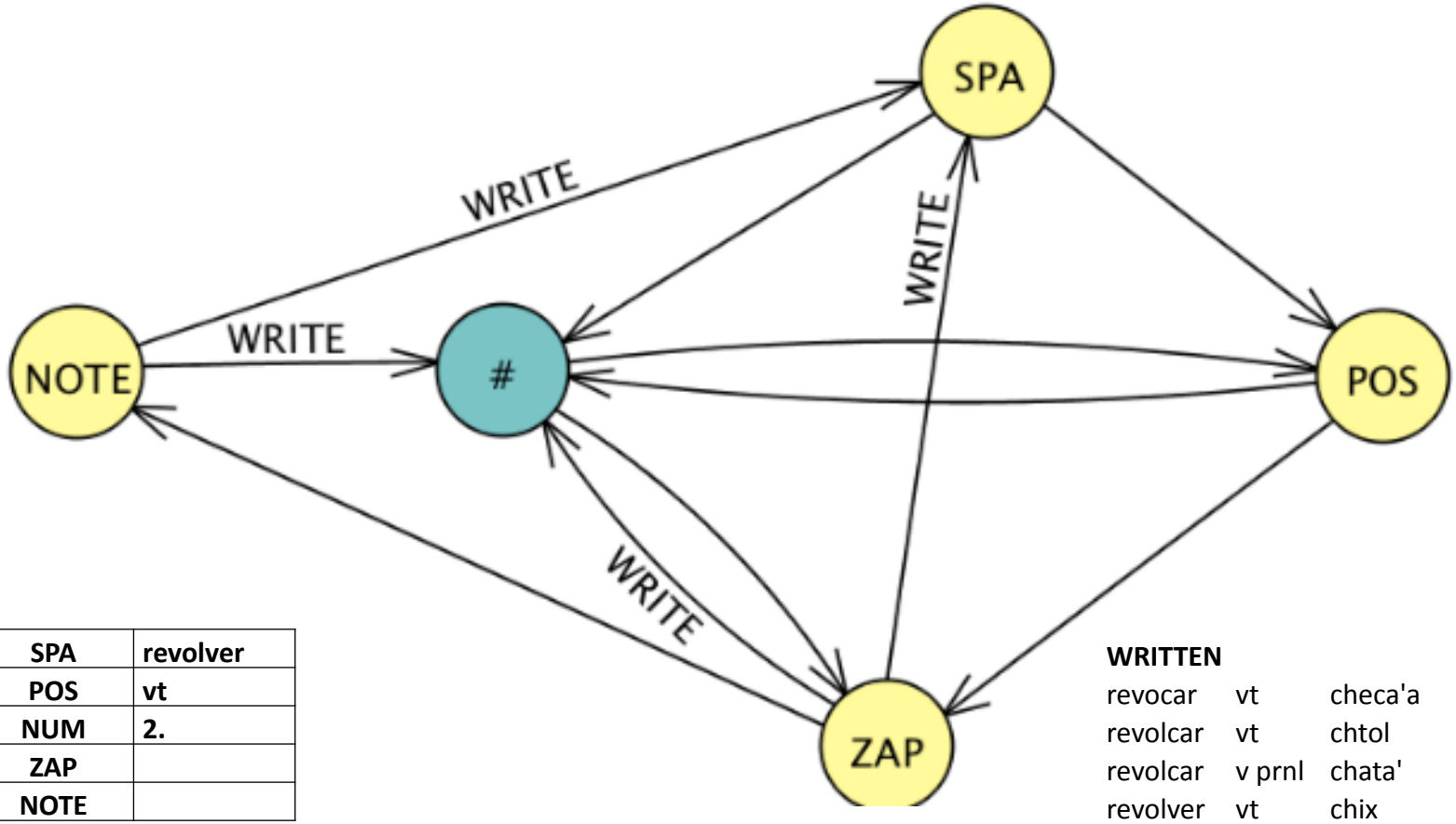


SPA	**revolver**
POS	**vt**
NUM	**2**
ZAP	
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'
revolver	vt	chix

revolver *vt* 1. chix 2. chta (líquido) **rey** …

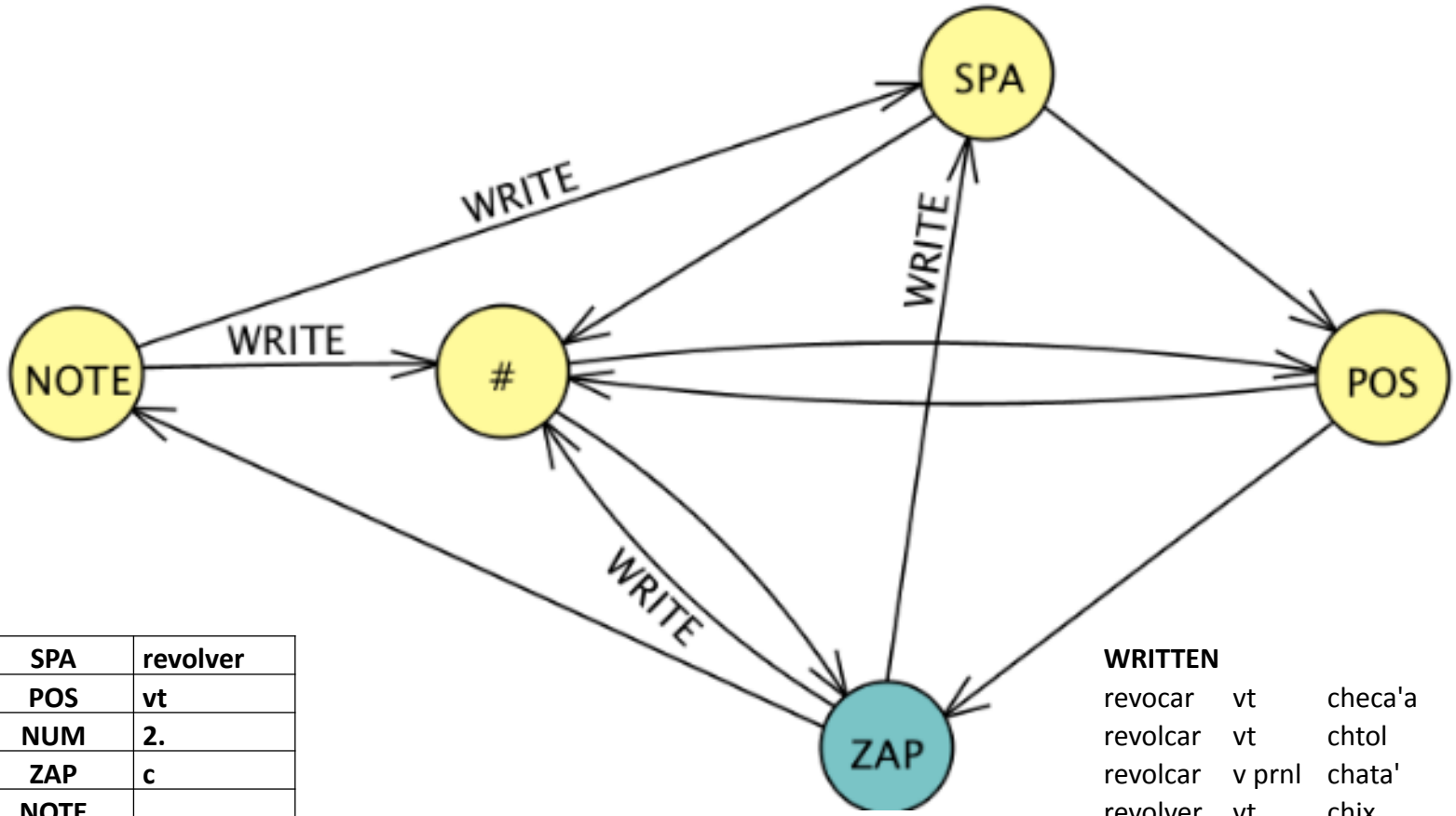


SPA	**revolver**
POS	**vt**
NUM	**2.**
ZAP	
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'
revolver	vt	chix

revolver *vt* 1. chix 2. chta (líquido)
rey …

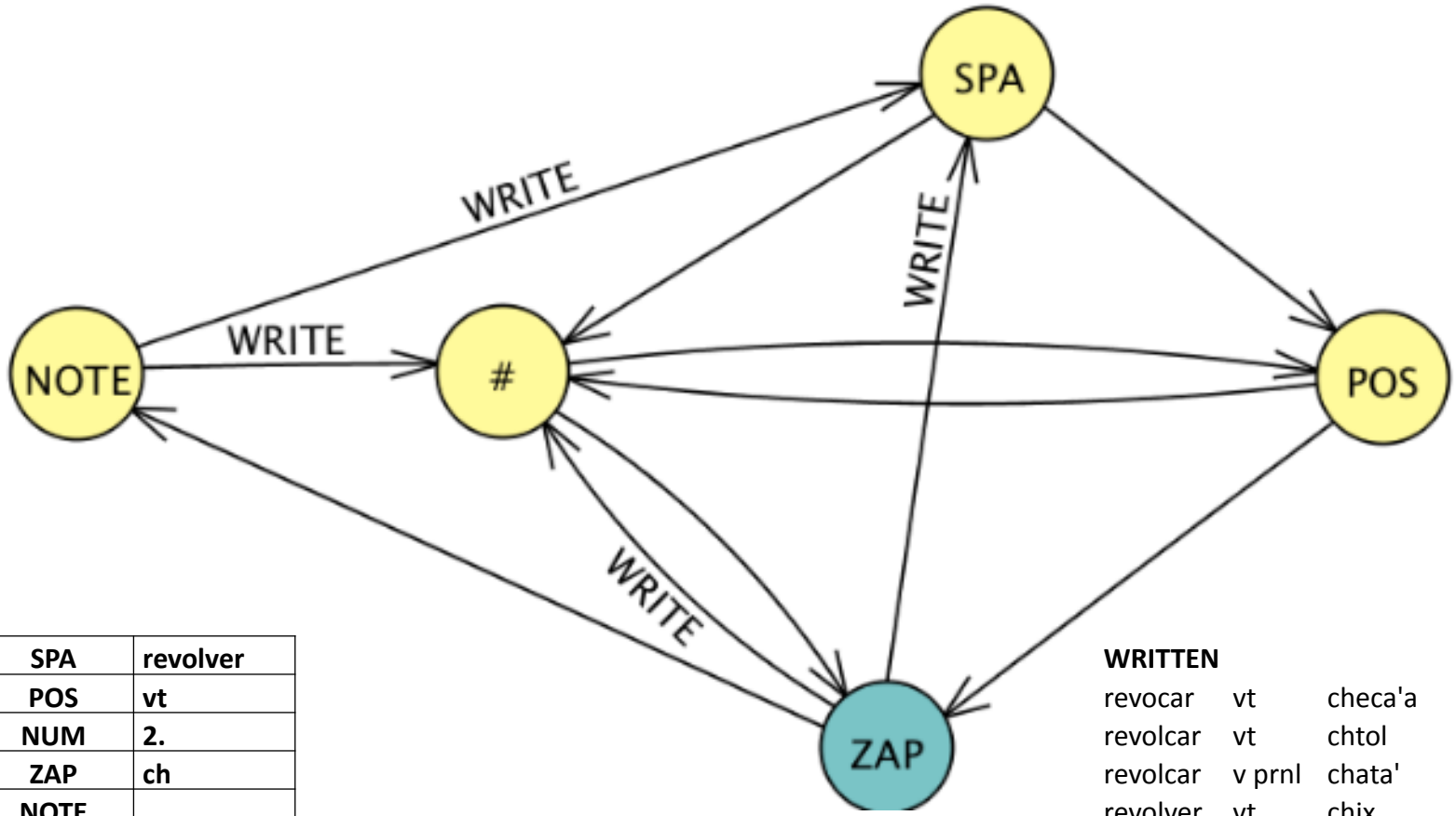


SPA	**revolver**
POS	**vt**
NUM	**2.**
ZAP	**c**
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'
revolver	vt	chix

revolver *vt* 1. chix 2. chta (líquido)
rey …

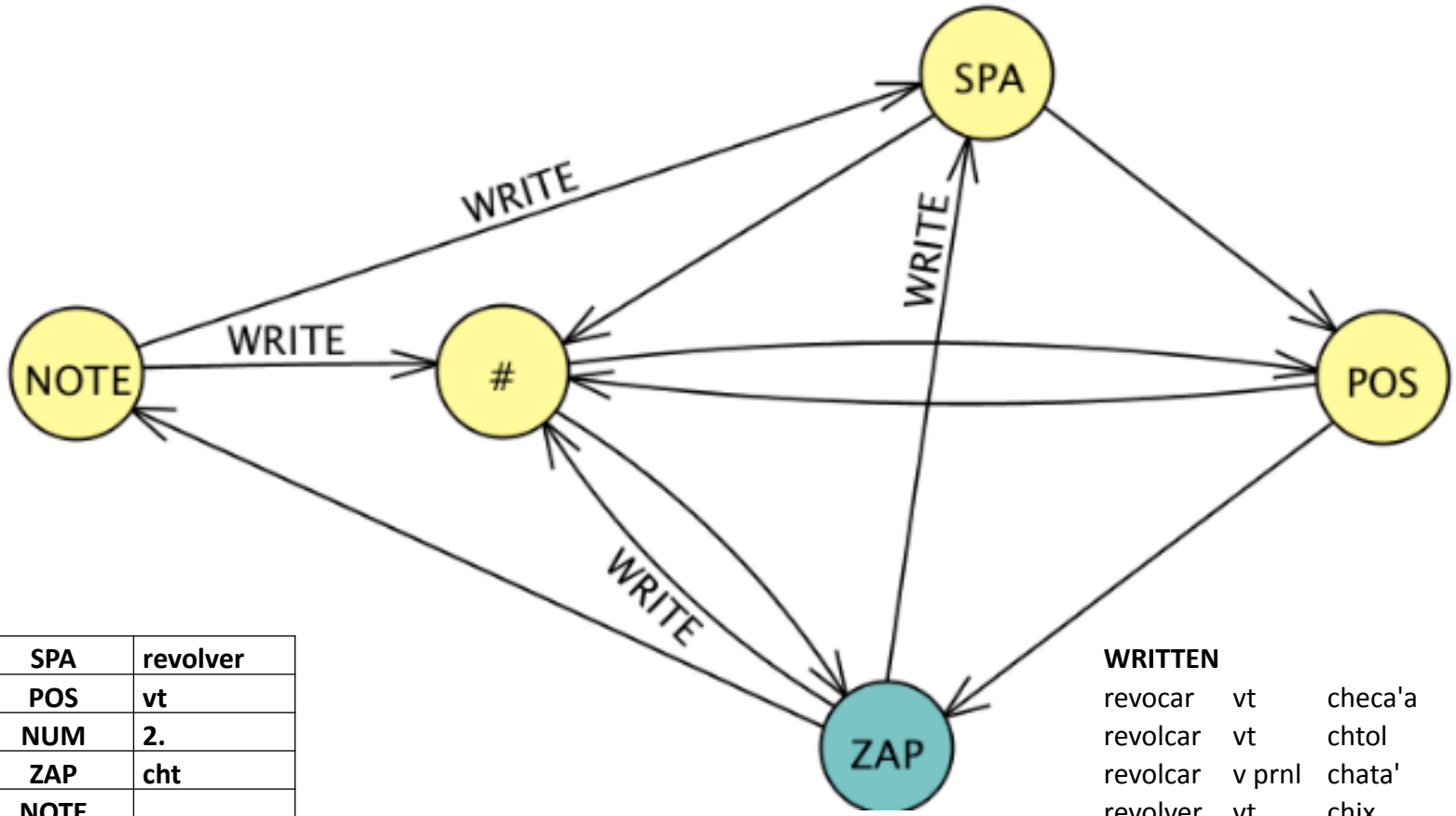


SPA	**revolver**
POS	**vt**
NUM	**2.**
ZAP	**ch**
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'
revolver	vt	chix

revolver *vt* 1. chix 2. cht̲a (líquido)
rey …

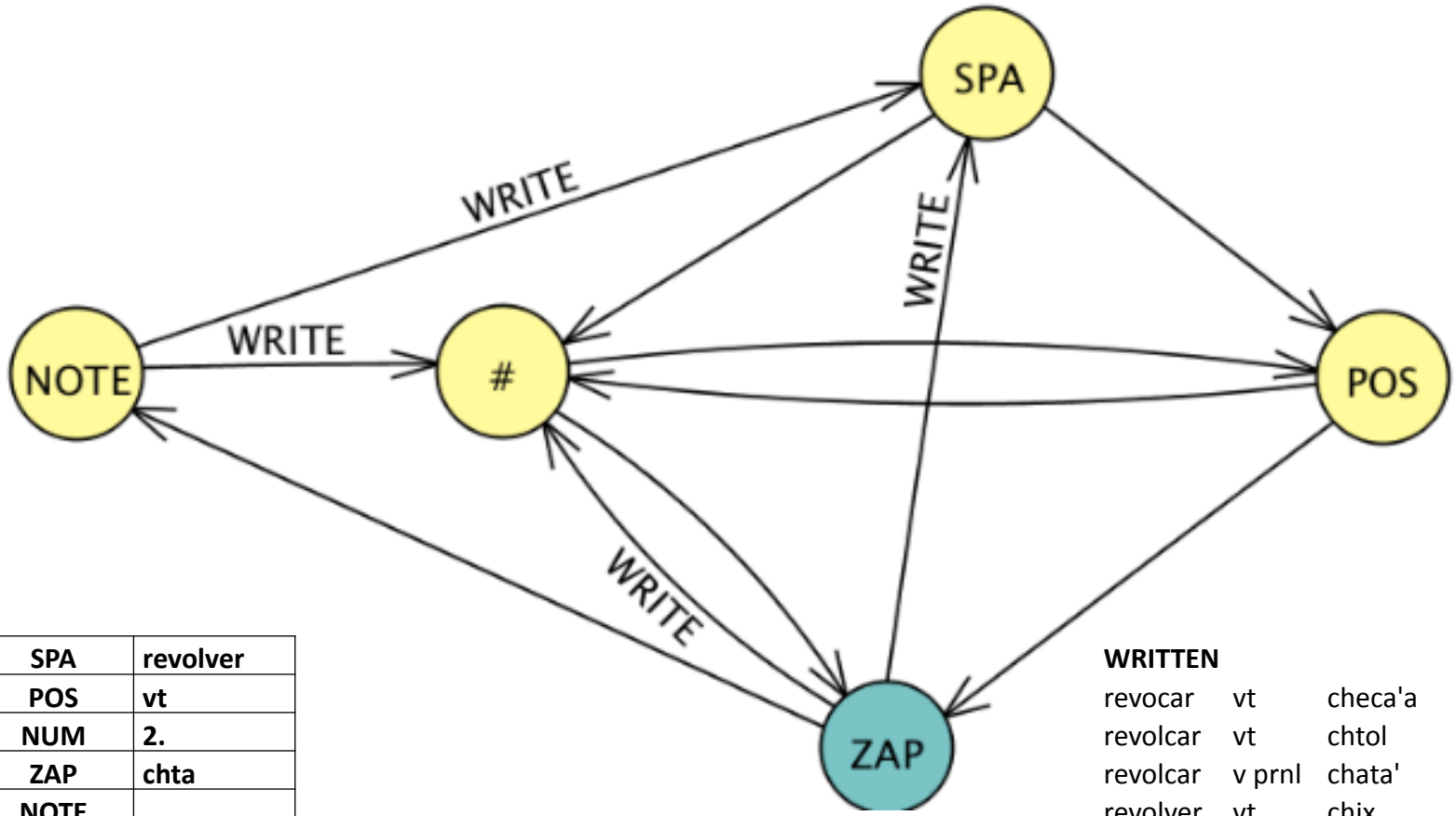


SPA	**revolver**
POS	**vt**
NUM	**2.**
ZAP	**cht**
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'
revolver	vt	chix

revolver *vt* 1. chix 2. chta (líquido) **rey** …

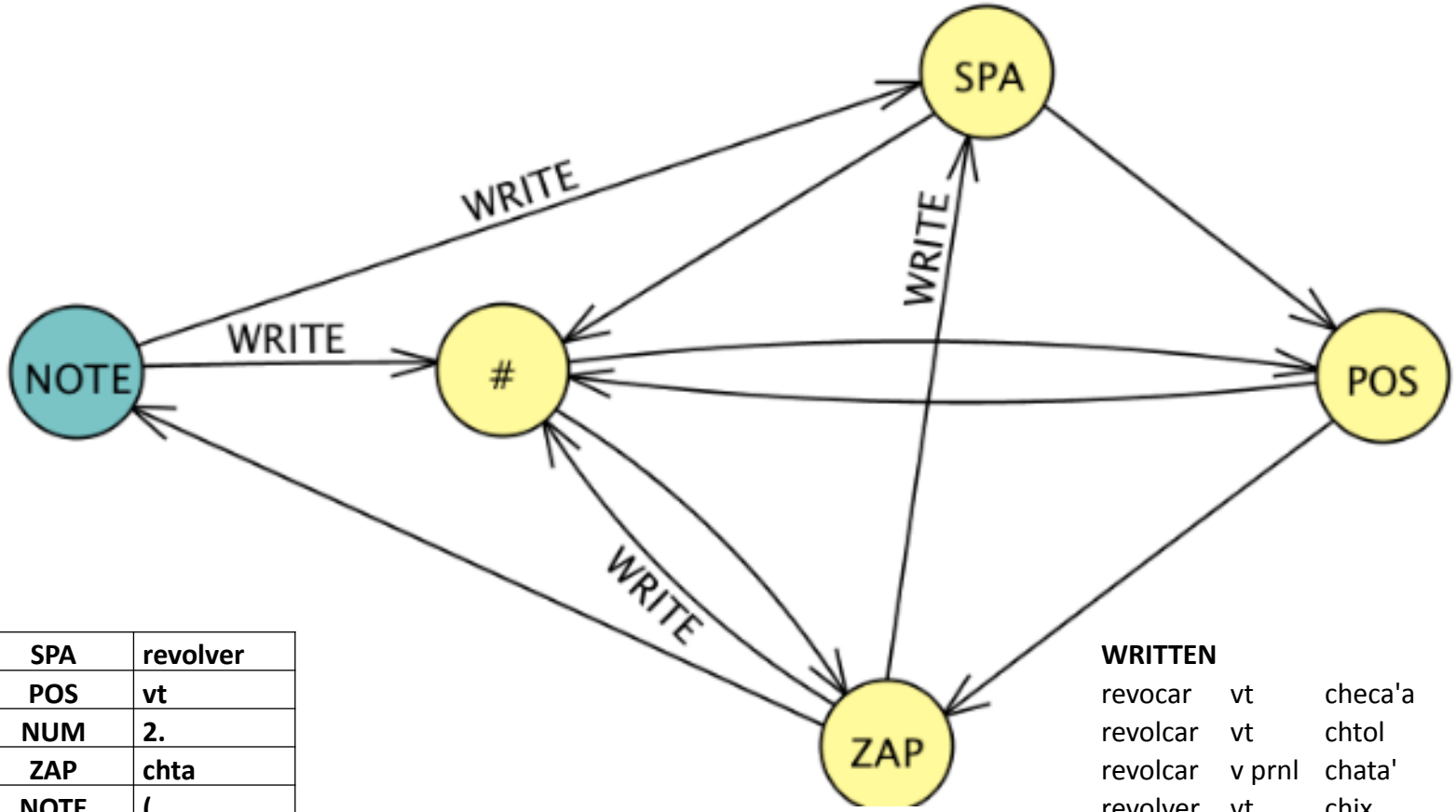


SPA	**revolver**
POS	**vt**
NUM	**2.**
ZAP	**chta**
NOTE	

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'
revolver	vt	chix

revolver *vt* 1. chix 2. chta (líquido) **rey** …

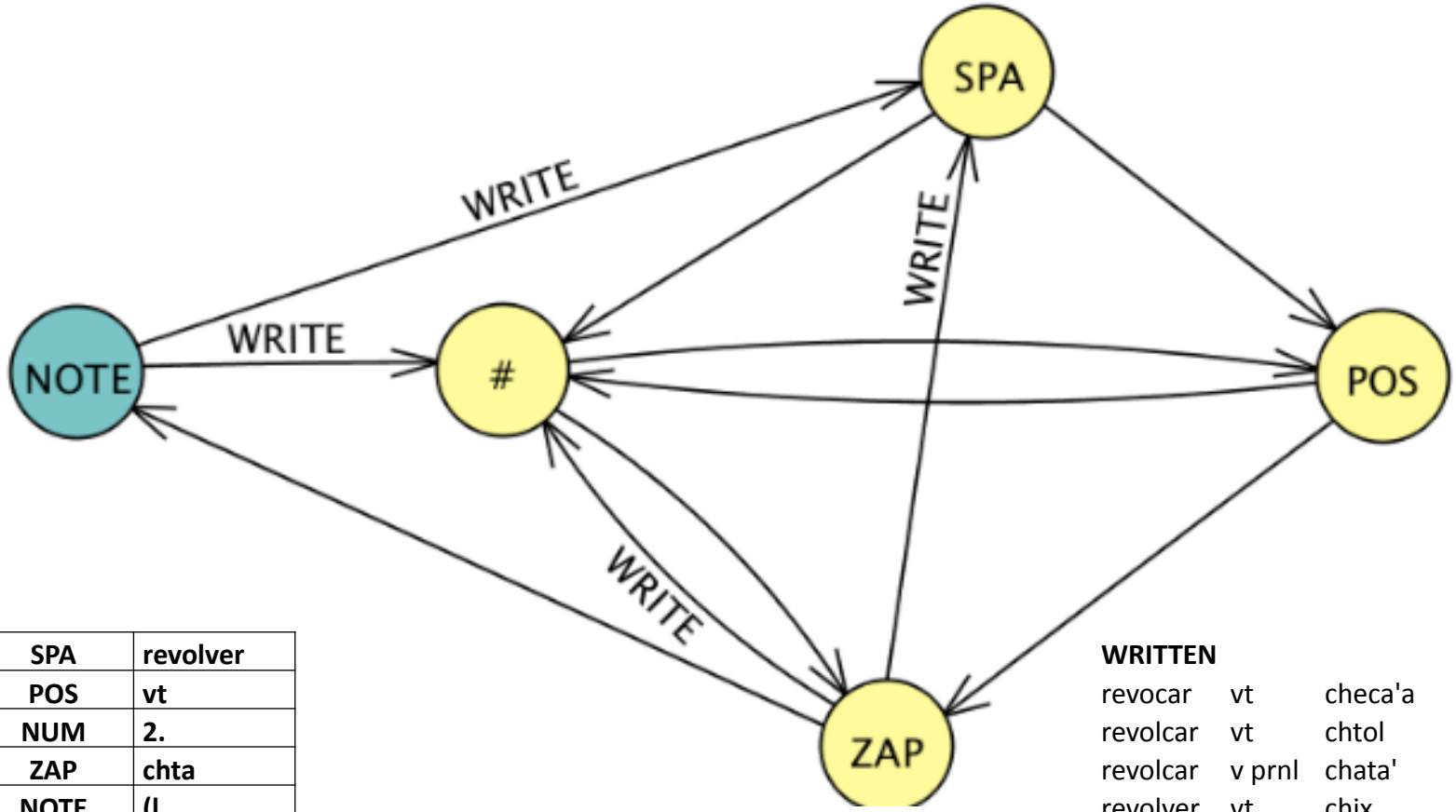


SPA	**revolver**
POS	**vt**
NUM	**2.**
ZAP	**chta**
NOTE	**(**

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'
revolver	vt	chix

revolver *vt* 1. chix 2. chta (líquido) **rey** …

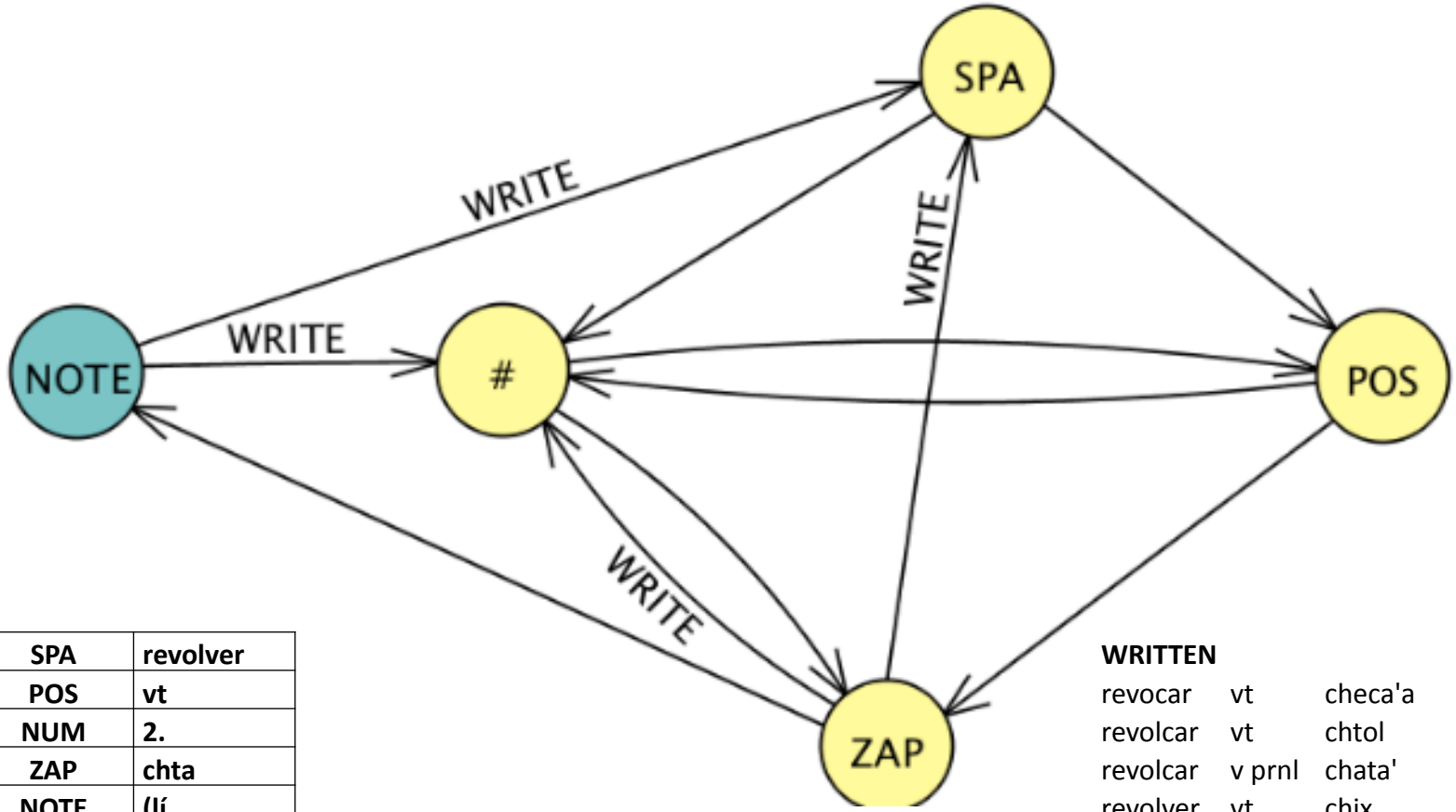


SPA	**revolver**
POS	**vt**
NUM	**2.**
ZAP	**chta**
NOTE	**(l**

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'
revolver	vt	chix

revolver *vt* 1. chix 2. chta (lí̱quido) **rey** …

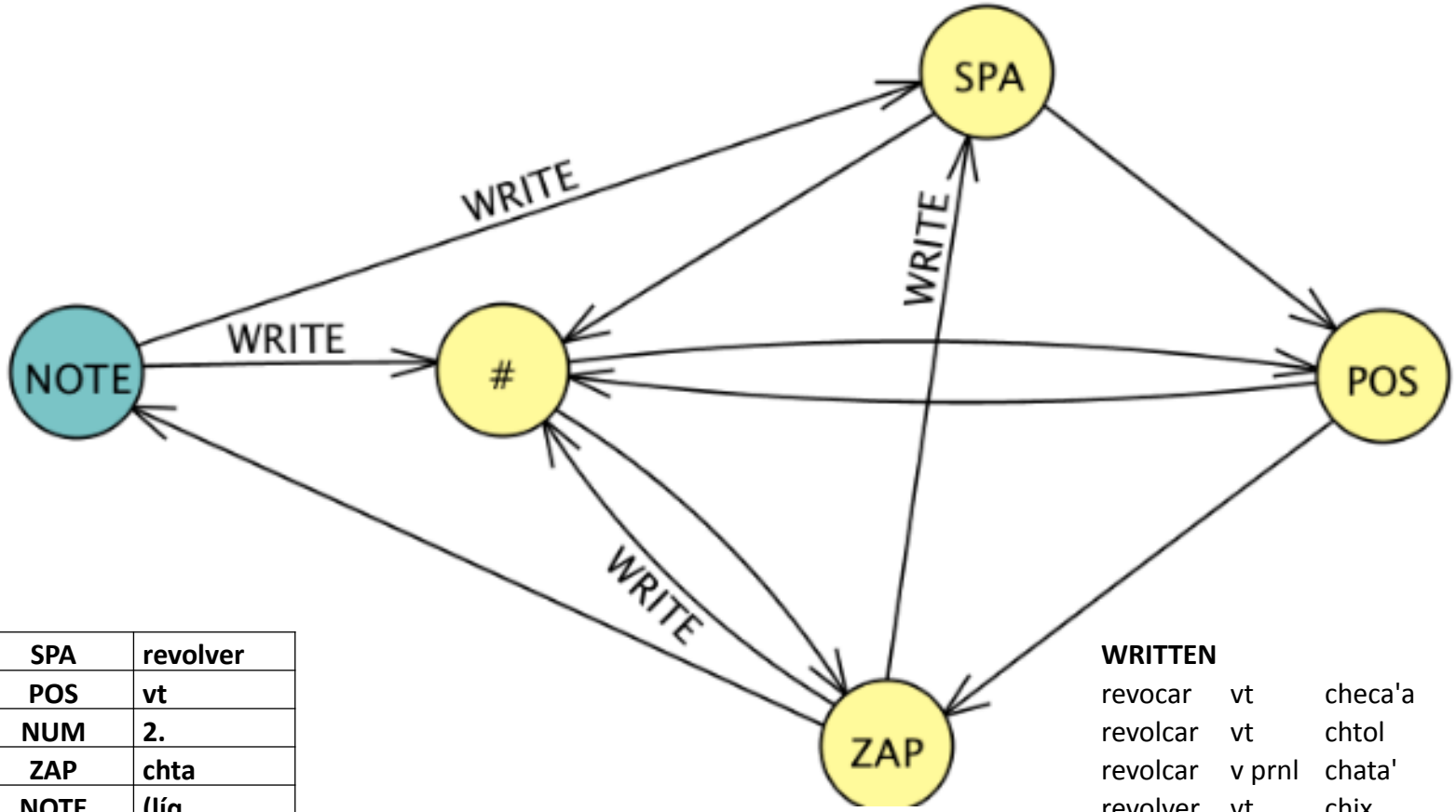


SPA	**revolver**
POS	**vt**
NUM	**2.**
ZAP	**chta**
NOTE	**(lí**

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'
revolver	vt	chix

revolver *vt* 1. chix 2. chta (líquido) **rey** …

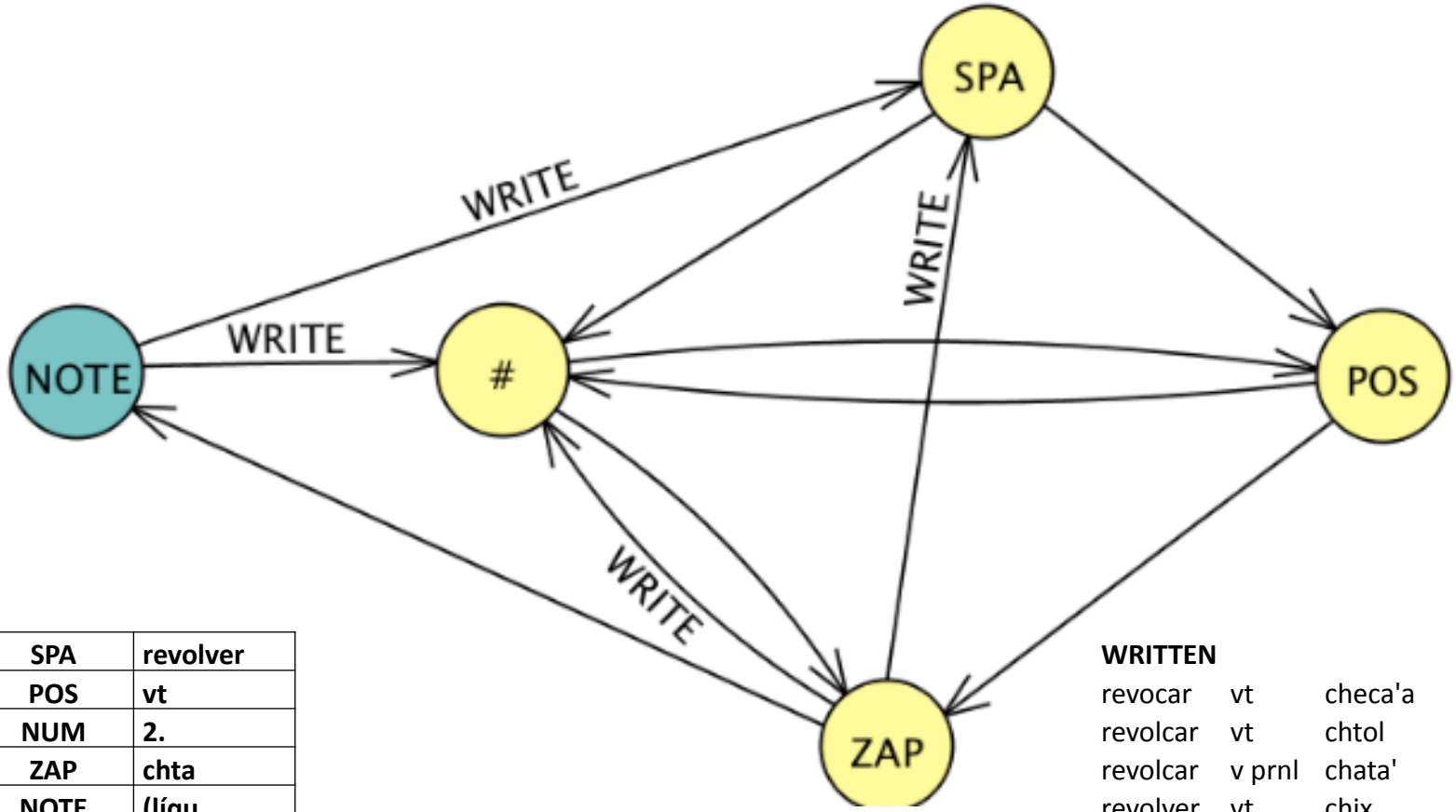


SPA	**revolver**
POS	**vt**
NUM	**2.**
ZAP	**chta**
NOTE	**(líq**

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'
revolver	vt	chix

revolver *vt* 1. chix 2. chta (líquido) **rey** …

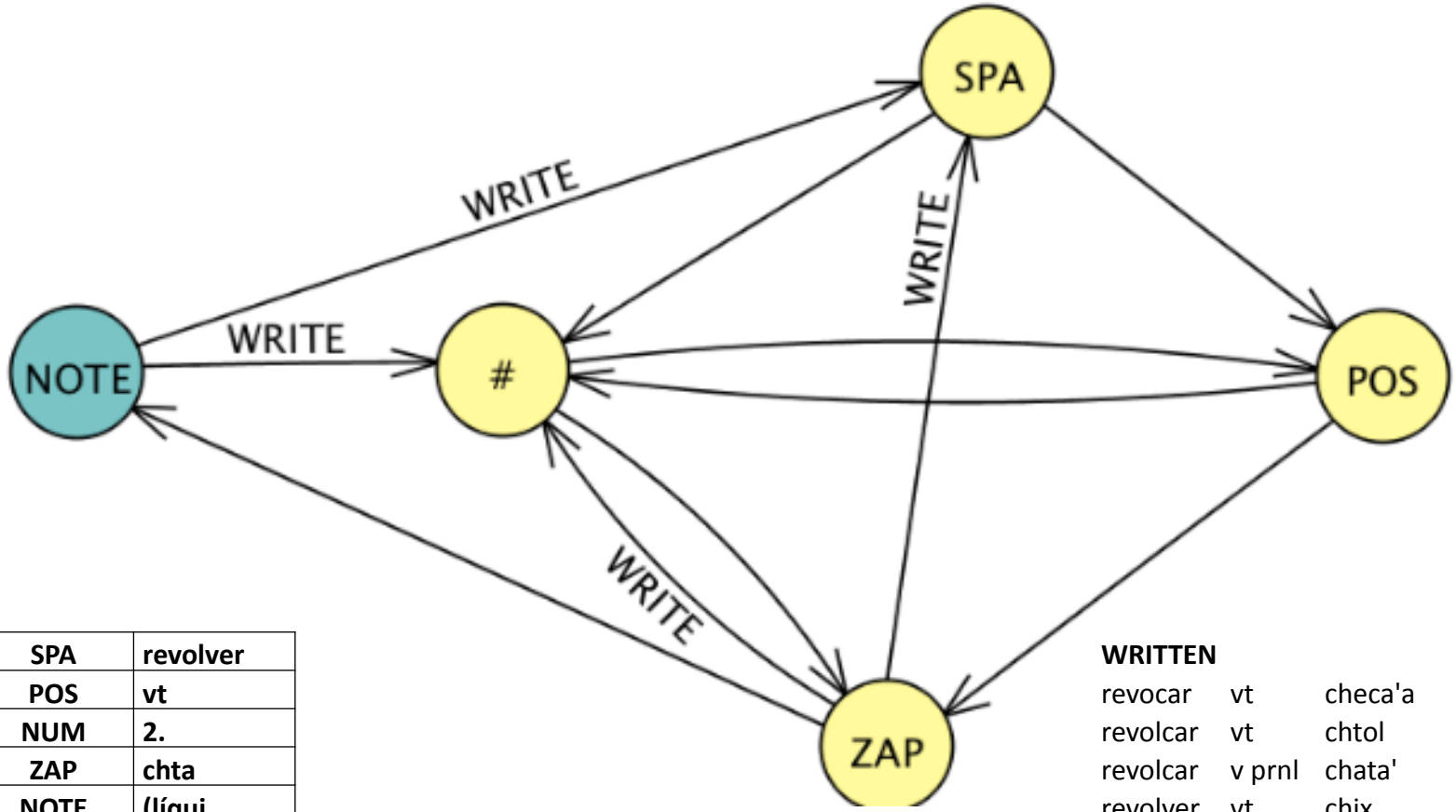


SPA	**revolver**
POS	**vt**
NUM	**2.**
ZAP	**chta**
NOTE	**(líqu**

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'
revolver	vt	chix

revolver *vt* 1. chix 2. chta (líquido) **rey** …

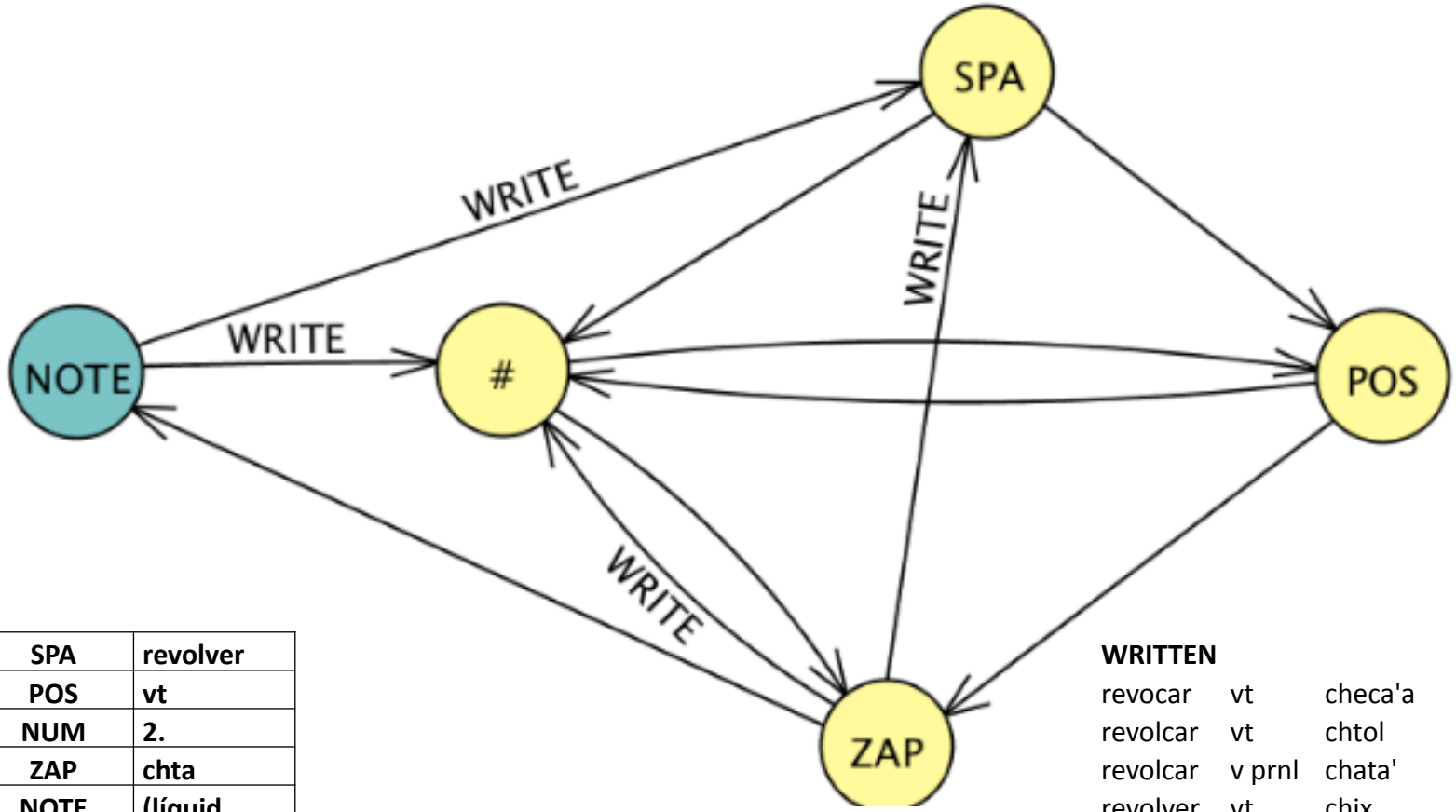


SPA	**revolver**
POS	**vt**
NUM	**2.**
ZAP	**chta**
NOTE	**(líqui**

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'
revolver	vt	chix

revolver *vt* 1. chix 2. chta (líquido)
rey …

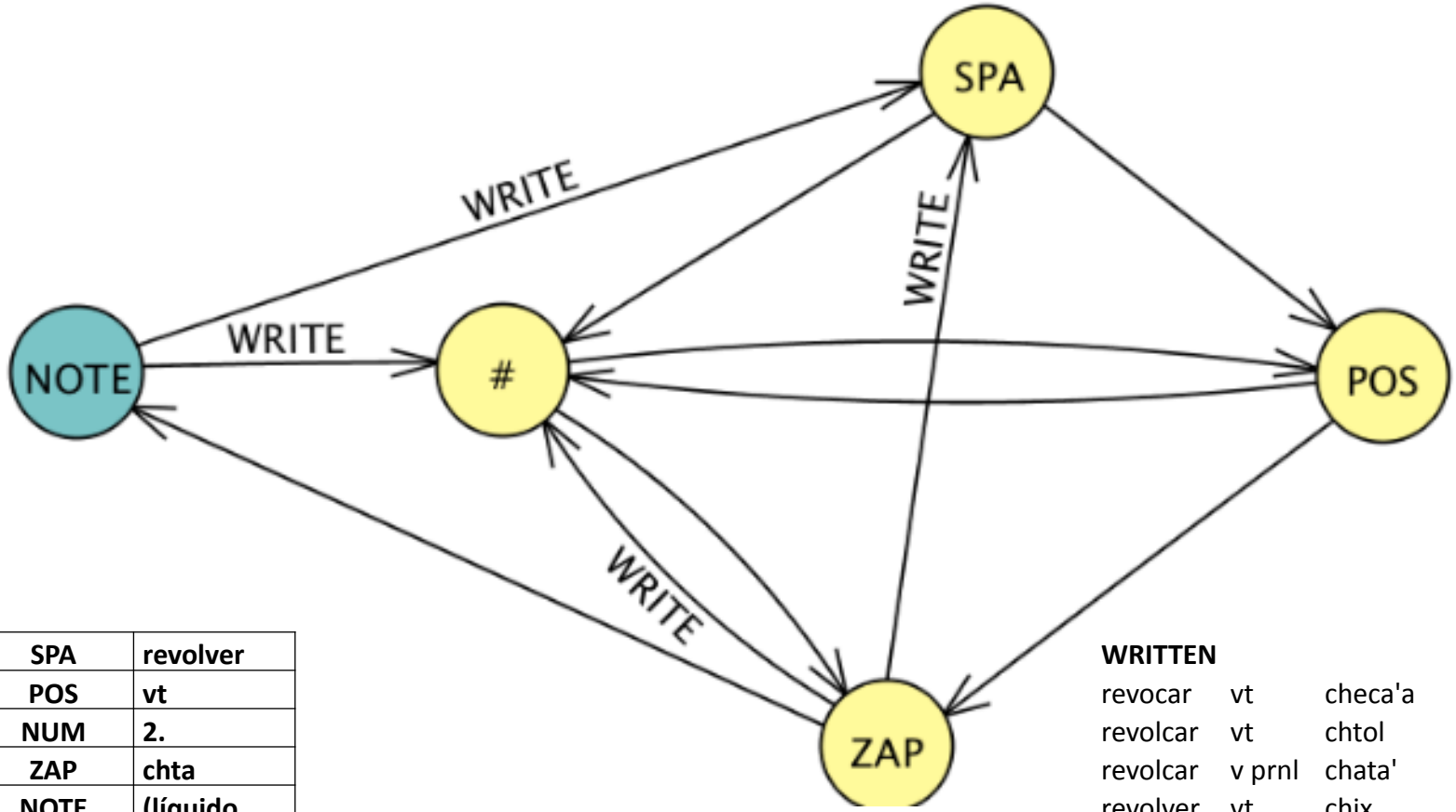


SPA	**revolver**
POS	**vt**
NUM	**2.**
ZAP	**chta**
NOTE	**(líquid**

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'
revolver	vt	chix

revolver *vt* 1. chix 2. chta (líquido) **rey** …

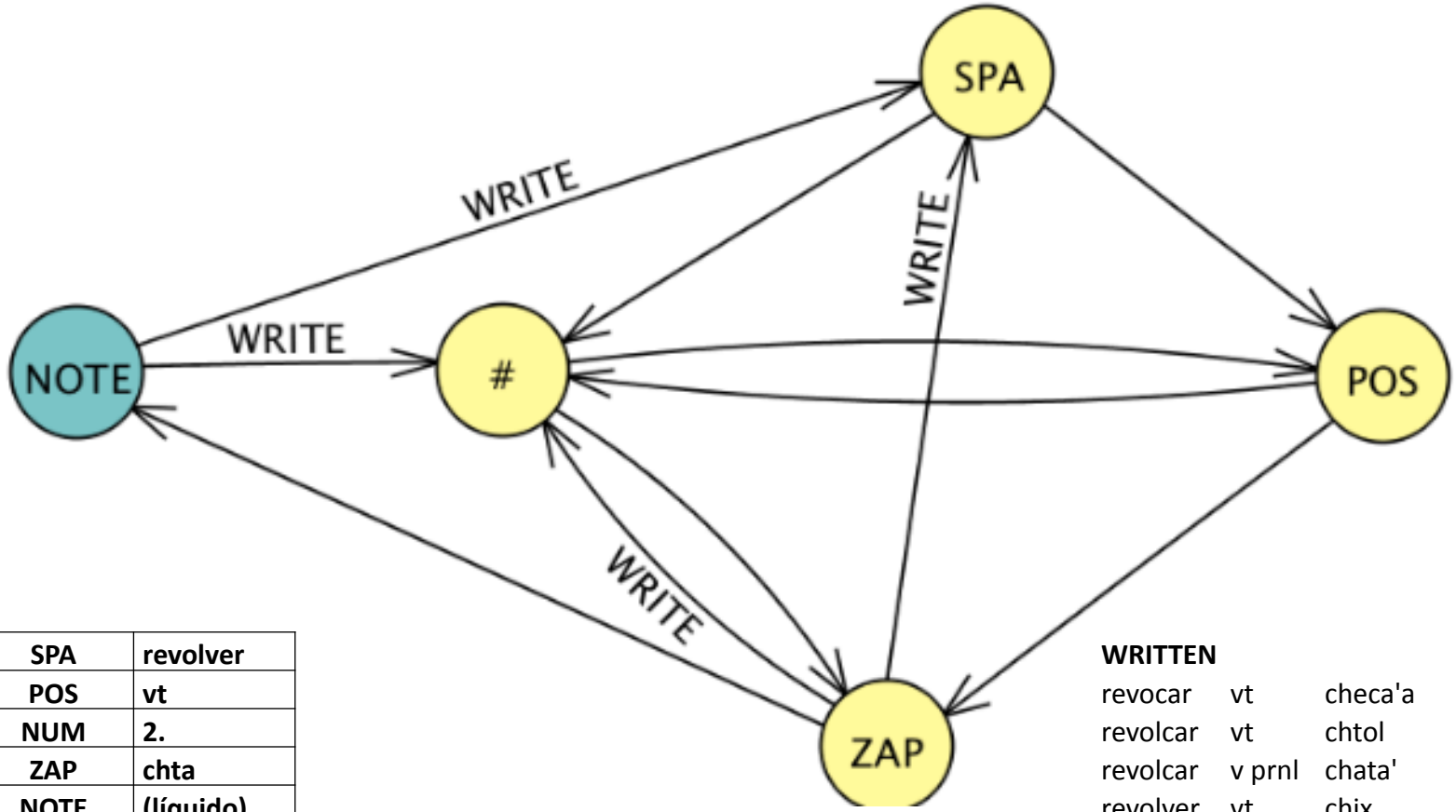


SPA	**revolver**
POS	**vt**
NUM	**2.**
ZAP	**chta**
NOTE	**(líquido**

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'
revolver	vt	chix

revolver *vt* 1. chix 2. chta (líquido) **rey** …



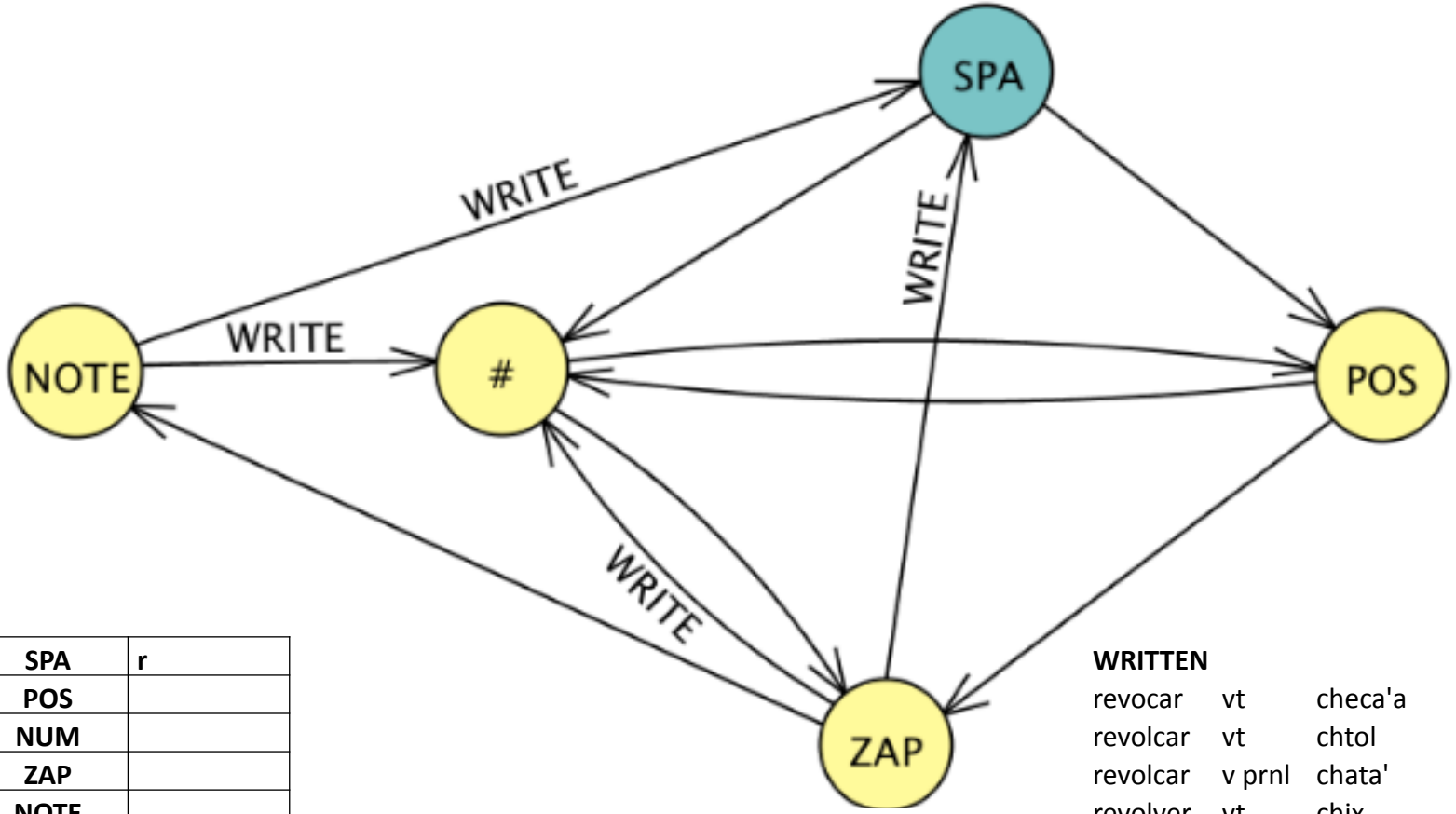
SPA	**revolver**
POS	**vt**
NUM	**2.**
ZAP	**chta**
NOTE	**(líquido)**

WRITTEN

revocar	vt	checa'a
revolcar	vt	chtol
revolcar	v prnl	chata'
revolver	vt	chix

revolver *vt* 1. chix 2. chta (líquido)

rey …



SPA	**r**
POS	
NUM	
ZAP	
NOTE	

WRITTEN

revocar	vt	checa'a	
revolcar	vt	chtol	
revolcar	v prnl	chata'	
revolver	vt	chix	
revolver	vt	chta	(líquido)

Methodology: A two-step tabularization process

- Use pdf2txt.py to create xml dump of pdf file
- Use xml-0to1.py (a Python script using lxml library) to:
  - parse the xml file (by specifying the transitions between the states)
  - fix any problems with character encoding
  - create a provisional tabularized output file
- Use xml-1to2.py to:
  - make any necessary 'manual' corrections
  - deal with entries with multiple word classes
  - other regex fixes to parenthesize punctuation, non-lemmatic forms
  - replace commas with synonym delimiters, etc
  - create a final tabularized output file
- Use the usual main.pl Perl script for serialization

Advantages

- Super fast and easy to set up
- Avoids having to grapple with complex regular expressions
- Can deal with cases that regular expression matches can't:

roncha *f* bloc, btoc
salir ronchas chtoc

- Can deal easily with words with more than one part-of-speech:

borracho *m, adj* borrašw
estar borracho *vi* chzože

Potential Pitfalls

- **Issue**: No unique font/size for each category
  Fix: Incorporate some simple regex syntax into conditioning criteria

- **Issue**: Dictionary headwords, page numbers are in the same typeface as the 'real' data
  Fix: Incorporate information from xml file about the positioning on the page of the main body of data (i.e. max and min height of text)

ropa *f* lache¹
rosa *f* 1. yej ros (*flor*)
2. color de ros
rosario *m* rsar

saciar *vt*
saciarse ²chaḷe¹, chelje, cheljenže¹
saco *m* sacw
sacudir 1. *vt* chbibe, chsese¹,

Potential Pitfalls

- **Issue**: Typeface for whitespace characters can mess up transitions
  Fix: Ignore all whitespace characters unless they are in same typeface as the preceding character

- **Issue**: The pdf document contains idiosyncratic typesetting errors (e.g. some parentheses are in regular font, others in italics)
  Fix: Can maybe correct consistent errors (such as parentheses), otherwise have to make manual fixes in second stage of tabularization

lavar *(algo)* *v*	rṵˈr´
lechuza *f*	güɨtʉʉ¯
leer *v*	ɨr´

sourcename-0.xml


```
<text font="LongZapSILDoulosL,Bold" bbox="132.117,369.738,137.560,383.538" size="13.800">v</text>
<text font="LongZapSILDoulosL,Bold" bbox="137.560,369.738,143.004,383.538" size="13.800">o</text>
<text font="LongZapSILDoulosL,Bold" bbox="143.004,369.738,146.118,383.538" size="13.800">l</text>
<text font="LongZapSILDoulosL,Bold" bbox="146.118,369.738,150.977,383.538" size="13.800">c</text>
<text font="LongZapSILDoulosL,Bold" bbox="150.977,369.738,156.420,383.538" size="13.800">a</text>
<text font="LongZapSILDoulosL,Bold" bbox="156.420,369.738,161.279,383.538" size="13.800">r</text>
<text font="LongZapSILDoulosL" bbox="161.280,369.658,164.304,382.560" size="12.902"> </text>
<text font="LongZapSILDoulosL" bbox="164.400,369.658,167.424,382.560" size="12.902"> </text>
<text font="LongZapSILDoulosL" bbox="167.520,369.658,170.544,382.560" size="12.902"> </text>
<text font="LongZapSILDoulosL" bbox="170.640,369.658,176.184,382.560" size="12.902">1</text>
<text font="LongZapSILDoulosL" bbox="176.150,369.658,179.074,382.560" size="12.902">.</text>
<text font="LongZapSILDoulosL" bbox="179.040,369.658,182.064,382.560" size="12.902"> </text>
<text font="LongZapSILDoulosL,Italic" bbox="182.160,369.392,186.647,381.111" size="11.719">v</text>
<text font="LongZapSILDoulosL,Italic" bbox="186.669,369.392,189.578,381.111" size="11.719">t</text>
<text font="LongZapSILDoulosL" bbox="189.600,369.446,192.336,381.120" size="11.674"> </text>
<text font="LongZapSILDoulosL" bbox="192.240,369.446,194.976,381.120" size="11.674"> </text>
<text font="LongZapSILDoulosL" bbox="194.880,369.658,197.904,382.560" size="12.902"> </text>
<text font="LongZapSILDoulosL" bbox="198.000,369.658,202.959,382.560" size="12.902">c</text>
<text font="LongZapSILDoulosL" bbox="202.953,369.658,208.497,382.560" size="12.902">h</text>
<text font="LongZapSILDoulosL" bbox="208.491,369.658,213.451,382.560" size="12.902">c</text>
<text font="LongZapSILDoulosL" bbox="213.445,369.658,218.989,382.560" size="12.902">h</text>
<text font="LongZapSILDoulosL" bbox="218.983,369.658,222.198,382.560" size="12.902">i</text>
<text font="LongZapSILDoulosL" bbox="222.192,369.658,227.585,382.560" size="12.902">(cid:158)</text>
<text font="LongZapSILDoulosL" bbox="227.579,369.658,230.603,382.560" size="12.902"> </text>
<text font="LongZapSILDoulosL" bbox="230.699,369.658,235.658,382.560" size="12.902">c</text>
<text font="LongZapSILDoulosL" bbox="235.652,369.658,241.196,382.560" size="12.902">h</text>
<text font="LongZapSILDoulosL" bbox="241.190,369.658,244.406,382.560" size="12.902">t</text>
<text font="LongZapSILDoulosL" bbox="244.400,369.658,249.843,382.560" size="12.902">o</text>
<text font="LongZapSILDoulosL" bbox="249.837,369.658,253.053,382.560" size="12.902">(cid:153)</text>
```

xml-0to1.py

```
import re
import codecs
from lxml import etree
from xmlparsefunctions import *

source = 'spa-zpq-Long'

# Enter font and size for entries in first language
lang1font = 'LongZapSILDoulosL,Bold'
lang1size = '13.800'

# Enter font and size for entries in second language
lang2font = 'LongZapSILDoulosL'
lang2size = '12.902'

# Enter font and size for part of speech (word class) indicator
posfont = 'LongZapSILDoulosL,Italic'
possize = '11.719'

# Enter font and size for numerical meaning delimiters
numdelfont = 'LongZapSILDoulosL'
numdelsize = '12.902'

# Enter font and size for parenthetical information
parnotefont = 'LongZapSILDoulosL,Italic'
parnotesize = '11.719'

# Enter correct unicode codes for incorrectly encoded characters
chardict = {u'ß': "'",
            u'ç': u'\u02c6',
            u'à': u'a\u0301\u0331'}

# Enter max and min values of y2 in bounding box
y2min = 135
y2max = 680
```

sourcename-1.txt

File Path: ~/PanLex/Source/zpq-spa-zpq-Long/secondary/spa-zpq

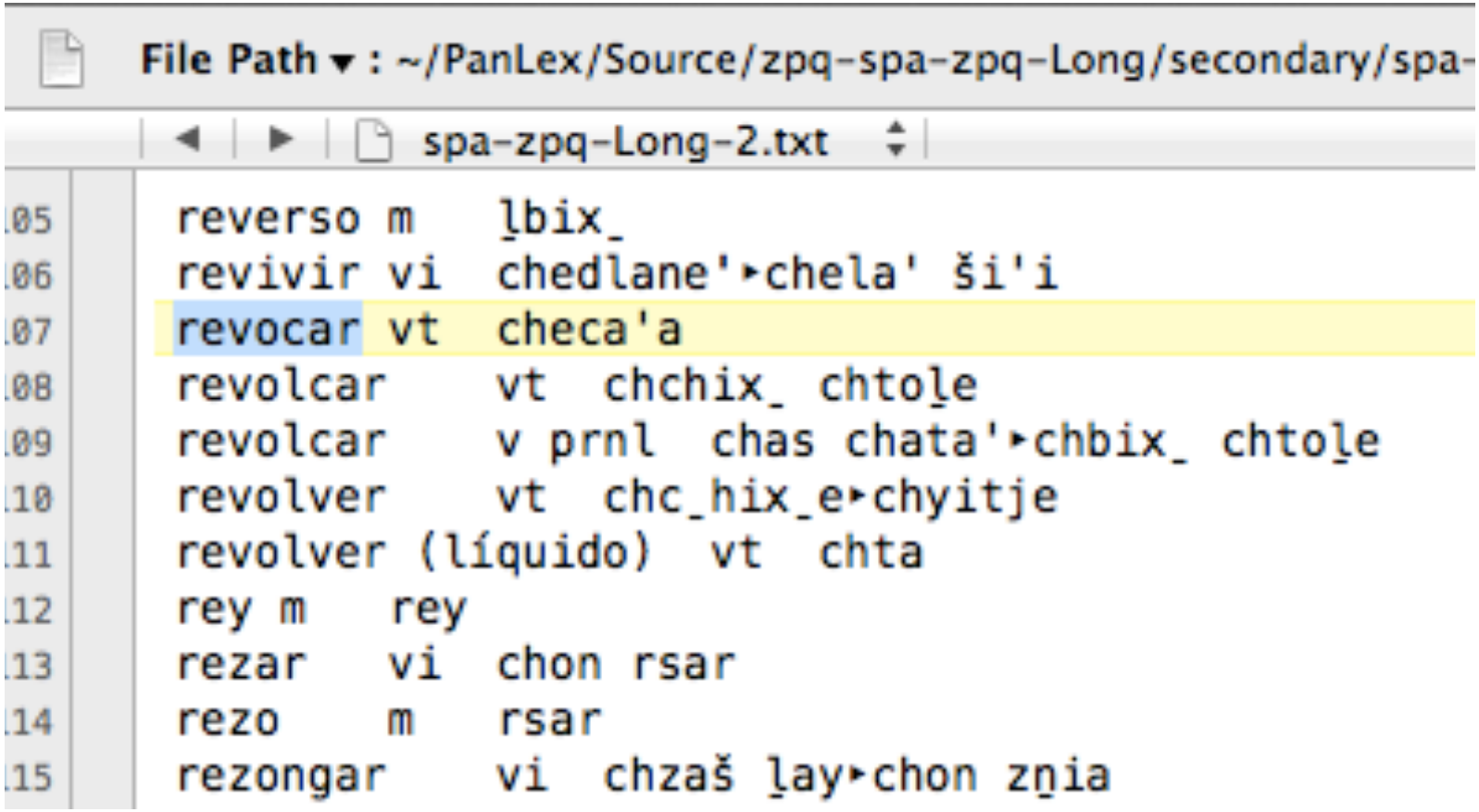
spa-zpq-Long-1.txt

```
reverso m    l̨bix_
revivir vi   chedlane', chela' ši'i
revocar vt   checa'a
revolcar     vt  chchix_ chtol̨e
revolcar     v prnl  chas chata', chbix_ chtol̨e
revolver     vt  chc_hix_e, chyitje
revolver     vt  chta    (líquido)
rey m    rey
rezar    vi  chon rsar
rezo     m   rsar
rezongar     vi  chzaš l̨ay, chon zn̨ia
```

xml-1to2.py

```
        note =      + note

## Hack some fixes for errors in -1.txt file (owing to incorrect typesetting in original pdf file)

    if lang1 == 'echuza':
        lang1 = 'lechuza'

    if lang2 == u"ry'r´ l":
        lang2 = u"ry'r´"


## Create a frequency table for the parts of speech attested in the source

    poslist.append(pos[0])

## Put punctuation marks for exclamatives and interrogatives inside parentheses

    lang1 = re.sub('!', '(!)', lang1)
    lang1 = re.sub(u'ǃ', u'(ǃ)', lang1)
    lang1 = re.sub('\?', '(?)', lang1)
    lang1 = re.sub(u'\¿', u'(¿)', lang1)

## Adjust for synonym delimiters in lang1 and lang2 expressions

    lang1 = re.sub(', ?',u'‣',lang1)
    lang2 = re.sub(', ?',u'‣',lang2)

## Common additional adjustments for expressions in Spanish

    lang1 = re.sub(r', ?-as?', '', lang1)              # removes the feminine forms
    lang1 = re.sub(r'^ser ', '(ser) ', lang1)
    lang1 = re.sub(r'^estar ', '(estar) ', lang1)
    lang1 = re.sub(r'^un ', '(un) ', lang1)

## Output to final text file

    f.write("\t".join([lang1 + note, pos[0], lang2]) + "\n")

    if len(pos) > 1:
        pos[1] = pos[1].strip()
```

sourcename-2.txt



Outstanding issues

- The basic Python scripts will get tabularizatin mostly correct, but there remains the time-consuming and laborious process of checking that no unexpected glitches have arisen
- Good place to start: the part-of-speech data in tabularization file
- xml-1to2.py will generate a frequency table of parts-of-speech in the final tabularization file
  - can check this against original pdf to see whether L1 or L2 expressions have crept in, or whether part-of-speech has been omitted
  - fine-tune script accordingly to catch this, or make manual fixes
- xml-0to1.py generates a list of all characters included in the xml to facilitate the re-encoding step. Would be nice if this process could be automated further