eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

Automatic scansion of poetry

Manex Agirrezabal Zabaleta

PhD dissertation

Dept. of Computer and Language Systems
University of the Basque Country (UPV / EHU)

Supervisors: Iñaki Alegria, Mans Hulden

June 19, 2017

O Captain! my Captain! our fearful trip is done,
The ship has weather'd every rack, the prize we sought is won,
The port is near, the bells I hear, the people all exulting,
While follow eyes the steady keel, the vessel grim and daring;
But O heart! heart! heart!
O the bleeding drops of red,
Where on the deck my Captain lies,
Fallen cold and dead.

...

Oh Captain! My Captain!
Walt Whitman

O Captain! my Captain! our fearful trip is done,
The ship has weather'd every rack, the prize we sought is won,
***The port is near**, the bells I hear, the people all exulting,*
While follow eyes the steady keel, the vessel grim and daring;
But O heart! heart! heart!
O the bleeding drops of red,
Where on the deck my Captain lies,
Fallen cold and dead.

...

Oh Captain! My Captain!
Walt Whitman

O Captain! my Captain! our fearful trip is done,
The ship has weather'd every rack, the prize we sought is won,
***The port is near**, **the bells I hear**, the people all exulting,*
While follow eyes the steady keel, the vessel grim and daring;
But O heart! heart! heart!
O the bleeding drops of red,
Where on the deck my Captain lies,
Fallen cold and dead.

...

Oh Captain! My Captain!
Walt Whitman

They said this day would never come
They said our sights were set too high
...

They said this day would never come
They said our sights were set too high
...

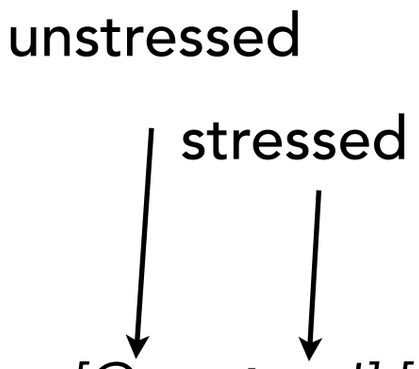
US election (2008)
Speech at Iowa Caucus
Barack Obama

One, two! One, two! And through and through
The vorpal blade went snicker-snack!
He left it dead, and with its head
He went galumphing back.

Jabberwocky
Lewis Carroll

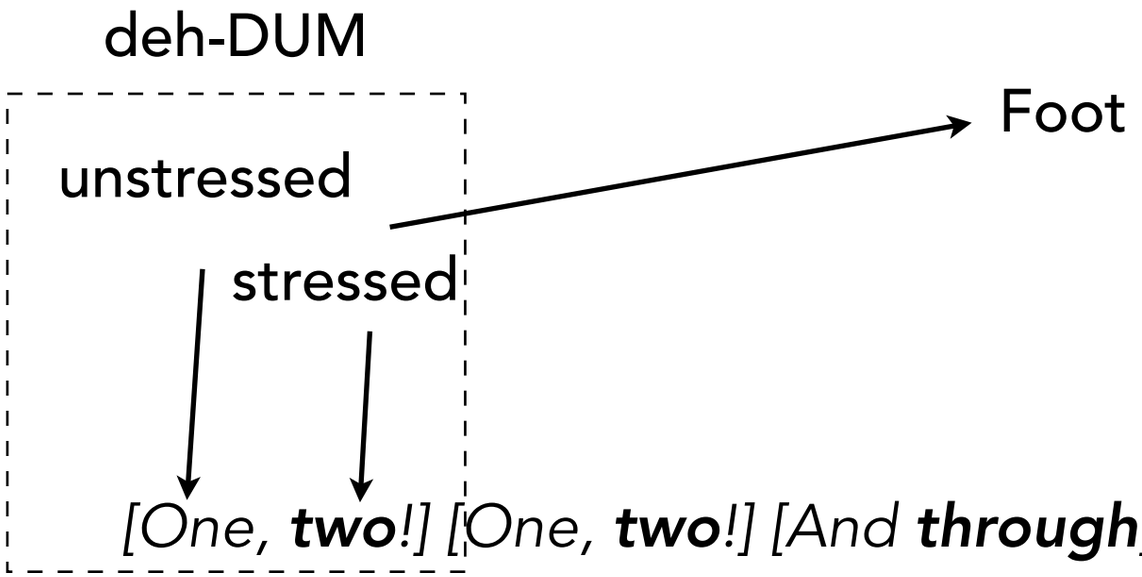
*[One, **two**!] [One, **two**!] [And **through**] [and **through**]*
*[The **vor**][pal **blade**] [went **snick**][er-**<u>snack</u>**!]*
*[He **left**] [it **dead**,] [and **with**] [its **head**]*
*[He **went**] [ga**lum**][phing **<u>back</u>**.]*

Jabberwocky
Lewis Carroll



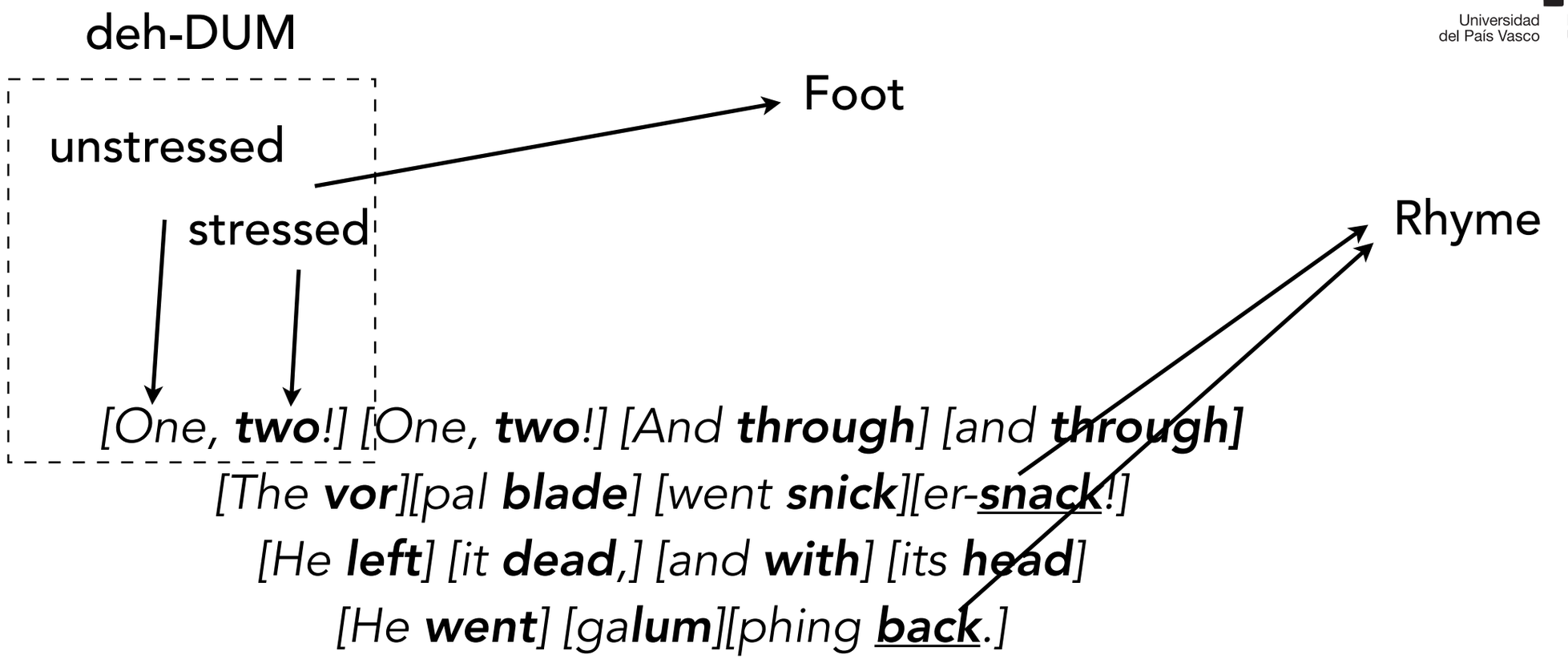
*[One, **two**!] [One, **two**!] [And **through**] [and **through**]*
*[The **vor**][pal **blade**] [went **snick**][er-**<u>snack</u>**!]*
*[He **left**] [it **dead**,] [and **with**] [its **head**]*
*[He **went**] [ga**lum**][phing **<u>back</u>**.]*

Jabberwocky
Lewis Carroll

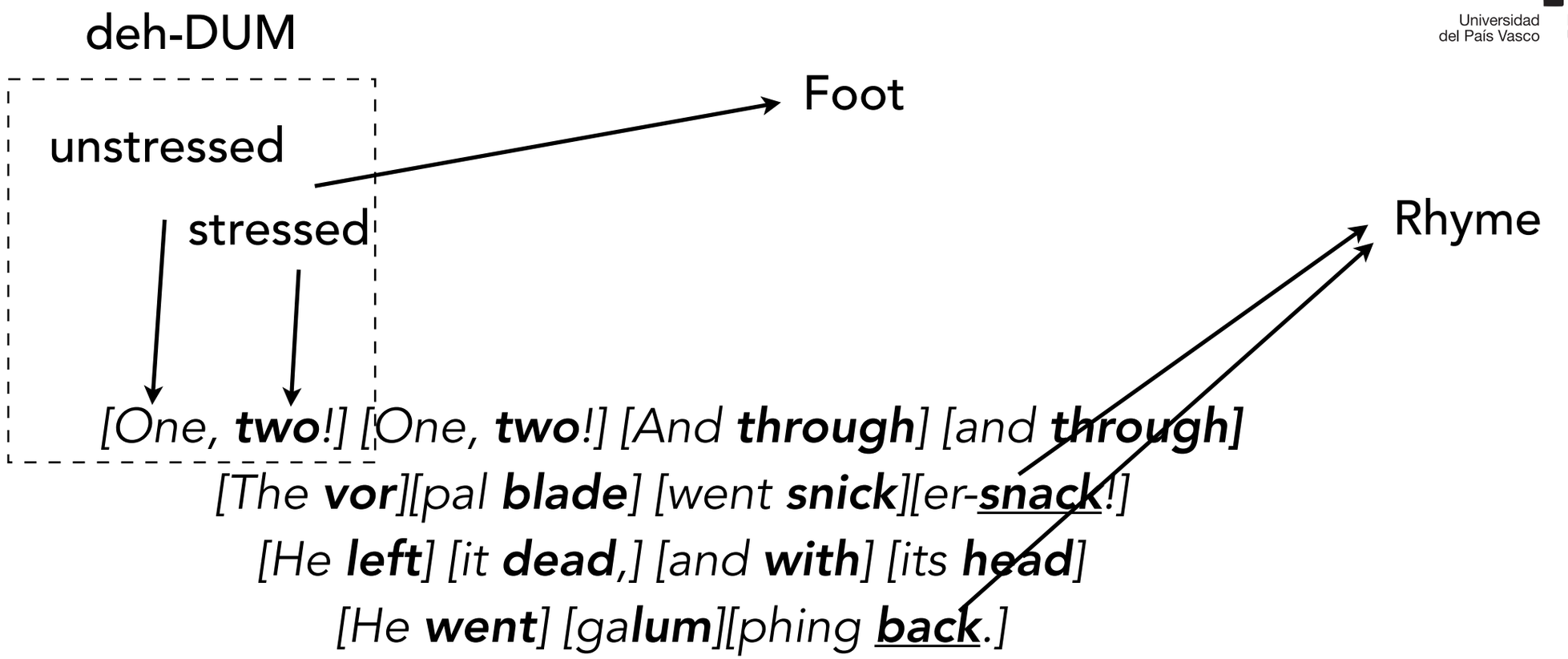


*[One, **two**!] [One, **two**!] [And **through**] [and **through**]*
*[The **vor**][pal **blade**] [went **snick**][er-**<u>snack</u>**!]*
*[He **left**] [it **dead**,] [and **with**] [its **head**]*
*[He **went**] [ga**lum**][phing **<u>back</u>**.]*

Jabberwocky
Lewis Carroll



Jabberwocky
Lewis Carroll



Jabberwocky
Lewis Carroll

Scansion involves marking all this information,
but in this work we mainly focus on the stress sequences

Uses of scansion systems

- Poetry Generation
- Authorship attribution
- Cataloging poems according to the meter
- Learn how to correctly recite a poem

Final goal: from marking stresses to finding structure in raw text

(1)

wo man	much	missed	how	you	call	to	me	call	to	me

Final goal: from marking stresses to finding structure in raw text

(1)

wo man	much	missed	how	you	call	to	me	call	to	me
/ x	/	\	/	/	/	x	/	/	x	/

Final goal: from marking stresses to finding structure in raw text

(1)

wo man		much	missed	how	you	call	to	me	call	to	me
/	x	/	\	/	/	/	x	/	/	x	/
/	x	x	/	x	x	/	x	x	/	x	x

Final goal: from marking stresses to finding structure in raw text

(1)

wo man	much	missed	how	you	call	to	me	call	to	me
/ x	/	\	/	/	/	x	/	/	x	/
/ x	x	/	x	x	/	x	x	/	x	x

Final goal: from marking stresses to finding structure in raw text

(1)

wo man		much	missed		how	you	call		to	me	call		to	me
/	x	/	\		/	/	/		x	/	/		x	/
/	x	x	/		x	x	/		x	x	/		x	x

(2)

al mas			di cho sas			que	del	mor tal			ve lo		
/	x	x	/		x	x	x	x	/		/	x	
/	x	x	/		x	x	x	x	/		/	x	

Final goal: from marking stresses to finding structure in raw text

(1)

wo man	much	missed	how	you	call	to	me	call	to	me
/ x	/	\	/	/	/	x	/	/	x	/
/	x x	/	x	x	/	x	x	/	x	x

(2)

al mas	di cho sas	que	del	mor tal	ve lo
/ x	x /	x	x	x /	/ x
/	x x	/	x	x x	/ x

(3)

Because I do not hope to know again
The infirm glory of the positive hour
Because I do not think
Because I know I shall not know
The one veritable transitory power
Because I cannot drink

Final goal: from marking stresses to finding structure in raw text

(1)

wo man	much	missed	how	you	call	to	me	call	to	me
/ x	/	\	/	/	/	x	/	/	x	/
/	x x	/	x	x	/	x	x	/	x	x

(2)

al mas	di cho sas	que	del	mor tal	ve lo
/ x	x /	x	x	x x	/ / x
/	x x	/	x	x x	/ / x

(3)

Because I do not hope to know again
The infirm glory of the positive hour
Because I do not think
Because I know I shall not know
The one veritable transitory power
Because I cannot drink

Final goal: from marking stresses to finding structure in raw text

(1)

wo man	much	missed	how	you	call	to	me	call	to	me
/ x	/	\	/	/	/	x	/	/	x	/
/	x x	/	x	x	/	x	x	/	x	x

(2)

al mas	di cho sas	que	del	mor tal	ve lo
/ x	x /	x	x	x x	/ / x
/	x x	/	x	x x	/ / x

(3)

Because I do not hope to know again
The infirm glory of the positive hour
Because I do not think
Because I know I shall not know
The one veritable transitory power
Because I cannot drink

Final goal: from marking stresses to finding structure in raw text

(1)

wo man	much	missed	how	you	call	to	me	call	to	me
/ x	/	\	/	/	/	x	/	/	x	/
/	x x	/	x	x	/	x	x	/	x	x

(2)

al mas		di cho sas			que	del	mor tal		ve lo	
/	x	x	/	x	x	x	x	/	/	x
/	x	x	/	x	x	x	x	/	/	x

(3)

Because I do not hope to know again
The infirm glory of the positive hour
Because I do not think
Because I know I shall not know
The one veritable transitory power
Because I cannot drink

Outline

- Research questions and Tasks
- Tradition of scansion
- Automatic scansion and Sequence modeling
- NLP techniques for scansion
- General results
- Discussion and Future work

Outline

- **Research questions and Tasks**
- Tradition of scansion
- Automatic scansion and Sequence modeling
- NLP techniques for scansion
- General results
- Discussion and Future work

Research questions

1. What do we need to know when analyzing a poem and how can we capture it?

2. Does language-specific linguistic knowledge contribute when analyzing poetry?

3. Is it possible to analyze a poem without any language-specific information? Is such analysis something that can be learnt?

Research questions

1. What do we need to know when analyzing a poem and how can we capture it?
2. Does language-specific linguistic knowledge contribute when analyzing poetry?
3. Is it possible to analyze a poem without any language-specific information? Is such analysis something that can be learnt?

Goal

To be able to correctly analyze poems in English and apply such knowledge to Spanish and Basque.

Tasks

- Develop a rule-based poetry scansion system for English
- Collect a corpus of scanned English poems to test the scansion system
- Train data-driven models using the English corpus. Use simple features and extended language-specific features to represent the poems
- Collect corpora in other languages and, when necessary, annotate them
- Extrapolate data-driven approaches to other available languages
- Try to infer poetic stress patterns directly from data without any labeled data

Outline

- Research questions and Tasks
- **Tradition of scansion**
- Automatic scansion and Sequence modeling
- NLP techniques for scansion
- General results
- Discussion and Future work

Scansion in English

- Accentual-syllabic poetry
  - Syllables
  - Stresses
- Repeating patterns of feet

Scansion in English

- Accentual-syllabic poetry
  - Syllables
  - Stresses
- Repeating patterns of feet

Iambic meter [x /]	Anapestic meter [x x /]
Come **live** with **me** and **be** my **love** And **we** will **all** the **plea**sures **prove**, That **val**leys, **groo**ves, **hills** and **fields**, **Woods**, or **steep**y **moun**tain **yields**.	and I **don't** like to **brag**, but I'm **tell**ing you **Liz** that **speak**ing of **cooks** I'm the **best** that there **is** why **on**ly last **Tues**day when **mo**ther was **out** I **real**ly cooked **some**thing worth **talk**ing a**bout**
Trochaic meter [/ x]	**Dactylic meter [/ x x]**
Can it **be** the **sun** des**cen**ding **O'er** the **le**vel **plain** of **wa**ter? **Or** the **Red** Swan **float**ing, **fly**ing, **Woun**ded **by** the **ma**gic **a**rrow,	**Wo**man much **missed**, how you **call** to me, **call** to me **Say**ing that **now** you are **not** as you **were** **When** you had **changed** from the **one** who was **all** to me, **But** as at **first**, when our **day** was **fair**.

Scansion in English

- Metrical variation

*Ad**mi**rer **as** I **think** I **am***
x / x / x / x /

*of **stars** that **do** not **give** a **damn**,*
x / x / x / x /

*I **can**not, **now** I **see** them, **say***
x / x / x / x /

*I **missed** one **ter**ribly **all day***
x / x / x x / /

The More Loving One
Wystan H. Auden

Scansion in English

- Metrical variation

*Ad**mi**rer **as** I **think** I **am***
x / x / x / x /

*of **stars** that **do** not **give** a **damn**,*
x / x / x / x /

*I **can**not, **now** I **see** them, **say***
x / x / x / x /

*I **missed** one **ter**ribly **all day***
x / x / x x / /

The More Loving One
Wystan H. Auden

Scansion in English

The Challenges of scansion:

1. Lexical stresses do not always apply
2. Dividing the stress pattern into feet
3. Dealing with Out-Of-Vocabulary words

Scansion in English

The Challenges of scansion:

1. **Lexical stresses do not always apply**
2. Dividing the stress pattern into feet
3. Dealing with Out-Of-Vocabulary words

LEXICAL STRESSES	
woman	/x
much	/
missed	\
how	/
you	/
call	/
to	x
me	/

wo	man	much	missed	how	you	call	to	me	call	to	me
/	x	/	\	/	/	/	x	/	/	x	/
/	x	x	/	x	x	/	x	x	/	x	x

Scansion in English

The Challenges of scansion:

1. Lexical stresses do not always apply
2. **Dividing the stress pattern into feet**
3. Dealing with Out-Of-Vocabulary words

Woman much missed how you call to me call to me

Scansion in English

The Challenges of scansion:

1. Lexical stresses do not always apply
2. **Dividing the stress pattern into feet**
3. Dealing with Out-Of-Vocabulary words

Woman much missed how you call to me call to me

[**Wo**man much] [**missed** how you] [**call** to me] [**call** to me]

Scansion in English

The Challenges of scansion:

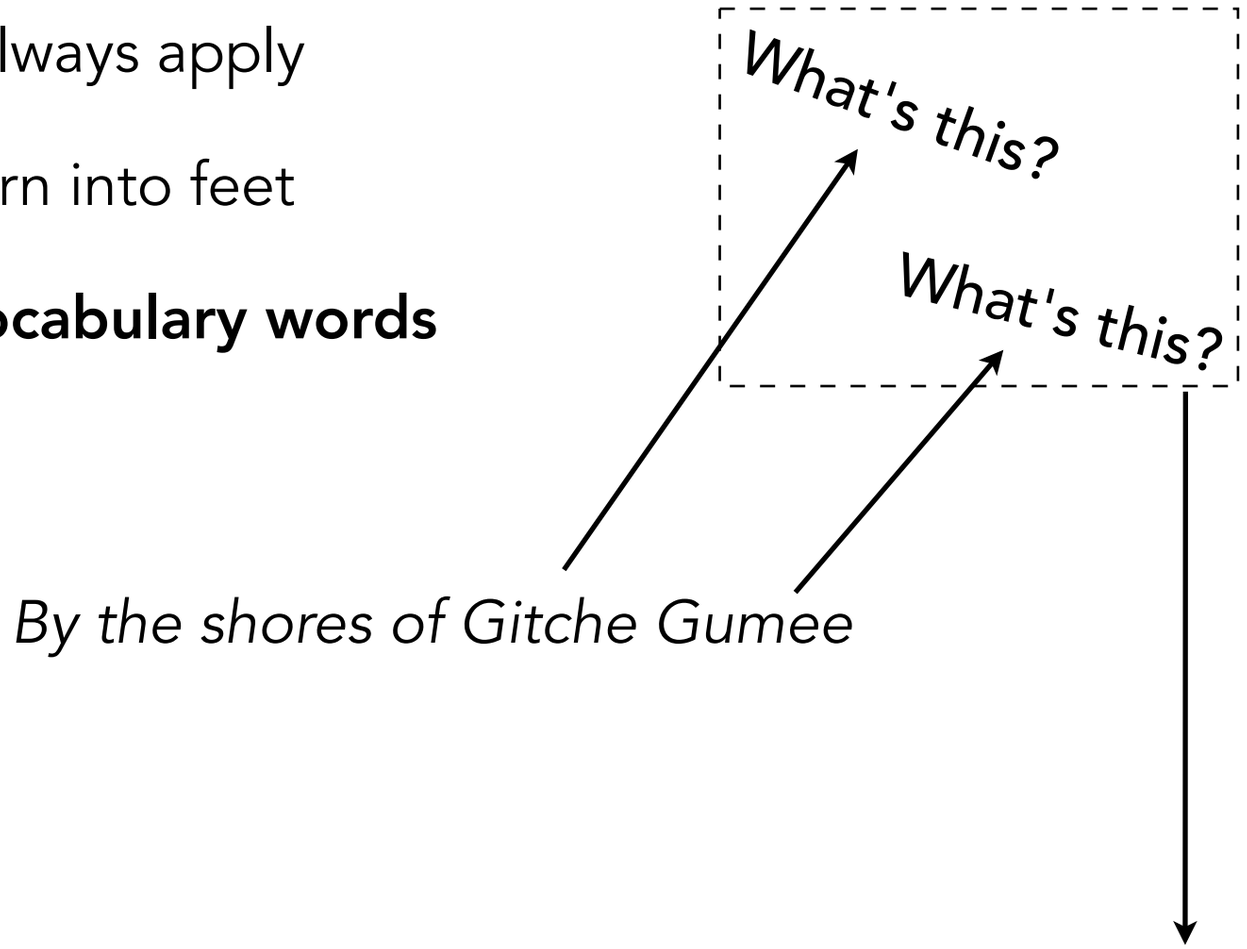
1. Lexical stresses do not always apply
2. Dividing the stress pattern into feet
3. **Dealing with Out-Of-Vocabulary words**

By the shores of Gitche Gumee

Scansion in English

The Challenges of scansion:

1. Lexical stresses do not always apply
2. Dividing the stress pattern into feet
3. **Dealing with Out-Of-Vocabulary words**



If there is no entry in the dictionary,
we have to somehow calculate their lexical stress

Scansion in English

English poetry Corpus

- 79 poems from For Better For Verse (4B4V) (Tucker, 2011)
  - Brought by the Scholar's Lab at the University of Virginia
- Interactive website to train people on the scansion of traditional poetry
- Statistics

	English corpus
No. syllables	10,988
No. distinct syllables	2,283
No. words	8,802
No. distinct words	2,422
No. lines	1,093

Scansion in English

English poetry Corpus

Sonnet 18 (1609)

William Shakespeare

U / U / U / U / U /
Shall I |compare |thee to |a sum|mer's day?
U U / / U U / / U /
Thou art |more love|ly and |more tem|perate:
/ / U / U / U / U /
Rough winds |do shake |the dar|ling buds |of May,
U / U / U / / / U /
And sum|mer's lease |hath all |too short |a date;

/ U / / U / U / U /
Sometimes |too hot |the eye |of heav|en shines,
U / U / U / U / U /
And of|ten is |his gold |complex|ion dimmed;
U / U / U / / U U /
And eve|ry fair |from fair |sometimes |declines,
U / U / U / U / U /
By chance |or na|ture's chang|ing course |untrimmed;

U / U / U / U / U /

Scansion in Spanish

- Accentual-syllabic poetry
  - Syllables
  - Stresses

Scansion in Spanish

- Accentual-syllabic poetry
  - Syllables
  - Stresses

- Classification according to the Syllables
  - Minor art verses
  - Major art verses
  - Composite verses
- According to the stresses
  - Last syllable stress (Oxytone verses)
  - Penultimate syllable stress (Paroxytone verses)
  - Antepenultimate syllable stress (Proparoxytone verses)

In this work we have focused on the Spanish Golden Age

The most common meter was the hendecasyllable.

Scansion in Spanish

- Accentual-syllabic poetry
  - Syllables
  - Stresses

__Fe__ria des__pués__ que del ar__nés__ do__ra__do
y la __to__ga pa__cí__fica des__nu__do
col__gó__ la es__pa__da y el lu__cien__te es__cu__do;
obede__cien__do a __Jú__piter sa__gra__do,

...

A los casamientos del Excelentísimo Duque de Feria
Lope de Vega

Scansion in Spanish

The challenge:

- Syllable contractions / Synaloephas

Cual suele la luna tras lóbrega nube
con franjas de plata bordarla en redor,
y luego si el viento la agita, la sube
disuelta a los aires en blanco vapor:
...

El estudiante de Salamanca
José de Espronceda

Scansion in Spanish

The challenge:

- Syllable contractions / Synaloephas

*Cual **sue**le la **lu**na tras **ló**brega **nu**be*
*con **fran**jas de **pla**ta bor**dar**la en re**dor**,*
*y **lue**go si el **vien**to la a**gi**ta, la **su**be*
*di**suel**ta a los **ai**res en **blan**co va**por**:*
...

El estudiante de Salamanca
José de Espronceda

Scansion in Spanish

The challenge:

- Syllable contractions / Synaloephas

*Cual **sue**le la **lu**na tras **ló**brega **nu**be*
*con **fran**jas de **pla**ta bor**dar**la_en re**dor**,*
*y **lue**go si_el **vien**to la_a**gi**ta, la **su**be*
*di**suel**ta_a los **ai**res en **blan**co va**por**:*
...

El estudiante de Salamanca
José de Espronceda

Scansion in Spanish

The challenge:

- Syllable contractions / Synaloephas

Not all syllables have a stress value.
How can we handle this?

Scansion in Spanish

The challenge:

- Syllable contractions / Synaloephas

- Heuristic:
  - Main trick: Add unstressed syllables and keep lexical stresses

y	lue	go	si_el	vien	to	la_a	gi	ta	la	su	be
x	/	x	x	/	x	x	/	x	x	/	x

y	lue	go	si	el	vien	to	la	a	gi	ta	la	su	be
x	/	x	x	x	/	x	x	x	/	x	x	/	x

Scansion in Spanish

Spanish poetry Corpus

- 137 sonnets from the Spanish Golden Age (Navarro-Colorado et al., 2015, 2016)
- Statistics

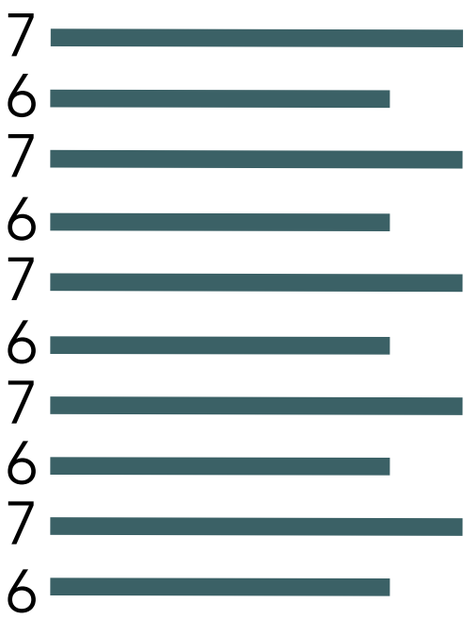
	Spanish corpus
No. syllables	24,524
No. distinct syllables	1,041
No. words	13,566
No. distinct words	3,633
No. lines	1,898

Scansion in Basque

- Basque poetry
  - Long-standing oral tradition
  - Syllabic

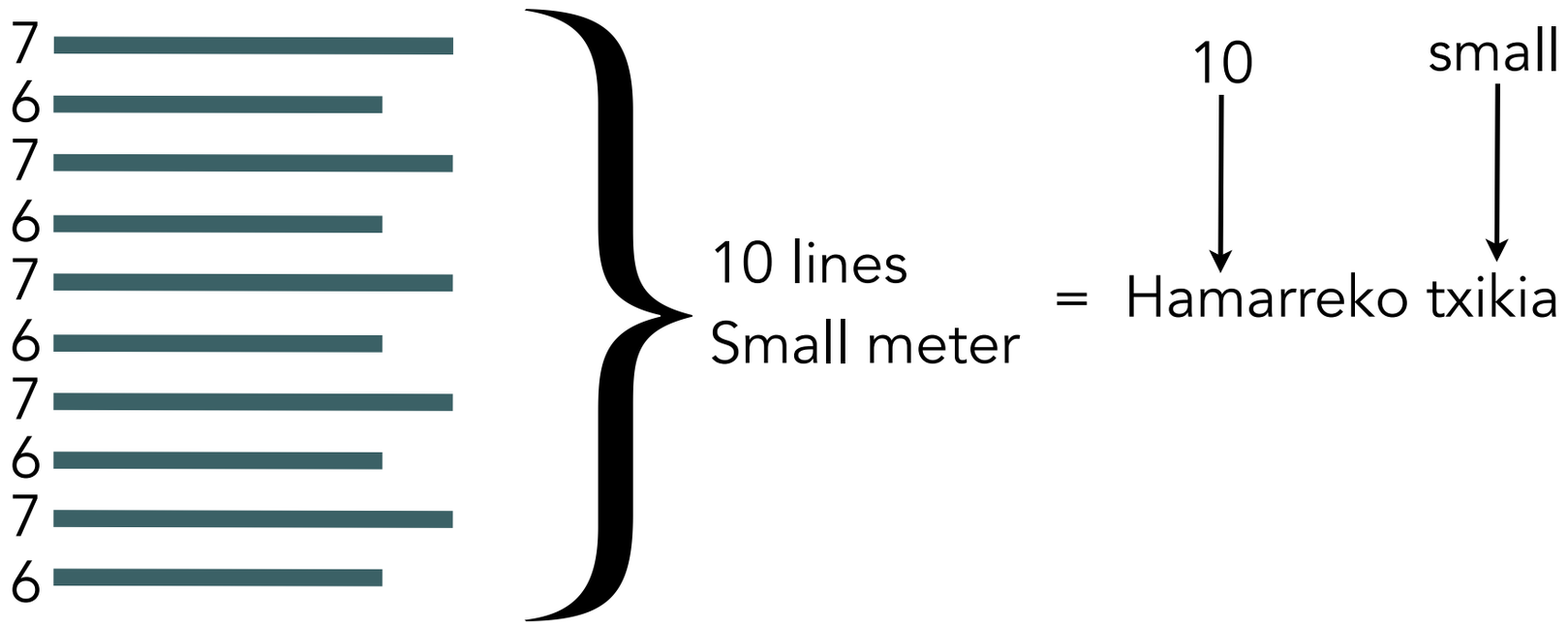
Scansion in Basque

- **Typical metrical structures**
  - Txikiak (small meters)
    - Odd lines, 7 syllables. Even lines, 6 syllables
  - Handiak (big meters)
    - Odd lines, 10 syllables. Even lines, 8 syllables
- The number of lines establishes the name



Scansion in Basque

- **Typical metrical structures**
  - Txikiak (small meters)
    - Odd lines, 7 syllables. Even lines, 6 syllables
  - Handiak (big meters)
    - Odd lines, 10 syllables. Even lines, 8 syllables
- The number of lines establishes the name



Scansion in Basque

- **Old Basque poetry**
  - Not isosyllabic (no regular syllable count per line)
  - The number of beats regular
  - Lekuona (1918): Not just syllable count, but a combination:
  - *"que aquel verso no se mide por silabas sino valiéndose de otra unidad…"*
  - **"that such verse is not measured by syllables but by another type of unit…"**
    - Syllables
    - Plausible feet

- Some researchers claim that rhythm plays an important role in Basque poetry.

- Others state that stress does not play an important role in Basque language.

Scansion in Basque

- My hypothesis

If we ask a group of people (that speak the same dialect) to tag a metrically regular poem, there should be an significant agreement.

Scansion in Basque

- **Challenges:**
  - Lack of metrically annotated corpus
  - Lack of coherent theorization about Basque stress in poetry

Scansion in Basque

Basque poetry Corpus

- 38 poems from the collection Urquizu Sarasua (2009)
  - Tokenized using Ixa-pipes (Agerri et al., 2014)
  - Syllabification based on (Agirrezabal et al., 2012):
    - Onset maximization
    - Sonority hierarchy
- Manually tagged by me

Scansion in Basque

Basque poetry Corpus

- 38 poems from the collection Urquizu Sarasua (2009)
  - Tokenized using Ixa-pipes (Agerri et al., 2014)
  - Syllabification based on (Agirrezabal et al., 2012):
    - Onset maximization
    - Sonority hierarchy
- Manually tagged by me

aplaudir	applause	aplikazio
a-plau	a-pplau	a-plik
ap-lau	ap-plau	ap-lik
apl-au	app-lau	apl-ik
	appl-au	

Scansion in Basque

Basque poetry Corpus

Ene Bizkaiko miatze gorri
zauri zarae mendi ezian!
Aurpegi balzdun miatzarijoi
ator pikotxa lepo-ganian.

Lepo-ganian pikotx zorrotza
eguzki-diz-diz ta mendiz bera.
...

Scansion in Basque

Basque poetry Corpus

```
▼<teiHeader type="text">
 ▼<fileDesc>
  ▼<titleStmt>
     <title>Langile eraildu bati</title>
     <author>Estepan Urkiaga -Lauaxeta-</author>
   </titleStmt>
  ▼<publicationStmt>
     <date>1935</date>
   </publicationStmt>
  ▼<sourceDesc default="false">
     <p/>
   </sourceDesc>
 </fileDesc>
</teiHeader>
▼<text id="POEM_MARKUP">
 ▼<body>
  ▼<lg n="1">
   ▼<l n="1" met="" real="+--+--+--+|-+--+-+--+">
      <!-- Ene Bizkaiko miatze gorri -->
      <seg type="syll" doc="NAF_FILE" targetId="w1">E</seg>
      <seg type="syll" doc="NAF_FILE" targetId="w1">ne</seg>
      <seg type="space"></seg>
      <seg type="syll" doc="NAF_FILE" targetId="w2">Biz</seg>
      <seg type="syll" doc="NAF_FILE" targetId="w2">kai</seg>
      <seg type="syll" doc="NAF_FILE" targetId="w2">ko</seg>
      <seg type="space"></seg>
      <seg type="syll" doc="NAF_FILE" targetId="w3">mi</seg>
      <seg type="syll" doc="NAF_FILE" targetId="w3">a</seg>
      <seg type="syll" doc="NAF_FILE" targetId="w3">tze</seg>
      <seg type="space"></seg>
      <seg type="syll" doc="NAF_FILE" targetId="w4">go</seg>
      <seg type="syll" doc="NAF_FILE" targetId="w4">rri</seg>
   </l>
  ▶<l n="2" met="" real="+--+--+--+|-+--+-+--+">...</l>
  ▶<l n="3" met="" real="-+--+-+--+">...</l>
  ▶<l n="4" met="" real="+--+--+--+|-+--+-+--+">...</l>
  </lg>
 ▼<lg n="2">
  ▼<l n="6" met="" real="-+--+-+--+">
     <!-- Lepo-ganian pikotx zorrotza -->
```

Scansion in Basque

Basque poetry Corpus

- Statistics

	Basque corpus
No. syllables	20,585
No. distinct syllables	920
No. words	7,866
No. distinct words	4,278
No. lines	1,963

Scansion

Summary of corpora

	English corpus	Spanish corpus	Basque corpus
No. of poems	79	137	38
No. syllables	10,988	24,524	20,585
No. distinct syllables	2,283	1,041	920
No. words	8,802	13,566	7,866
No. distinct words	2,422	3,633	4,278
No. lines	1,093	1,898	1,963

Outline

- Research questions and Tasks
- Tradition of scansion
- **Automatic scansion and Sequence modeling**
- NLP techniques for scansion
- General results
- Discussion and Future work

Automatic scansion

- **Rule-based scansion:**
  - Logan (1988), Gervas (2000), Hartman (1996), Plamondon (2006), McAleese (2007), Bobenhausen and Hammerich (2016), Navarro-Colorado (2015, 2017) and Delmonte (2016)
- **Data-driven scansion:**
  - Hayward (1991), Greene et al. (2010), Hayes et al. (2012) and Estes and Hench (2016)

- **Automatic poetry analysis:**
  - Kaplan and Blei (2007), Kao and Jurafsky (2012) and McCurdy et al. (2015)

Sequence modeling

- Greedy prediction
  - Each prediction is done independently, no matter which the output is
- Structured prediction
  - Output transition probabilities come into play

- Poetic scansion as sequence modeling

Sequence modeling

- Greedy prediction
  - Each prediction is done independently, no matter which the output is
- Structured prediction
  - Output transition probabilities come into play

- Poetic scansion as sequence modeling

To swell the gourd and plump the hazel shells

x / x / x / x / x /

	to	swell	the	gourd	and	plump	the	ha	zel	shells
S2S	x	/	x	/	x	/	x	/	x	/

	to	swell	the	gourd	and	plump	the	hazel	shells
W2SP	x	/	x	/	x	/	x	/x	/

Outline

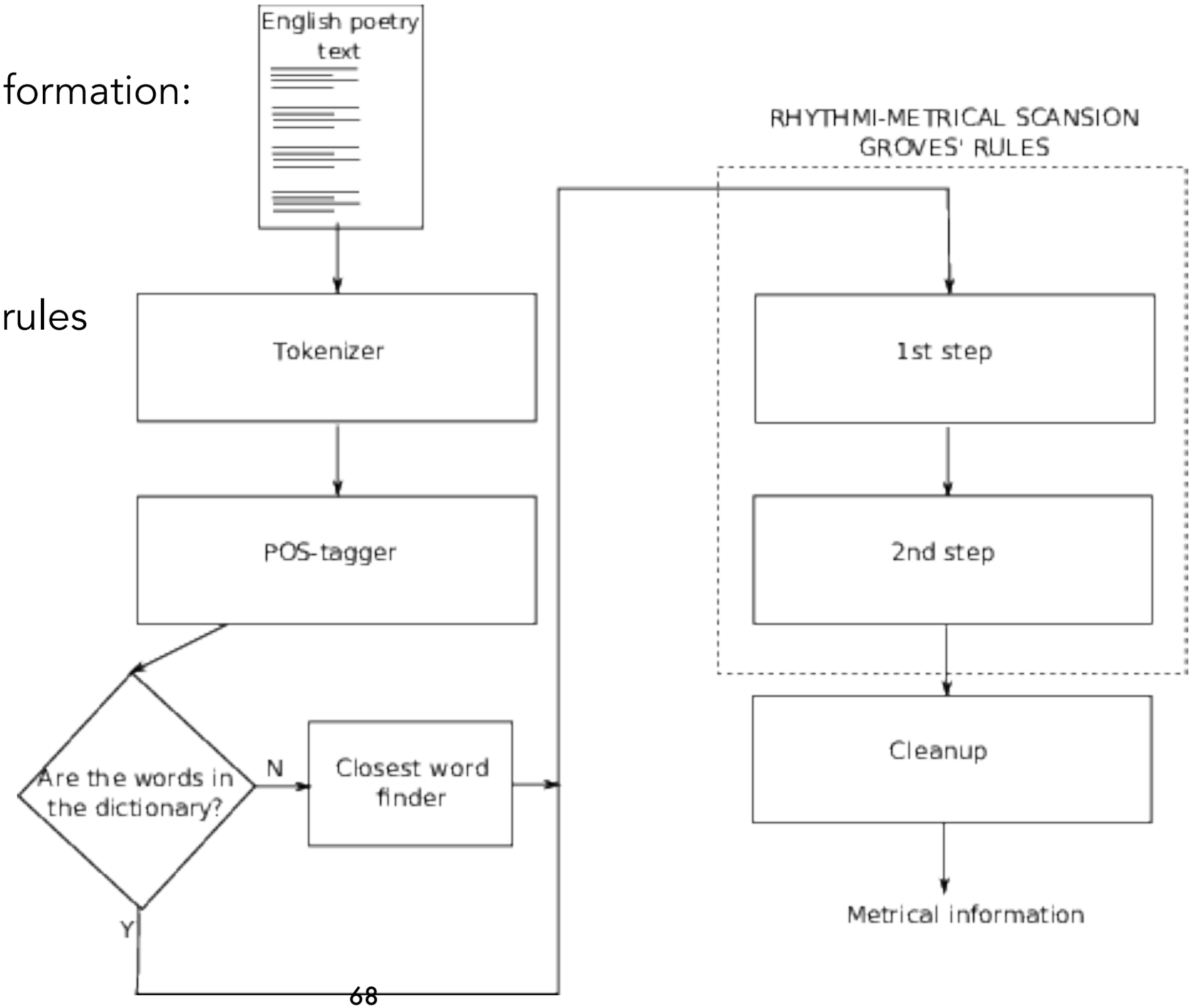
- Research questions and Tasks
- Tradition of scansion
- Automatic scansion and Sequence modeling
- **NLP techniques for scansion**
- General results
- Discussion and Future work

NLP techniques for scansion

- Two ways:
  - Following some rules (by experts)
  - Learning from patterns in the observed data
    - Supervised methods
      - Greedy prediction
      - Structured prediction
      - Neural Networks
    - Unsupervised methods

ZeuScansion: a tool for scansion of English poetry

- Rule-based system
- Two main pieces of information:
  - Lexical stress
  - POS-tag
- Stress assignment:
  - Following Groves' rules



ZeuScansion: a tool for scansion of English poetry

- Groves' rules (Groves, 1998):
  1. Primarily stressed syllable in content words **get primary stress**
  2. Secondary stress of polysyllabic content words, secondary stress in compound words and primarily stressed syllable of polysyllabic function words **get secondary stress**

I dwell in possibility

TOKENIZE	I	dwell	in	possibility
POS-tagger	PRP	VBP	IN	NN
Lexical stress	x	/	x	\x/xx
Beginning	x	x	x	xxxxx
1st step	x	/	x	xx/xx
2nd step	x	/	x	\x/xx

ZeuScansion: a tool for scansion of English poetry

- Groves' rules (Groves, 1998):
  1. Primarily stressed syllable in content words **get primary stress**
  2. Secondary stress of polysyllabic content words, secondary stress in compound words and primarily stressed syllable of polysyllabic function words **get secondary stress**

I dwell in possibility

TOKENIZE	I	dwell	in	possibility
POS-tagger	PRP	VBP	IN	NN
Lexical stress	x	/	x	\x/xx
Beginning	x	x	x	xxxxx
1st step	x	/	x	xx/xx
2nd step	x	/	x	\x/xx

ZeuScansion: a tool for scansion of English poetry

- Groves' rules (Groves, 1998):
  1. Primarily stressed syllable in content words **get primary stress**
  2. Secondary stress of polysyllabic content words, secondary stress in compound words and primarily stressed syllable of polysyllabic function words **get secondary stress**

I dwell in possibility

TOKENIZE	I	dwell	in	possibility
POS-tagger	PRP	VBP	IN	NN
Lexical stress	x	/	x	\x/xx
Beginning	x	x	x	xxxxx
1st step	x	/	x	xx/xx
2nd step	x	/	x	\x/xx

ZeuScansion: a tool for scansion of English poetry

- Groves' rules (Groves, 1998):
  1. Primarily stressed syllable in content words **get primary stress**
  2. Secondary stress of polysyllabic content words, secondary stress in compound words and primarily stressed syllable of polysyllabic function words **get secondary stress**

I dwell in possibility

TOKENIZE	I	dwell	in	possibility
POS-tagger	PRP	VBP	IN	NN
Lexical stress	x	/	x	\x/xx
Beginning	x	x	x	xxxxx
1st step	x	/	x	xx/xx
2nd step	x	/	x	\x/xx

ZeuScansion: a tool for scansion of English poetry

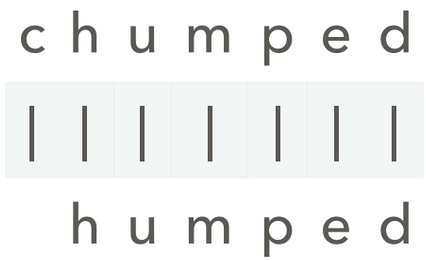
- When we do not know the lexical stress
- We find a similarly spelled word, expecting that it will be pronounced similarly
- Closest Word Finder
  - FST-based system that finds the closest spelled word in the dictionary.

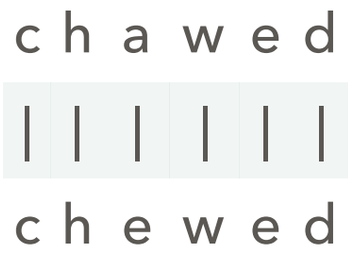
We chumped and chawed the buttered toast

chumped and **chawed** are not in the dictionary.

We must find a similarly pronounced word.

ZeuScansion: a tool for scansion of English poetry





The similarly pronounced words presented by the Closest Word Finder are **humped** and **chewed**.

We chumped and chawed the buttered toast

We humped and chewed the buttered toast

ZeuScansion: a tool for scansion of English poetry

Barred with streaks of red and yellow	*/ x \ x / x / *
Streaks of blue and bright vermilion	*\ x / x / x / x*
Shone the face of Pau-Puk-Keewis	*/ x / x ?*
From his forehead fell his tresses	*x x / \ / x \ x*
Smooth and parted like a woman's	*/ x \ x x x \ x*
...	...

Syllable	1	2	3	4	5	6	7	8
Count (stressed)	14	0	19	1	14	0	12	1
Normalized	0.74	0	1	0.05	0.74	0	0.63	0.05
Average Stress	/	x	/	x	/	x	/	x

ZeuScansion: a tool for scansion of English poetry

Predominant stress: / x / x / x / x

How can we split it?

4 trochees	2 amphibrachs	3 iambs
[/ x] [/ x] [/ x] [/ x]	/ [x / x] / [x / x]	/ [x /] [x /] [x /] x

Name	Feet	N° matches	Score
trochee	[/ x]	4	4
amphibrach	[x / x]	2	3
iamb	[x /]	3	3

ZeuScansion: a tool for scansion of English poetry

Results on English data

	Per syllable (%)	Per line (%)
ZeuScansion	86.17	29.37
Scandroid	**87.42**	**34.49**

Global analysis

	Correctly classified (%)
The song of Hiawatha	32.03
Shakespeare's Sonnets	70.13

ZeuScansion: a tool for scansion of English poetry

These results have been published in:

Agirrezabal, M., Astigarraga, A., Arrieta, B., & Hulden, M. (2016)
ZeuScansion: a tool for scansion of English poetry
Journal of Language Modelling, 4(1), 3-28.

Agirrezabal, M., Arrieta, B., Astigarraga, A., and Hulden, M. (2013)
ZeuScansion: a tool for scansion of English poetry
Finite State Methods and Natural Language Processing Conference, 18-24.

Supervised Learning

Features

- 10 basic features (almost language agnostic):
  - Syllable position within the word
  - Syllable position within the line
  - Number of syllables in the line
  - Syllable's phonological weight
  - Word length
  - Last char, last 2 chars, ..., last 5 chars of the word

Supervised Learning

Features

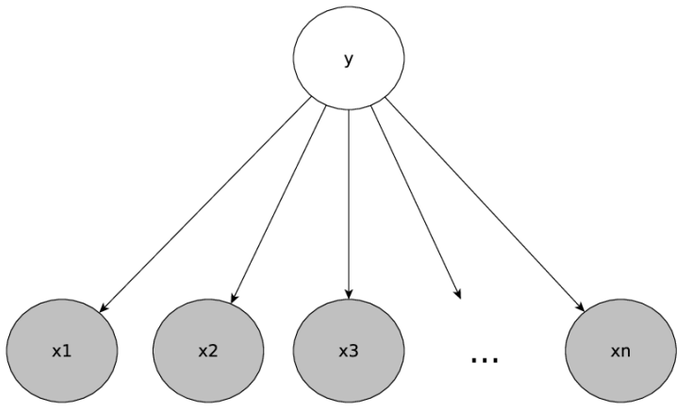
- Additional features:
  - Syllable (t±10)
  - Word (t±5)
  - Part-of-speech tag (t±5)
  - Lexical stress (t±5)*

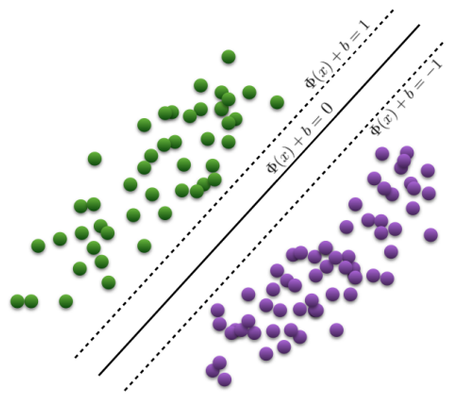
*In the case of OOV words, we calculate their lexical stress using an SVM-based implementation presented in Agirrezabal et al., 2014.

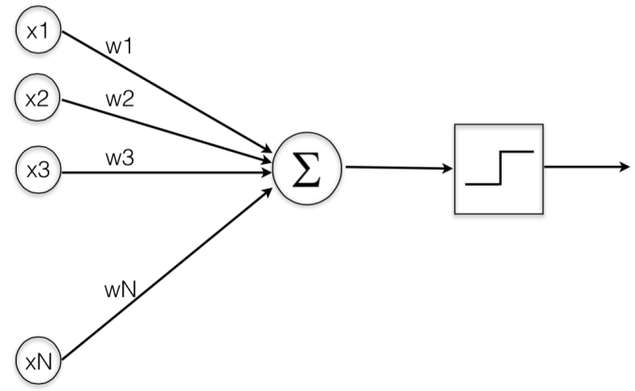
Supervised Learning

Greedy prediction / Structured prediction

- Greedy Predictors:
  - Naive Bayes
  - Averaged Perceptron
  - Linear Support Vector Machines







- Structured predictors
  - Hidden Markov Models (HMM)
  - Conditional Random Fields (CRF)

Supervised Learning

Greedy prediction

Results on English data

10 features

	Per syllable (%)	Per line (%)
ZeuScansion	**86.17**	**29.37**
Naive Bayes	78.06	9.53
Linear SVM	83.50	22.31
Perceptron	85.04	28.79

64 features

	Per syllable (%)	Per line (%)
ZeuScansion	86.17	29.37
Naive Bayes	80.96	13.51
Linear SVM	87.42	34.45
Perceptron	**89.12**	**40.86**

Supervised Learning

Structured prediction

Results on English data

	#FTs	Per syllable (%)	Per line (%)
ZeuScansion	-	86.17	29.37
Scandroid	-	87.42	34.49
HMM (just syll)	-	90.39	48.51
CRF (just syll)	1	88.01	43.85
CRF	10	89.32	47.28
CRF	64	**90.94**	**51.22**

Supervised Learning

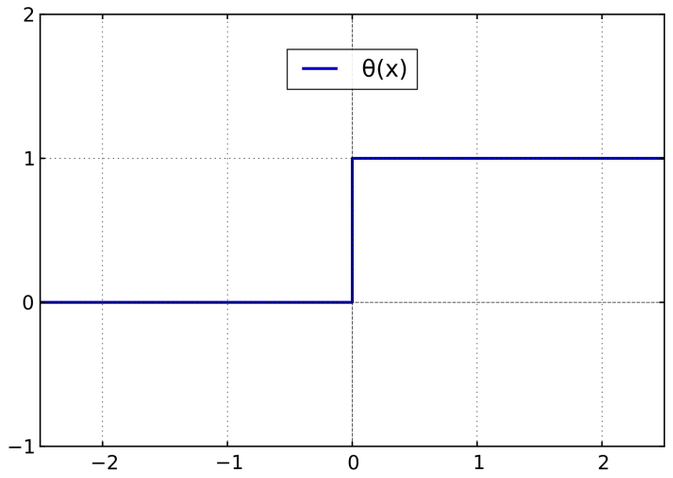
These results have been published in:

Agirrezabal, M., Alegria, I., & Hulden, M. (2016, December).
Machine Learning for the Metrical Analysis of English Poetry.
International Conference on Computational Linguistics (COLING 2016), 772-781

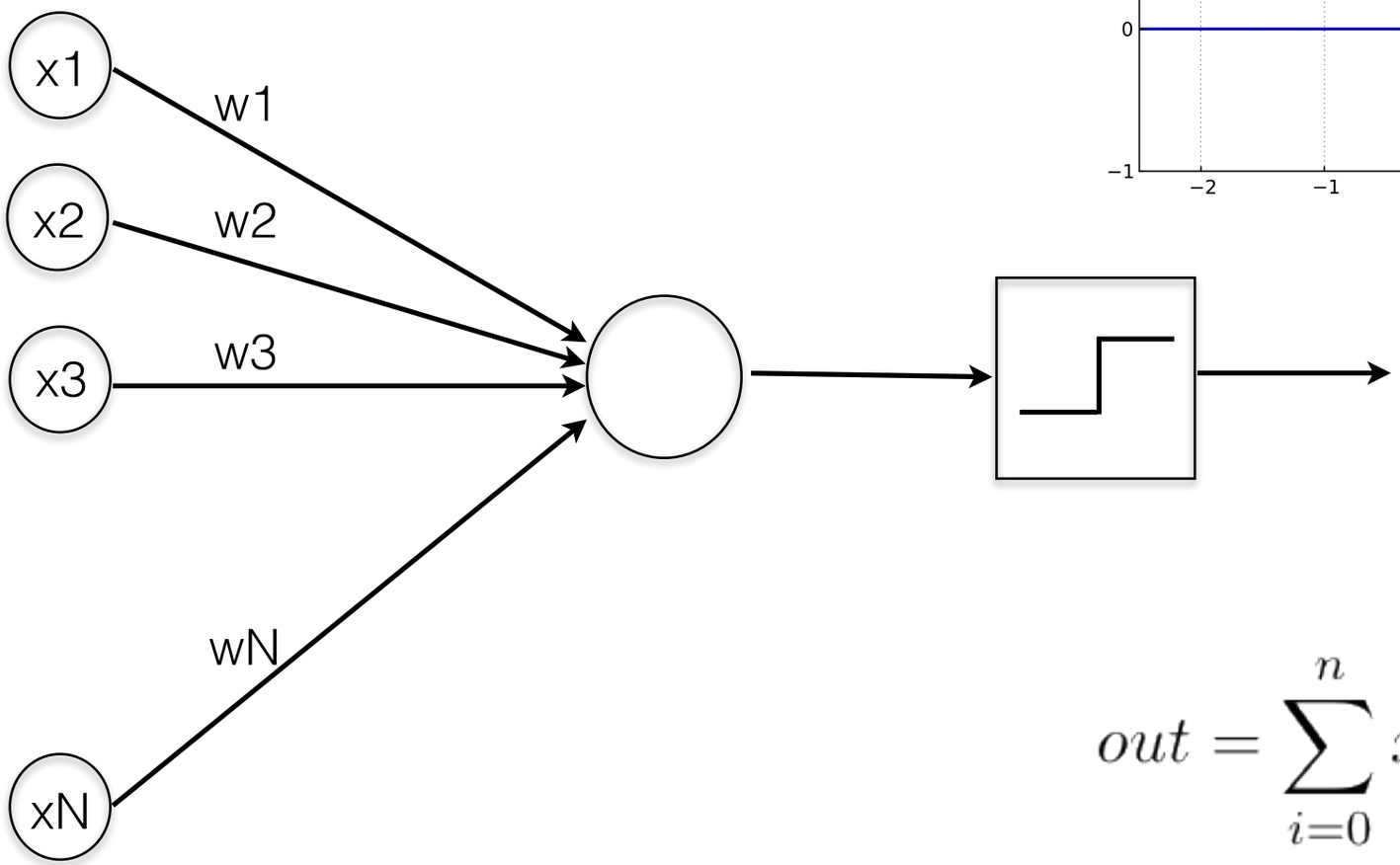
Supervised Learning

Neural Networks

Heaviside step function



Perceptron



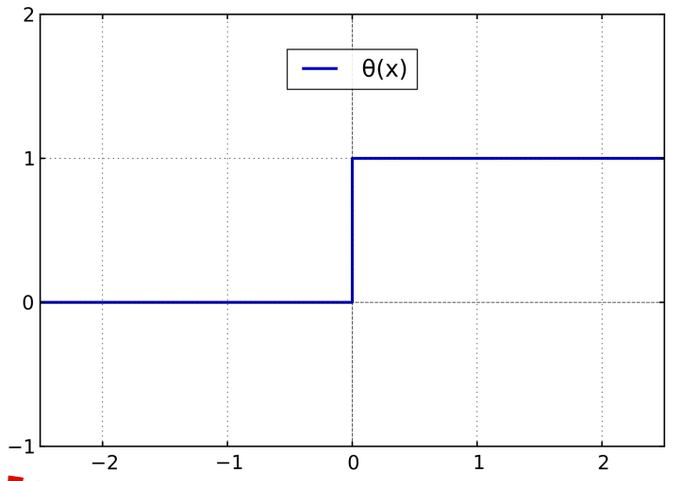
$$out = \sum_{i=0}^{n} x_i w_i$$

$$y = H(out)$$

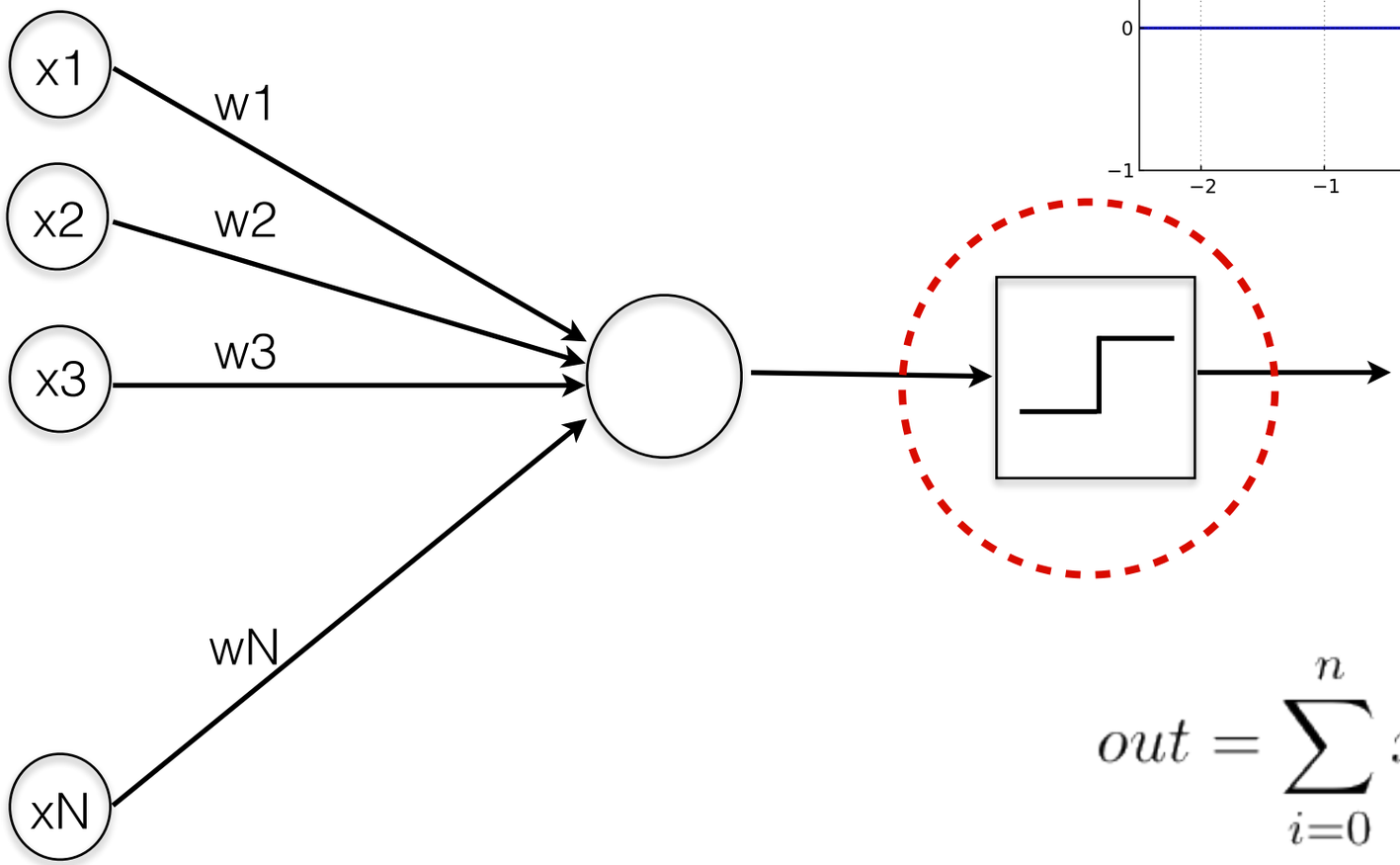
Supervised Learning

Neural Networks

Heaviside step function



Perceptron



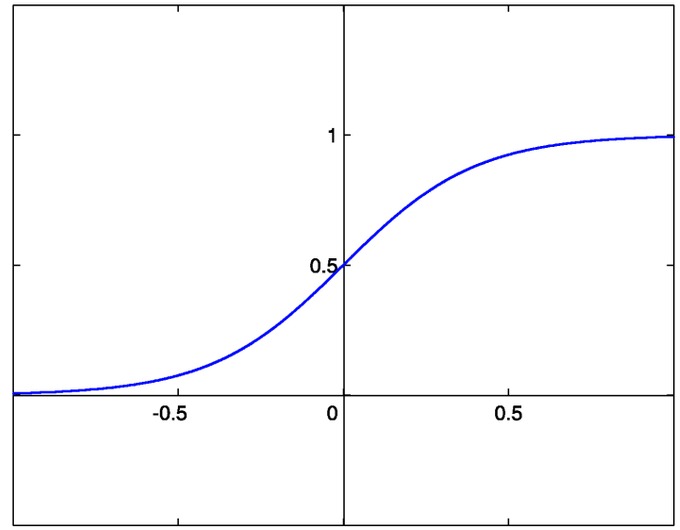
$$out = \sum_{i=0}^{n} x_i w_i$$

$$y = H(out)$$

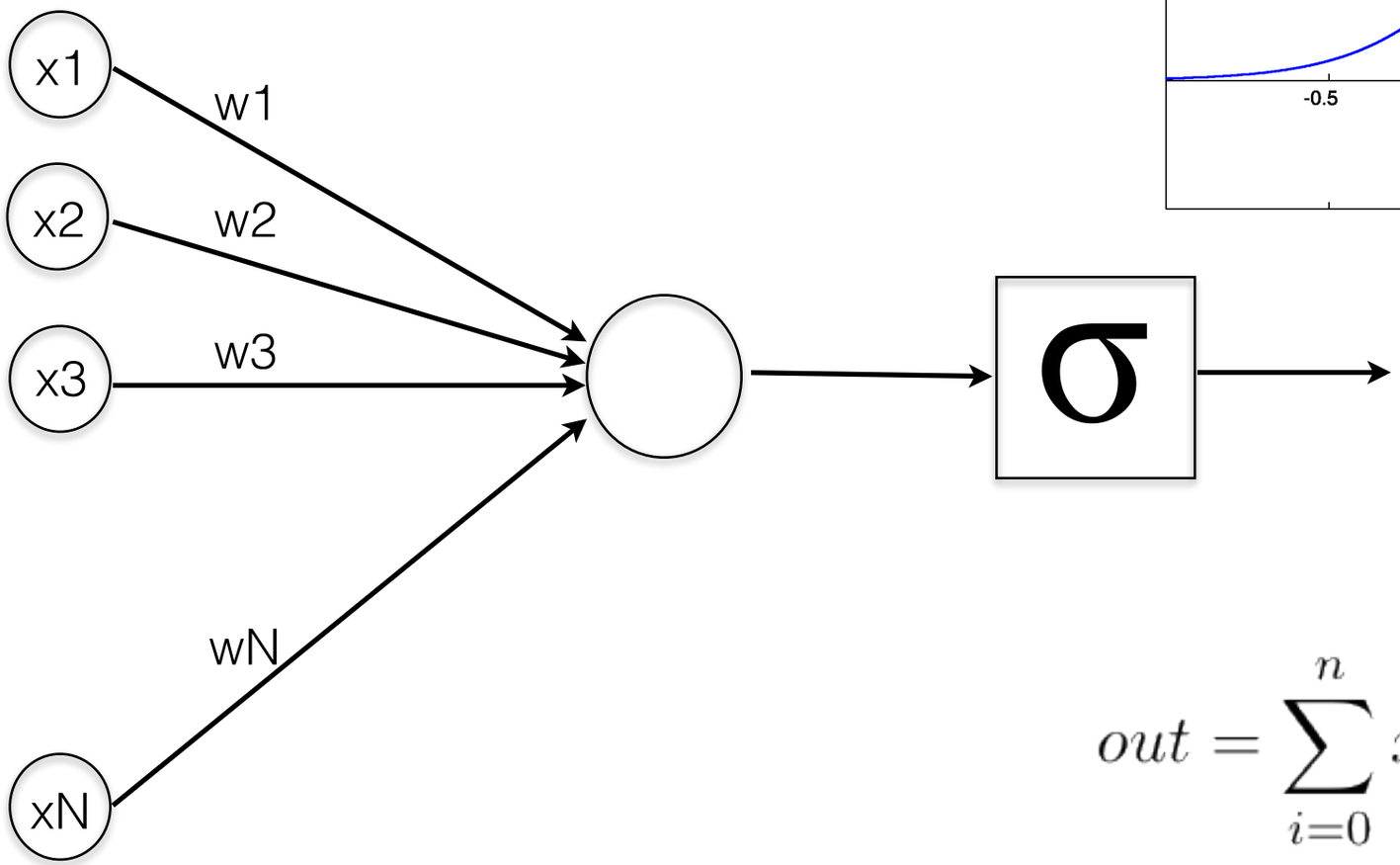
Supervised Learning

Neural Networks

Sigmoid function



Logistic Regression



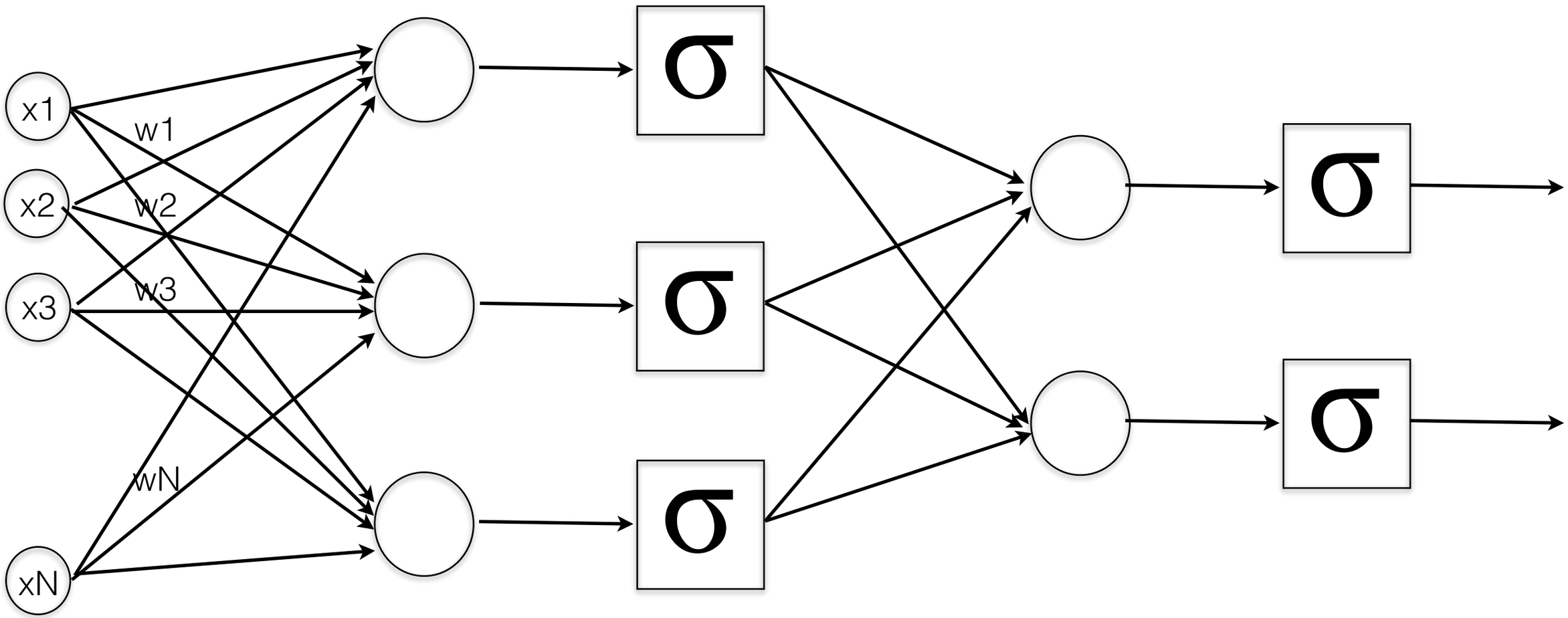
$$out = \sum_{i=0}^{n} x_i w_i$$

$$y = H(out)$$

Supervised Learning

Neural Networks

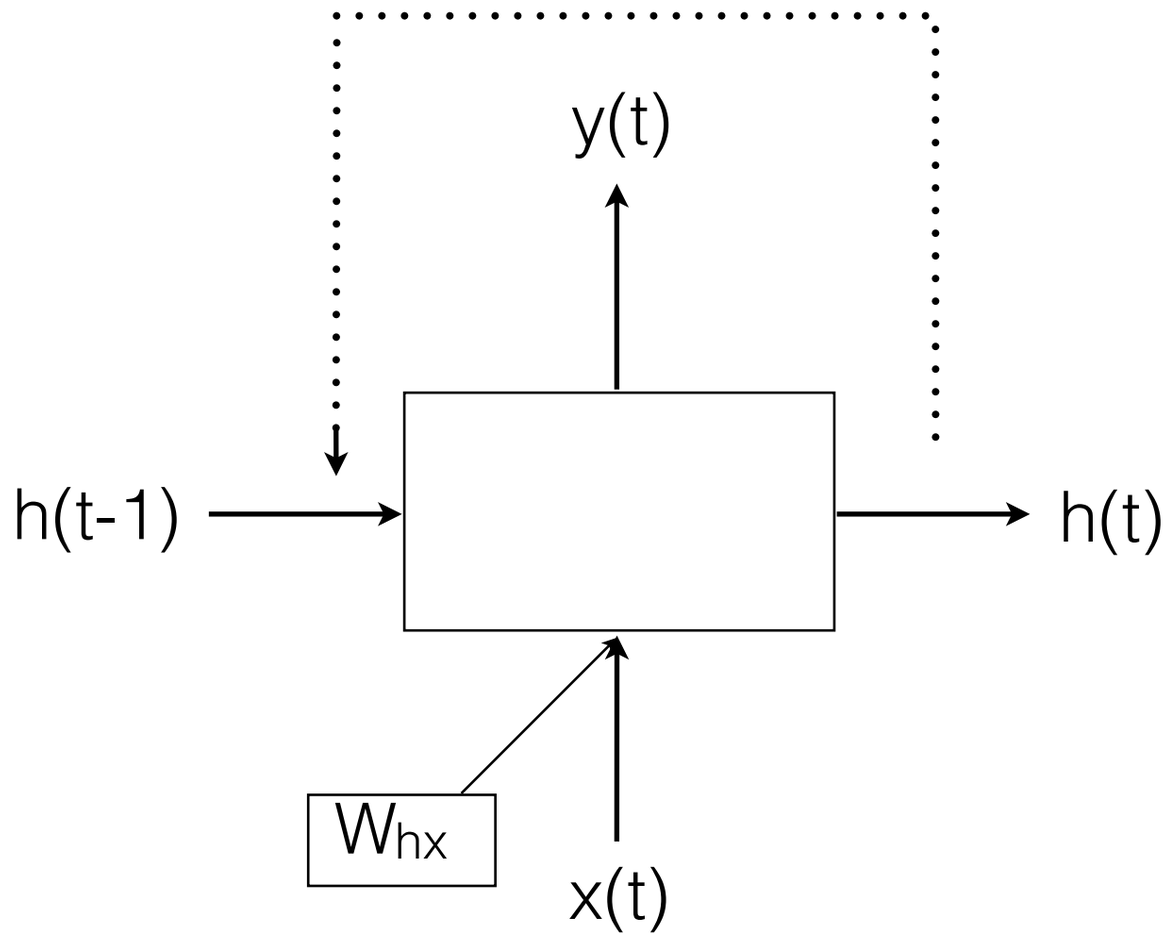
Multilayer Perceptron (2 layers)



Supervised Learning

Neural Networks

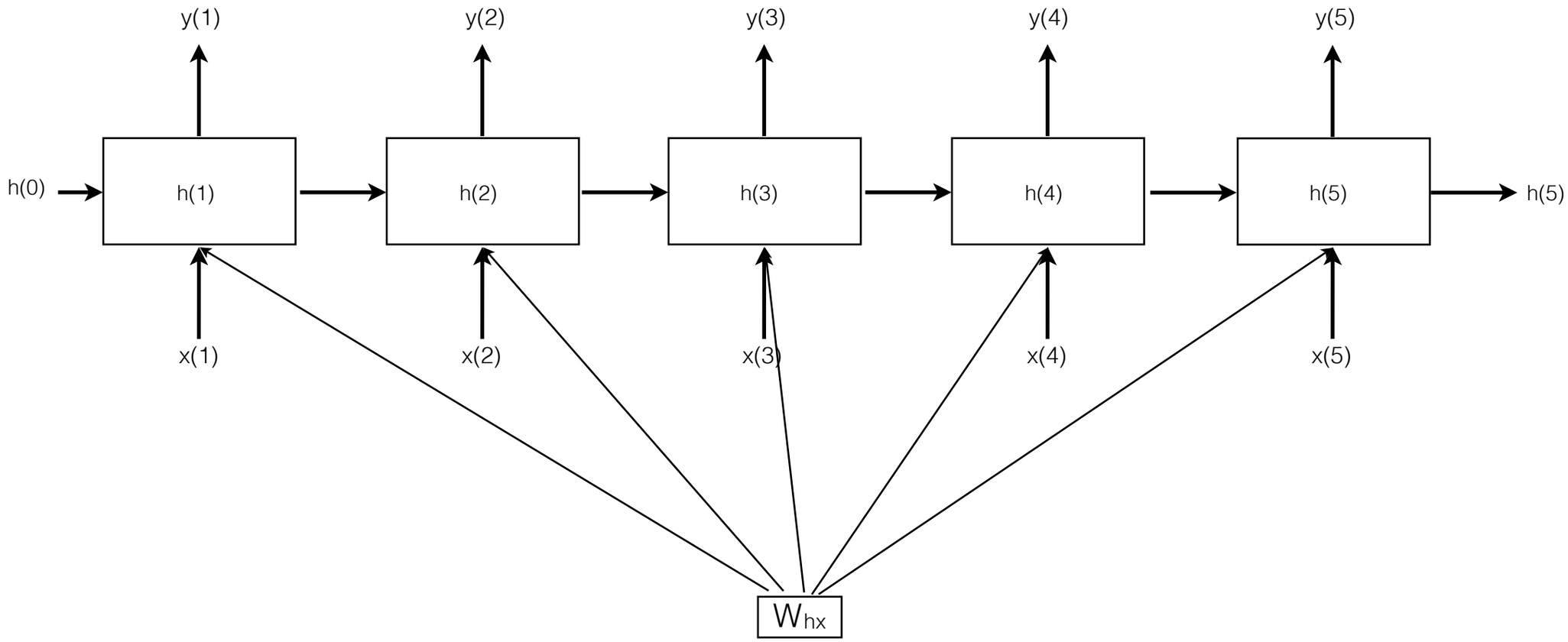
Recurrent Neural Network (recursive representation)



Supervised Learning

Neural Networks

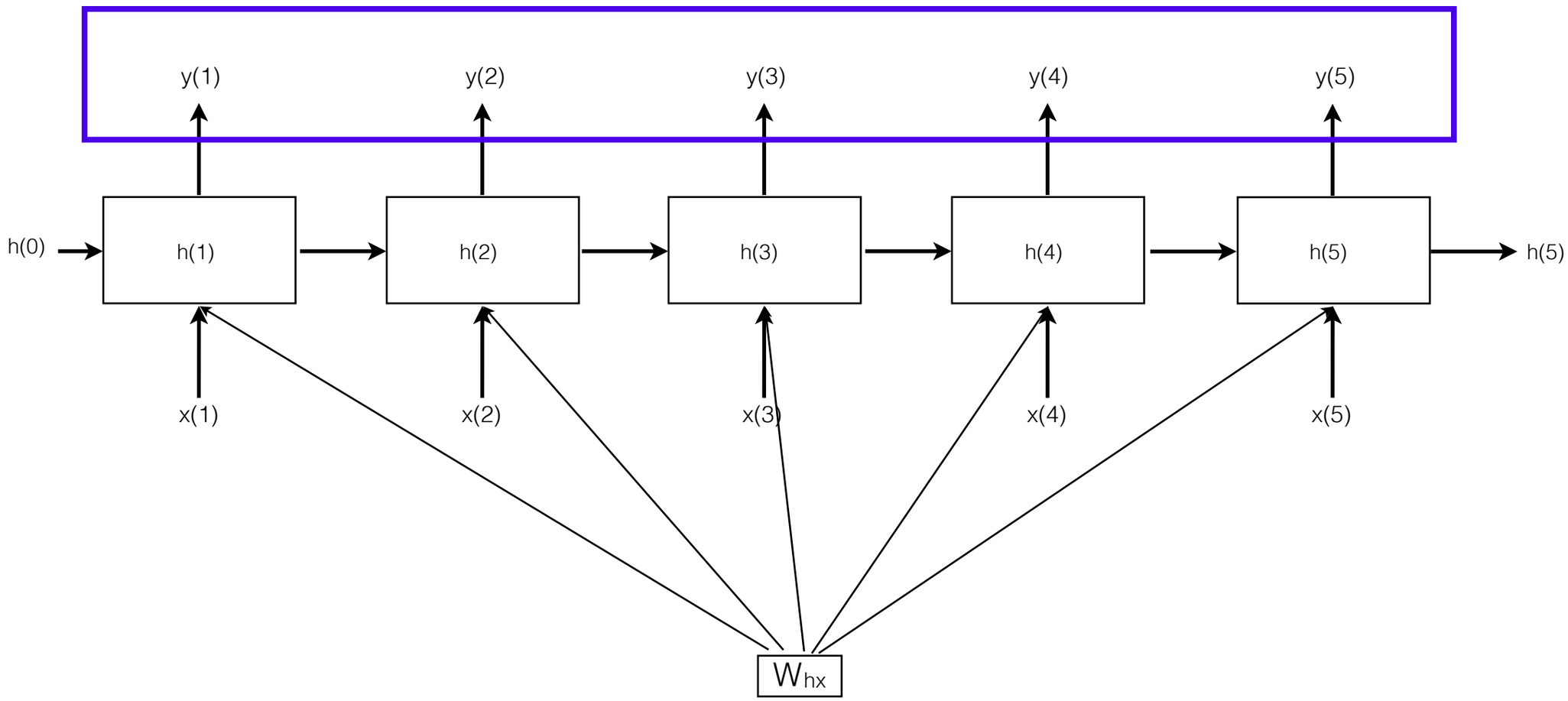
Recurrent Neural Network (unfolded)



Supervised Learning

Neural Networks

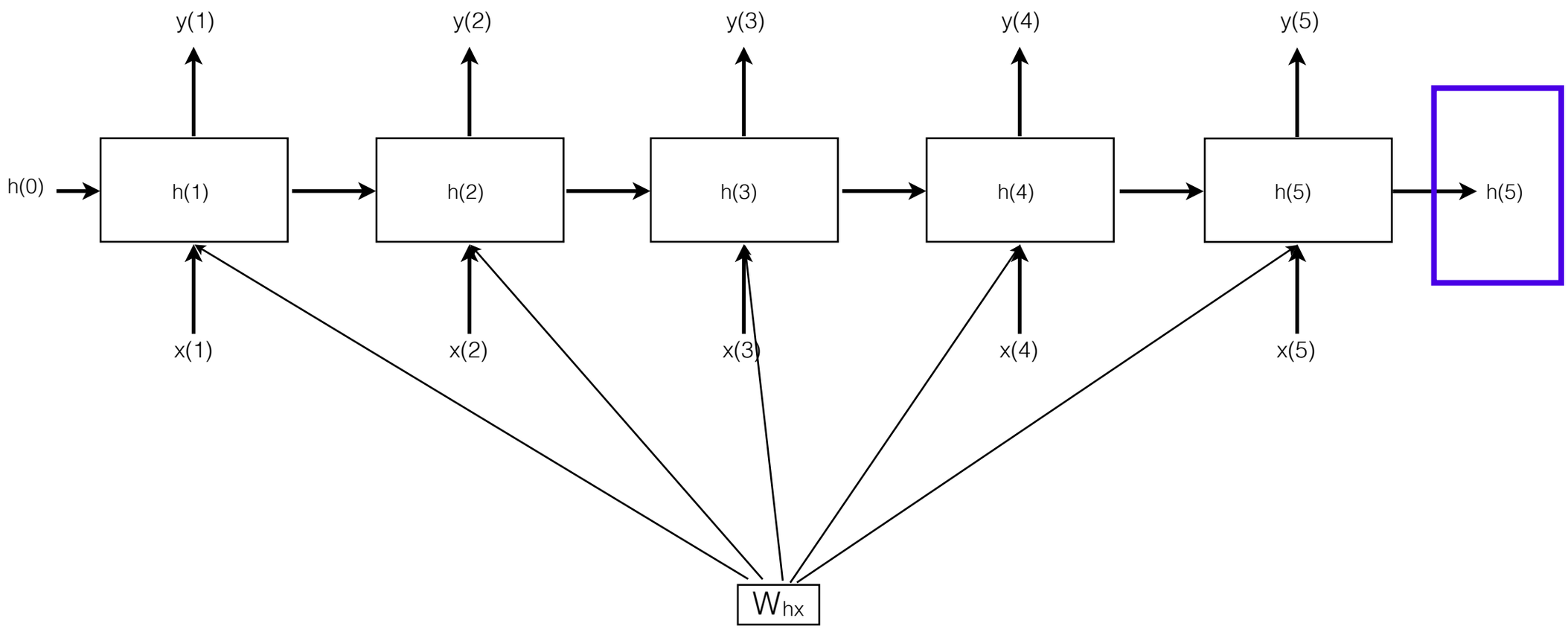
Recurrent Neural Network (unfolded)



Supervised Learning

Neural Networks

Recurrent Neural Network (unfolded)



Supervised Learning

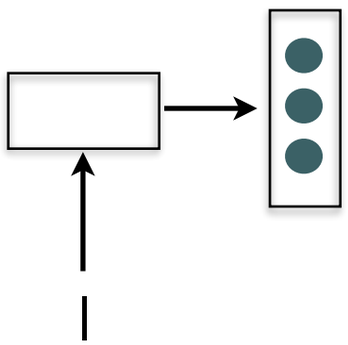
Neural Networks

- Encoder-Decoder model
- Widely used
- Succesful in tasks such as:
  - Machine Translation (Sutskever et al., 2014)
  - Morphological Reinflection (Kann and Schütze, 2016)

Supervised Learning

Neural Networks

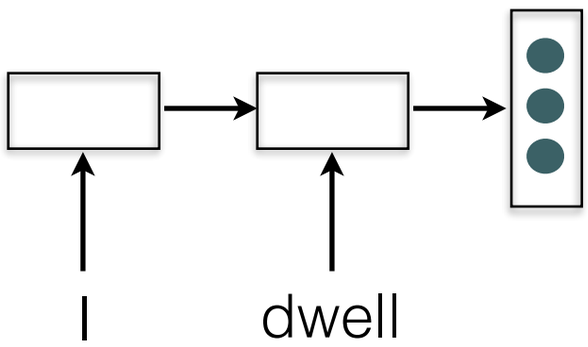
- Encoder-Decoder model
- Widely used
- Succesful in tasks such as:
  - Machine Translation (Sutskever et al., 2014)
  - Morphological Reinflection (Kann and Schütze, 2016)



Supervised Learning

Neural Networks

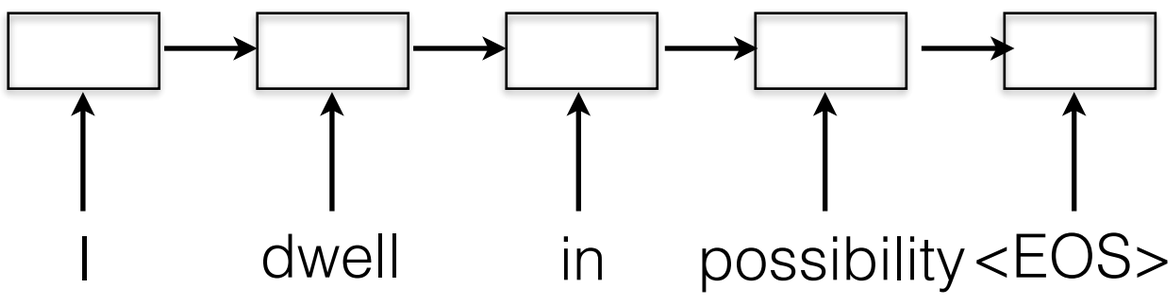
- Encoder-Decoder model
- Widely used
- Succesful in tasks such as:
  - Machine Translation (Sutskever et al., 2014)
  - Morphological Reinflection (Kann and Schütze, 2016)



Supervised Learning

Neural Networks

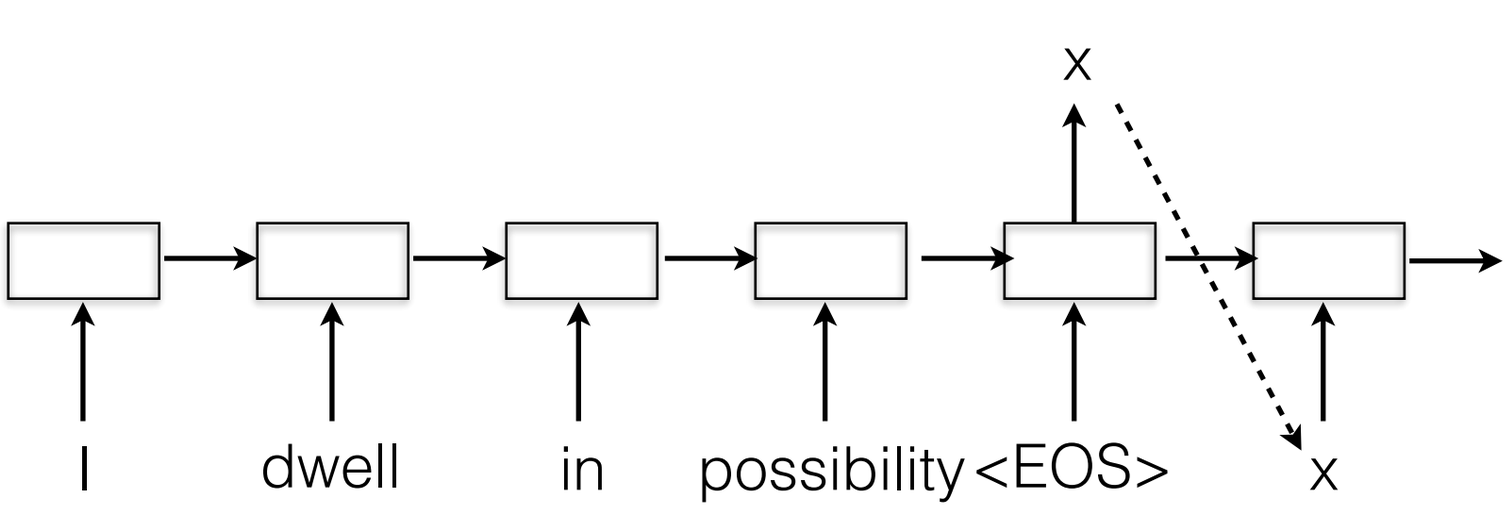
- Encoder-Decoder model
- Widely used
- Succesful in tasks such as:
  - Machine Translation (Sutskever et al., 2014)
  - Morphological Reinflection (Kann and Schütze, 2016)



Supervised Learning

Neural Networks

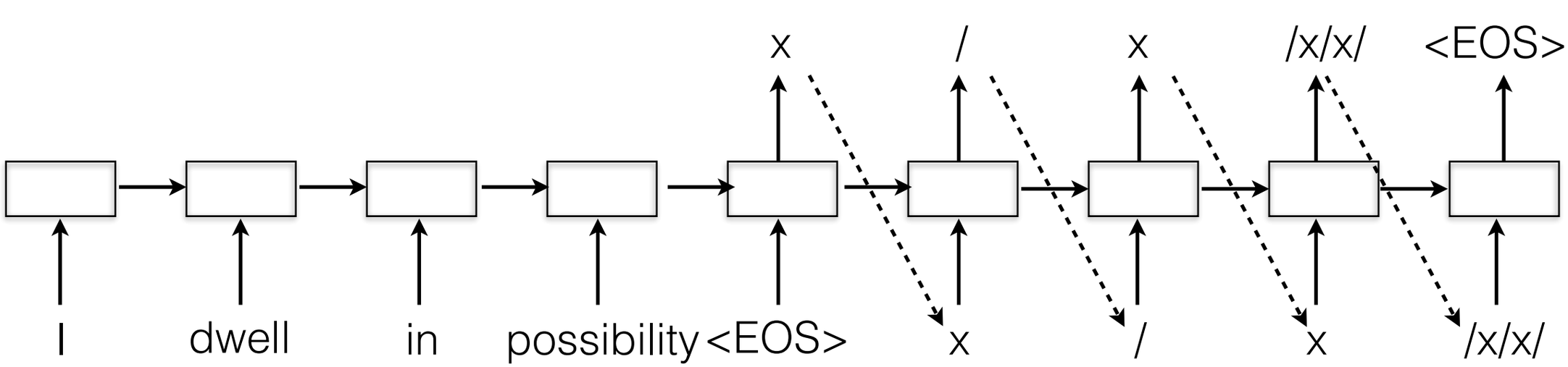
- Encoder-Decoder model
- Widely used
- Succesful in tasks such as:
  - Machine Translation (Sutskever et al., 2014)
  - Morphological Reinflection (Kann and Schütze, 2016)



Supervised Learning

Neural Networks

- Encoder-Decoder model
- Widely used
- Succesful in tasks such as:
  - Machine Translation (Sutskever et al., 2014)
  - Morphological Reinflection (Kann and Schütze, 2016)



Supervised Learning

Encoder-Decoder

Results on English data (development set)

	Per syllable (%)	Per line (%)
S2S	84.52	30.93
W2SP	85.44	34.00

Supervised Learning

Neural Networks

- Bi-LSTM+CRF (Lample et al., 2016)
- Gets information from input characters and words with Bi-LSTMs
- The information goes through a CRF layer to model the output dependencies
- Succesful in tasks such as:
  - Named Entity Recognition
  - Poetry scansion
- Advantages:
  - Words' character sequence
  - Interaction between words
  - Conditional dependencies between outputs

Supervised Learning

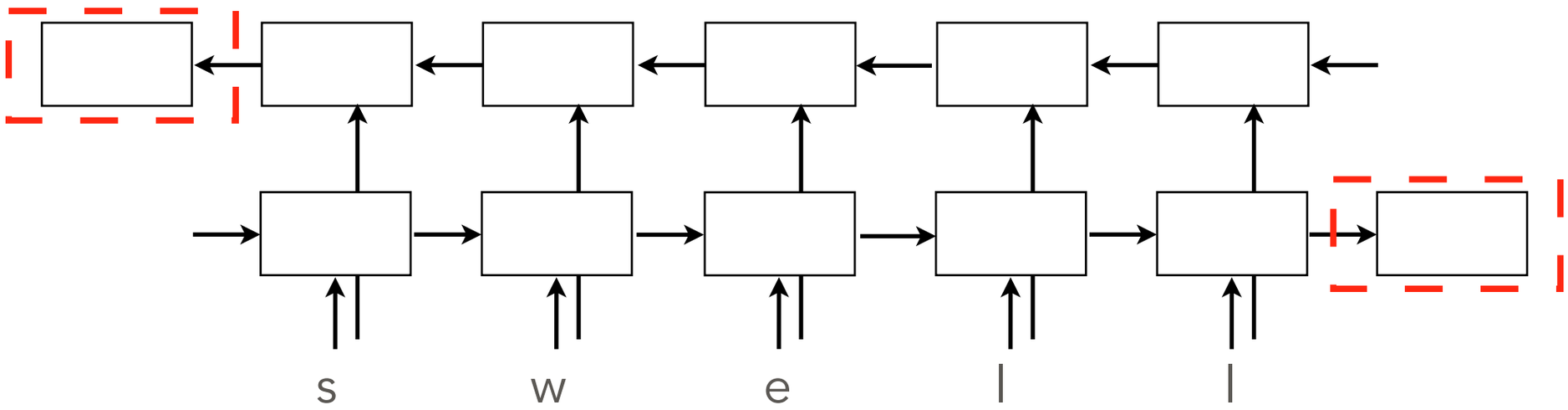
Neural Networks

- Words are modeled using three pieces of information:
  - Forward LSTMs output
  - Backward LSTMs output
  - Word embedding

These vectors are concatenated

LOOKUP table

...	
dwell	0.176 0.635 .... 0.121
...	
swear	0.477 0.233 ... 0.654
sweat	0.264 0.925 ... 0.137
...	0.187 0.649 ... 0.319
swell	0.934 0.197 ... 0.194
...	

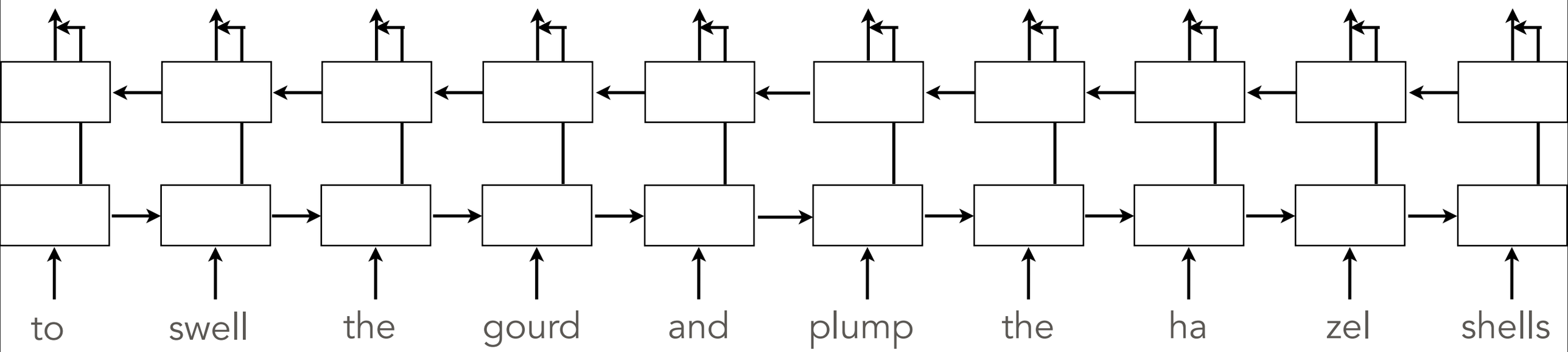


Supervised Learning

Neural Networks

- In the sentence level
- Previous vectors are combined with:
  - Left context (forward LSTM)
  - Right context (backward LSTM)

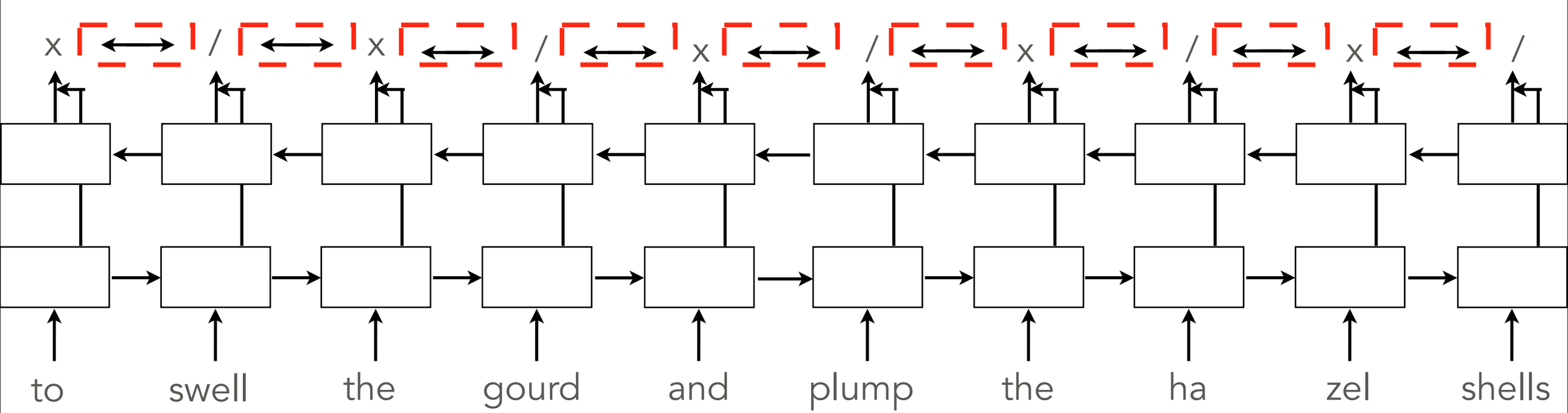
The information of the two sentence-level LSTMs is concatenated.



Supervised Learning

Neural Networks

- Dependencies among outputs are modeled with a CRF layer



Supervised Learning

Bi-LSTM+CRF

Results on English data (development set)

	Per syllable (%)	Per line (%)
W2SP	90.80	53.29
S2S	93.06	61.95

Supervised Learning

Bi-LSTM+CRF

Results on English data (development set)

	Per syllable (%)	Per line (%)
W2SP	90.80	53.29
S2S	93.06	61.95
S2S+WB	**94.49**	**69.97**

Supervised Learning

Bi-LSTM+CRF

Results on English data (development set)

	Per syllable (%)	Per line (%)
W2SP	90.80	53.29
S2S	93.06	61.95
S2S+WB	**94.49**	**69.97**

Results on English data (test set)

	Per syllable (%)	Per line (%)
W2SP	89.39	44.29
S2S	91.26	55.28
S2S+WB	**92.96**	**61.39**

Supervised Learning

Results on English data (test set)

	#FTs	Per syllable (%)	Per line (%)
Perceptron	10	85.04	28.79
Perceptron	64	89.12	40.86
HMM	-	90.39	48.51
CRF	10	89.32	47.28
CRF	64	90.94	51.22
Bi-LSTM+CRF (W2SP)	-	89.39	44.29
Bi-LSTM+CRF (S2S)	-	91.26	55.28
Bi-LSTM+CRF (S2S+WB)	-	**92.96**	**61.39**

Unsupervised Learning

We did several experiments:

1. Simple cross-lingual experiment
2. Clustering algorithms
    1. K-Means
    2. Expectation-Maximization
3. Hidden Markov Models

Unsupervised Learning

We did several experiments:

1. Simple cross-lingual experiment (best result 71.65%)
2. Clustering algorithms with 64 feature templates (results below 55%)
    1. K-Means
    2. Expectation-Maximization
3. Hidden Markov Models

Results on English data

	Per syllable (%)	Per line (%)
HMM (4 states)	66.28	7.29
HMM (8 states)	74.65	9.91
HMM (16 states)	**76.51**	**12.53**
HMM (32 states)	74.03	8.07

Outline

- Research questions and Tasks
- Tradition of scansion
- Automatic scansion and Sequence modeling
- NLP techniques for scansion
- **General results**
- Discussion and Future work

General results

Supervised learning methods (test set)

		English		Spanish		Basque	
	#FTs	Per syllable (%)	Per line (%)	Per syllable (%)	Per line (%)	Per syllable (%)	Per line (%)
ZeuScansion		86.17	29.37	-	-	-	-
Perceptron	**10**	85.04	28.79	74.39	0.44	71.77	9.74
Perceptron	**64**	89.12	40.86	91.49	35.71	69.86	8.47
HMM	**-**	90.39	48.51	92.32	45.08	80.97	24.10
CRF	**10**	89.32	47.28	84.89	18.61	81.19	26.23
CRF	**64**	90.94	51.22	92.87	55.44	80.52	**26.93**
Bi-LSTM+CRF (W2SP)	**-**	89.39	44.29	**98.95**	**90.84**	**83.19**	23.75
Bi-LSTM+CRF (S2S)	**-**	91.26	55.28	95.13	63.68	79.38	20.32
Bi-LSTM+CRF (S2S+WB)	**-**	**92.96**	**61.39**	98.74	88.82	79.66	24.67

Outline

- Research questions and Tasks
- Tradition of scansion
- Automatic scansion and Sequence modeling
- NLP techniques for scansion
- General results
- **Discussion and Future work**

Discussion and Future work

- Analysis and development of methods for automatic poetic scansion
  - Rule-based
  - Data-driven
- Main investigation in English
- Best resulting models to Spanish and Basque

Discussion and Future work

Conclusions

- ZeuScansion: promising results
- Data-driven approaches
  - Previous results improved upon
  - Structural information
- Supervised learning: >80% for all languages
- Generally, best results with BiLSTM+CRF
  - No hand-crafted fetures
  - They model the phonological structure of words/syllables
- Almost direct extrapolation to Spanish and similar results
  - This shows the robustness of the models for the problem of Scansion
- Preliminary experiments for Basque
- Promising results in unsupervised learning

Discussion and Future work

Research questions

1.- What do we need to know when analyzing a poem and how can we capture it?

ZeuScansion: Lexical stress and POS-tag

Additional features improve results significantly

Output dependencies improve results

Bi-LSTMs as feature extractors

Discussion and Future work

Research questions

2.- Does language-specific linguistic knowledge contribute when analyzing poetry?

Lexical stresses and POS-tags boost the accuracy of the predictors

Word structure information is helpful (word boundary)

Cross-lingual experiment, low results.

Discussion and Future work

Research questions

3.- Is it possible to analyze a poem without any language-specific information? Is such analysis something that can be learnt?

Results of 75% without using tagged information

The results of these models should be included as features

Discussion and Future work

Contributions

- ZeuScansion: Rule-based system
- Data-driven approaches: Revealed important aspects when analyzing poetry
- New dataset of Basque poetry

Discussion and Future work

Future work

- Independence between lines
- Inclusion of HMM results as features (semi supervised learning)
- Apply this to poetry generation
- Check the validity of this work with acoustic information

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

Automatic scansion of poetry

Manex Agirrezabal Zabaleta

PhD dissertation

Dept. of Computer and Language Systems
University of the Basque Country (UPV / EHU)

Supervisors: Iñaki Alegria, Mans Hulden

June 19, 2017

Scansion in Basque

- Old Basque poetry
  - Not isosyllabic
  - The number of beats regular
  - Lekuona (1918): Not just syllable count, but a combination:
    - Syllables
    - Plausible feet

no diptongo. Con lo cual creo que se prueba suficientemente la falta de isocronía de las sílabas del Euskera: procede, pues, rechazar el sistema de la sílaba, unidad de la medida del verso, y consiguientemente también el sistema del número de sílabas como base real y sólida de la versificación vasca.

- Some researchers claim that rhythm plays an impor

crónos, aunque su número de sílabas no sea igual. Dedúcese de aquí, que la unidad métrica de aquel verso no es la sílaba, que aquel verso no se mide por sílabas, sino valiéndose de otra unidad, en la cual, por precisión, dos sílabas equivalgan a una, y una equivalga a dos. Es decir, que nos hallamos en pleno terreno, en que

- Others state that stress does not play an important role in Basque language.

dos. Es decir, que nos hallamos en pleno terreno, en que la cantidad silábica es variable, y en el cual no se puede prescindir de tomar por unidad métrica el pie rítmico, pues en él únicamente es donde dos breves valen por una larga, y el espondeo por ejemplo (pie de dos largas) equivale al dáctilo (pie de larga y dos breves).

ZeuScansion: a tool for scansion of English poetry

Word change rules:

1. At the end of the word, higher cost (Word splitter)
2. We only allow a maximum of 2 character changes
3. Change characters in the following order:
    1. 1 vowel
    2. 1 consonant
    3. 2 vowels
    4. 1 vowel and 1 consonant
    5. 2 consonants

Word splitter:
chumped: chum | ped
chawed: cha | wed

ZeuScansion: a tool for scansion of English poetry

c h u m p e d
| | | | | | |
h u m p e d

c h a w e d
| | | | | |
c h e w e d

The similarly pronounced words presented by the Closest Word Finder are **humped** and **chewed**.

TOKENIZE	POS-tagger	1st step	2nd step	CleanUp
we	we+PRP	we+x+PRP	we+x+PRP	x
chumped	chumped+VBD	humped+/+VBD	humped+/+VBD	/
and	and+CC	and+x+CC	and+x+CC	x
chawed	chawed+VBD	chewed+/+VBD	chewed+/+VBD	/
the	the+DT	the+x+DT	the+x+DT	x
buttered	buttered+JJ	buttered+/x+JJ	buttered+/x+JJ	/x
toast	toast+NN	toast+/+NN	toast+/+NN	/

ZeuScansion: a tool for scansion of English poetry

Once stresses are marked,
ZeuScansion tries do identify the predominant meter of the poem,
by finding plausible feet.

Barred with streaks of red and yellow	/ x \ x / x / \
Streaks of blue and bright vermilion	\ x / x / x / x
Shone the face of Pau-Puk-Keewis	/ x / x ?
From his forehead fell his tresses	x x / \ / x \ x
Smooth and parted like a woman's	/ x \ x x x \ x
Shining bright with oil and plaited	\ x / x / x \ x
Hung with braids of scented grasses	/ x \ x \ x \ x
As among the guests assembled	/ x \ x \ x \ x
To the sound of flutes and singing	x x / x \ x \ x
To the sound of drums and voices	x x / x \ x \ x
Rose the handsome Pau-Puk-Keewis	/ x / x ?
And began his mystic dances	x x \ x / x \ x