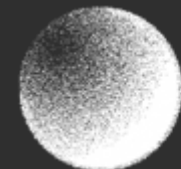


Modelagem de linguagens naturais para a captura de relacionamentos sintático-semânticos

Bruno Menegola
bmenegola@inf.ufrgs.br

Orientadora: Aline Villavicencio
Colaboradoras: Maria Alice Pimenta Parente e Maity Siqueira

Grupo de Processamento de Linguagem Natural
Instituto de Informática / UFRGS



Informática
UFRGS

Modelagem de linguagens naturais para a captura de relacionamentos sintático-semânticos

- Organização:
 - Introdução e motivação
 - Arquitetura proposta
 - Conclusões e trabalhos futuros

- Como as pessoas encontram as palavras que querem usar?

- Como as pessoas encontram as palavras que querem usar?
- As palavras estão armazenadas no léxico mental

- Como as pessoas encontram as palavras que querem usar?
- As palavras estão armazenadas no léxico mental
- Entretanto o número e a organização das palavras é diferente de um dicionário

- Como as pessoas encontram as palavras que querem usar?
- As palavras estão armazenadas no léxico mental
- Entretanto o número e a organização das palavras é diferente de um dicionário
- Crianças e adultos organizam e recuperam as palavras da mesma forma?
- Diferentes culturas organizam e recuperam as palavras da mesma forma?

- O objetivo é criar modelos computacionais

- O objetivo é criar modelos computacionais
- Foco no desenvolvimento de modelos que capturem relacionamentos sintático-semânticos:
 - Sinonímia
Ex.: criar ↔ produzir
 - Hiperonímia
Ex.: voar ↔ locomover
andar ↔ locomover
 - Hiponímia
Ex.: cortar ↔ picar
cortar ↔ serrar

- Dados provém de pesquisas feitas com crianças e adultos do Brasil e da China realizadas por colaboradores do Instituto de Psicologia da UFRGS

- Dados provém de pesquisas feitas com crianças e adultos do Brasil e da China realizadas por colaboradores do Instituto de Psicologia da UFRGS
- Cada pessoa realizou duas tarefas:
 - Descrever a ação exibida (primeira resposta)
 - Sugerir outro verbo que pudesse substituir o primeiro (segunda resposta)

- Dezesete filmes foram exibidos

- Dezessete filmes foram exibidos

Filme: ballon			
resposta	adultos	resposta	crianças
1	Estourar(35)	1	Estourar(34)
1	Dar(2)	1	Quebrar(2)
1	Bater(1)	1	Esmagar(1)
1	Furar(1)	1	Fazer(1)
2	furar(11)	1	Pegar(1)
2	explodir(9)	2	bater(2)
2	dar(5)	2	estourar(9)
2	bater(3)		
2	estourar(2)		
2	pressionar(2)		
2	tirar(2)		
2	esmagar(1)		
2	esvaziar(1)		
2	fazer(1)		
2	rasgar(1)		
2	romper(1)		

Pain3

resposta	adultos
----------	---------

1	Esfarelar(12)
---	---------------

1	Quebrar(11)
---	-------------

1	Esmigalhar(6)
---	---------------

1	Despedaçar(4)
---	---------------

1	Partir(2)
---	-----------

1	Desfarelar(1)
---	---------------

1	Desmanchar(1)
---	---------------

1	Destruir(1)
---	-------------

1	Tirar(1)
---	----------

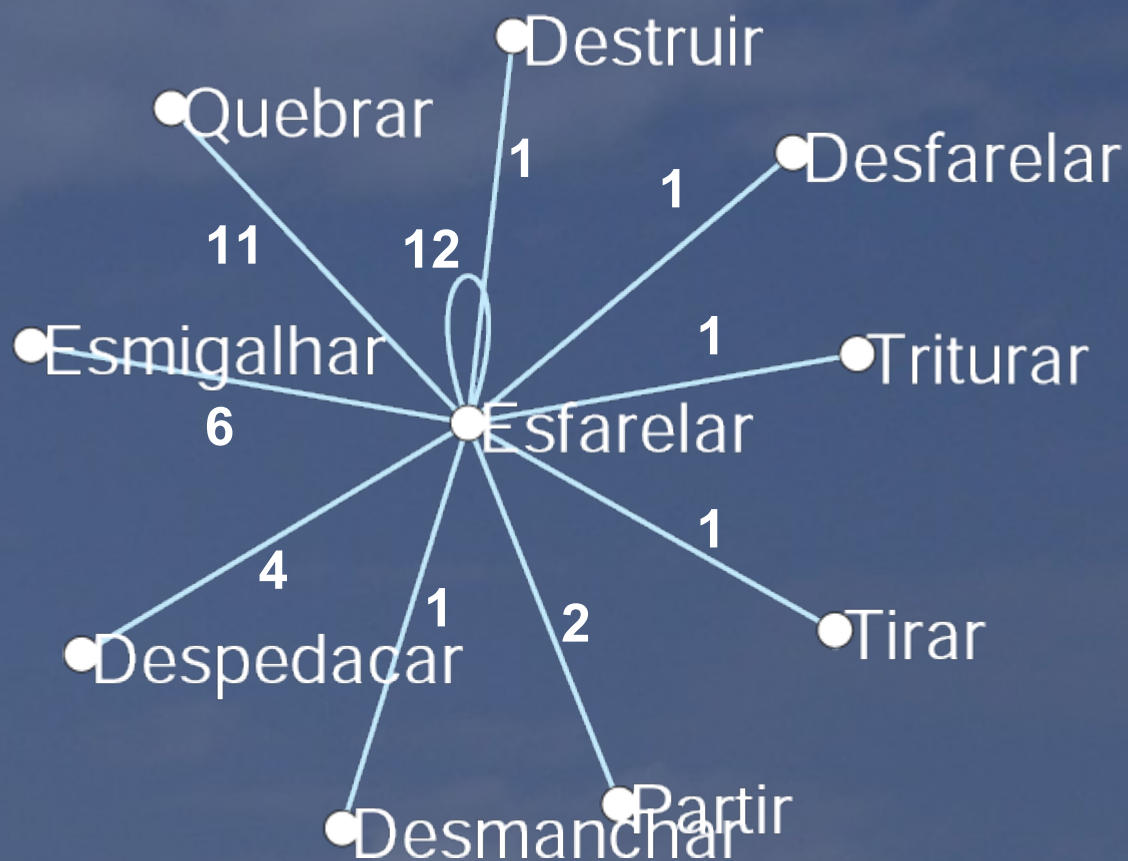
1	Triturar(1)
---	-------------

● Esfarelar

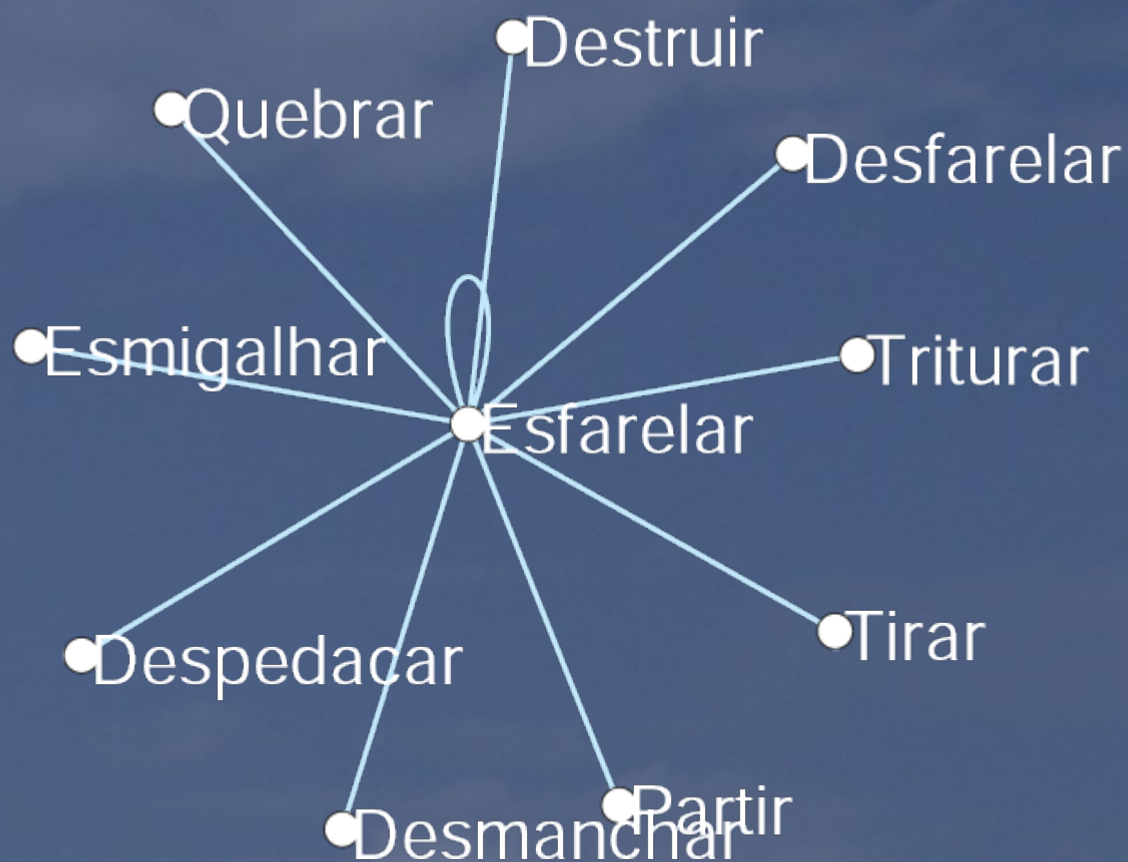
resposta	Pain3 adultos
1	Esfarelar(12)
1	Quebrar(11)
1	Esmigalhar(6)
1	Despedaçar(4)
1	Partir(2)
1	Desfarelar(1)
1	Desmanchar(1)
1	Destruir(1)
1	Tirar(1)
1	Triturar(1)

12
● Esfarelar

resposta	Pain3 adultos
1	Esfarelar(12)
1	Quebrar(11)
1	Esmigalhar(6)
1	Despedaçar(4)
1	Partir(2)
1	Desfarelar(1)
1	Desmanchar(1)
1	Destruir(1)
1	Tirar(1)
1	Triturar(1)

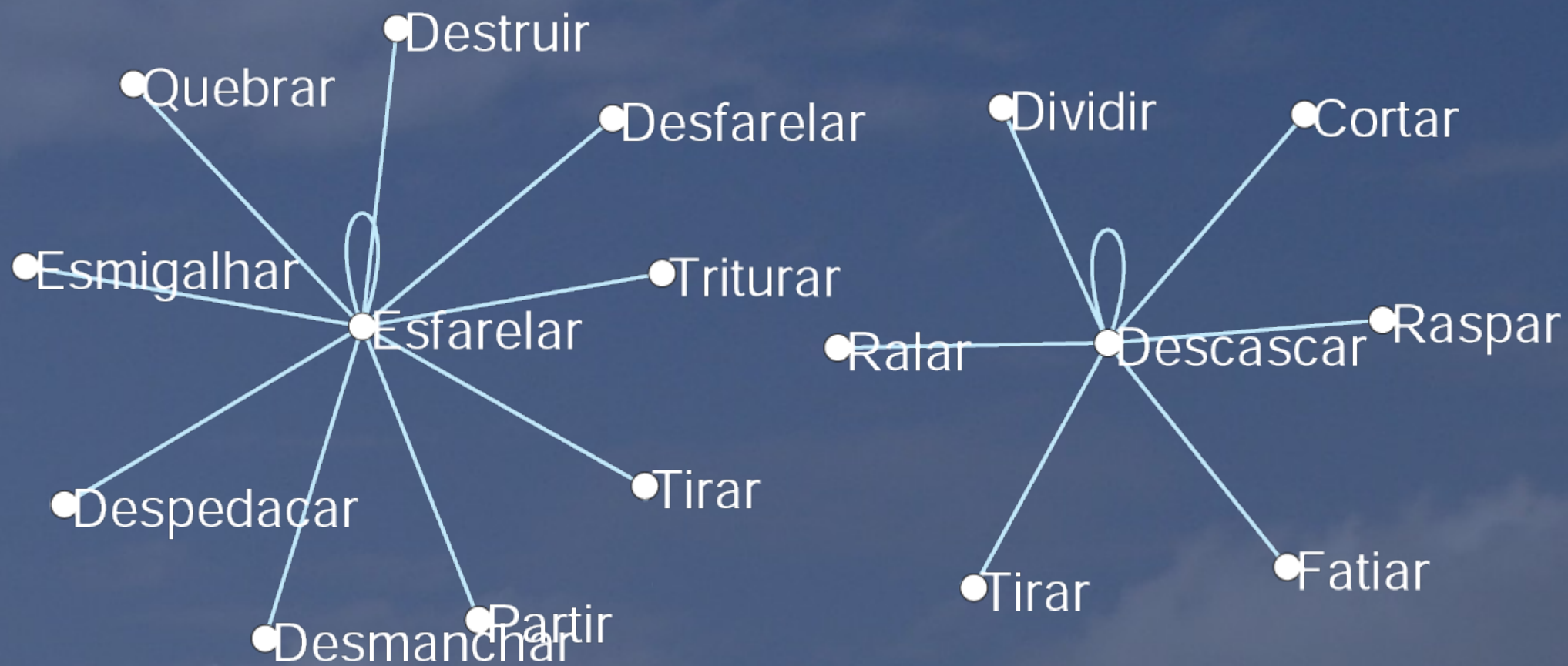


resposta	Pain3 adultos
1	Esfarelar(12)
1	Quebrar(11)
1	Esmigalhar(6)
1	Despedaçar(4)
1	Partir(2)
1	Desfarelar(1)
1	Desmanchar(1)
1	Destruir(1)
1	Tirar(1)
1	Triturar(1)

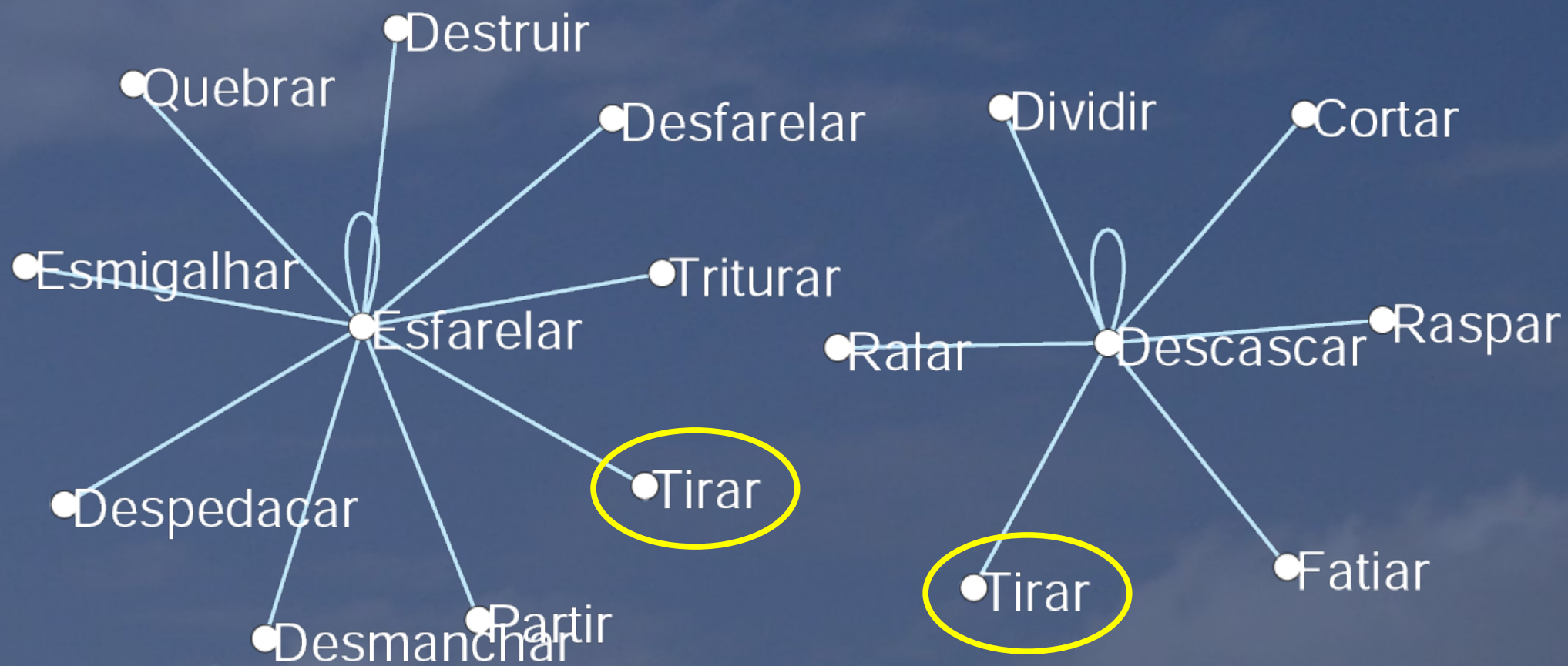


resposta	Pain3 adultos
1	Esfarelar(12)
1	Quebrar(11)
1	Esmigalhar(6)
1	Despedaçar(4)
1	Partir(2)
1	Desfarelar(1)
1	Desmanchar(1)
1	Destruir(1)
1	Tirar(1)
1	Triturar(1)

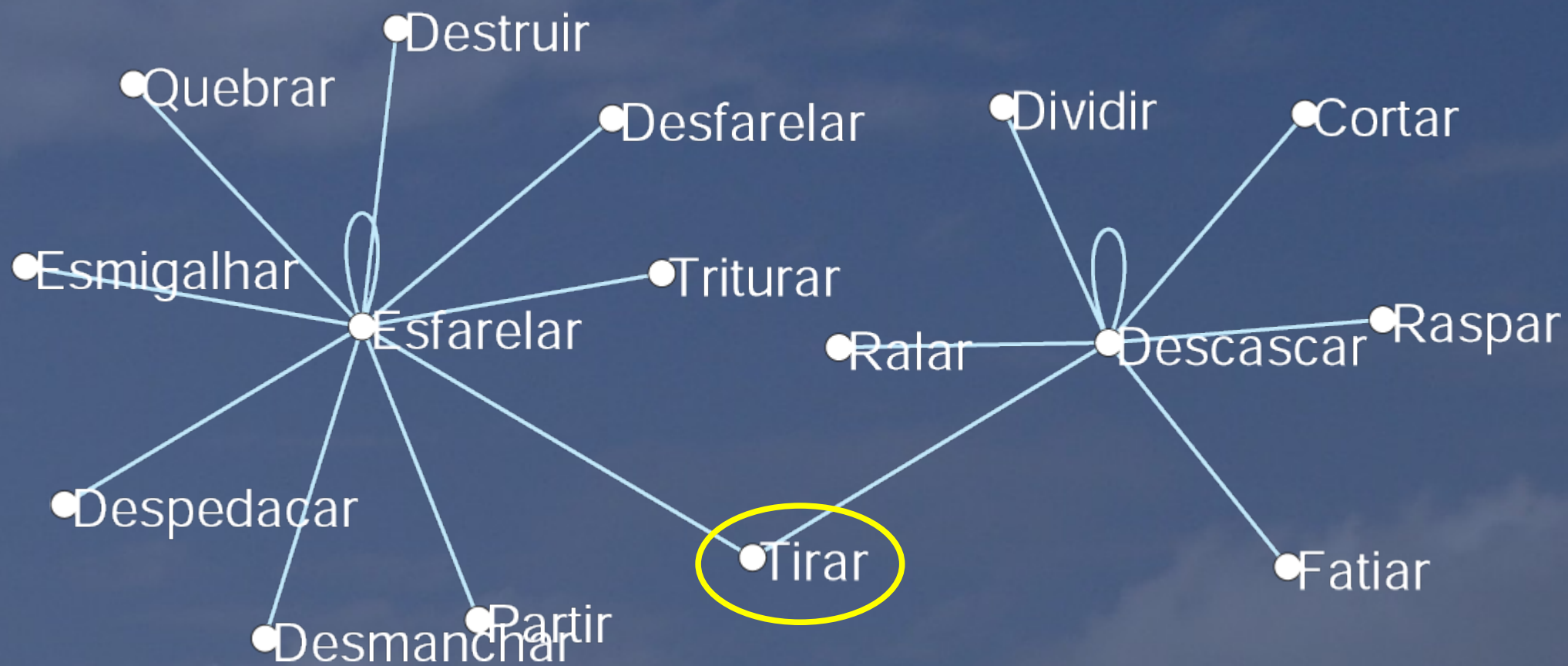
Grafos



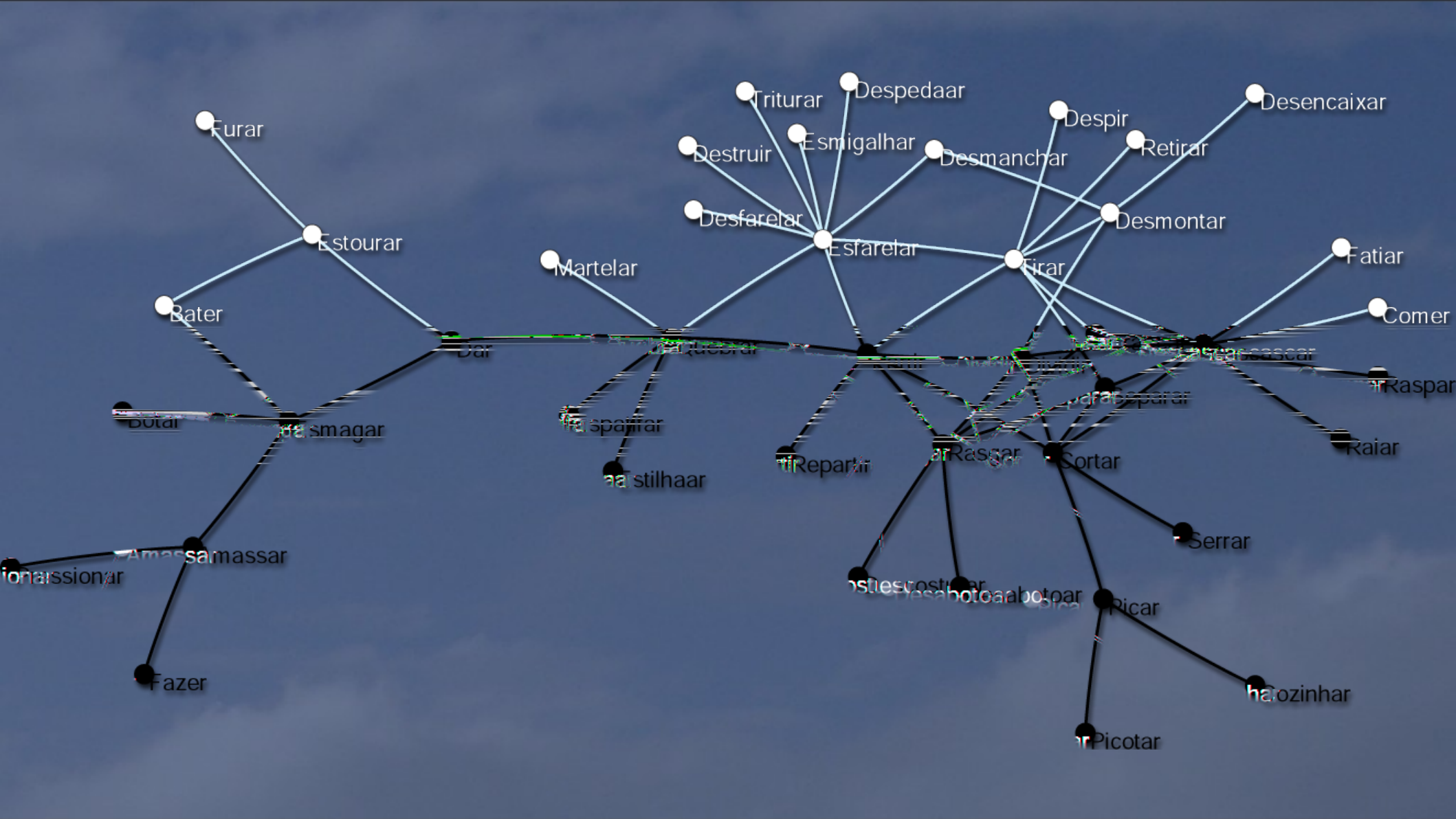
Grafos



Grafos



Grafos



Tabelas

Verbo	Número de ligações	Citações	Freq. Total (670)	Freq. tomate (36)	Freq. arbore (36)	Freq. ballon (39)	...
Amassar	15	2	45	7			
Arrancar	3	3	5		2		
Bater	2	2	2	1			1
Botar	1	1	1	1			
Comer	1	1	1				
Cortar	17	5	58				
Cozinhar	1	1	1				
Dar	3	3	6	3			2
Desabotoar	1	1	2				
Descascar	32	4	105		13		
Descosturar	1	1	5				
Desencaixar	1	1	1				
Desfarelar	1	1	1				
Desmanchar	2	2	3				
...							

Tabelas

Verbo	Número de ligações	Citações	Freq. Total (670)	Freq. tomate (36)	Freq. arbore (36)	Freq. ballon (39)	...
Amassar	15	2	45	7			
Arrancar	3	3	5		2		
Bater	2	2	2	1			1
Botar	1	1	1	1			
Comer	1	1	1				
Cortar	17	5	58				
Cozinhar	1	1	1				
Dar	3	3	6	3			2
Desabotoar	1	1	2				
Descascar	32	4	105		13		
Descosturar	1	1	5				
Desencaixar	1	1	1				
Desfarelar	1	1	1				
Desmanchar	2	2	3				
...							

Tabelas

Verbo	Número de ligações	Citações	Freq. Total (670)	Freq. tomate (36)	Freq. arbore (36)	Freq. ballon (39)	...
Amassar	15	2	45	7			
Arrancar	3	3	5		2		
Bater	2	2	2	1			1
Botar	1	1	1	1			
Comer	1	1	1				
Cortar	17	5	58				
Cozinhar	1	1	1				
Dar	3	3	6	3			2
Desabotoar	1	1	2				
Descascar	32	4	105		13		
Descosturar	1	1	5				
Desencaixar	1	1	1				
Desfarelar	1	1	1				
Desmanchar	2	2	3				
...							

Tabelas

Verbo	Número de ligações	Citações	Freq. Total (670)	Freq. tomate (36)	Freq. arbore (36)	Freq. ballon (39)	...
Amassar	15	2	45	7			
Arrancar	3	3	5		2		
Bater	2	2	2	1			1
Botar	1	1	1	1			
Comer	1	1	1				
Cortar	17	5	58				
Cozinhar	1	1	1				
Dar	3	3	6	3			2
Desabotoar	1	1	2				
Descascar	32	4	105		13		
Descosturar	1	1	5				
Desencaixar	1	1	1				
Desfarelar	1	1	1				
Desmanchar	2	2	3				
...							

Tabelas

Verbo	Número de ligações	Citações	Freq. Total (670)	Freq. tomate (36)	Freq. arbore (36)	Freq. ballon (39)	...
Amassar	15	2	45	7			
Arrancar	3	3	5		2		
Bater	2	2	2	1			1
Botar	1	1	1	1			
Comer	1	1	1				
Cortar	17	5	58				
Cozinhar	1	1	1				
Dar	3	3	6	3			2
Desabotoar	1	1	2				
Descascar	32	4	105		13		
Descosturar	1	1	5				
Desencaixar	1	1	1				
Desfarelar	1	1	1				
Desmanchar	2	2	3				
...							

- Indica o quanto uma palavra é freqüente e utilizável

- Indica o quanto uma palavra é freqüente e utilizável

$$P(x) = \frac{freq(x)}{\sum_{y \in G} freq(y)}$$
$$con(x) = \sum_{y \in G} \delta(x, y)$$
$$gen(x) = P(x) * con(x)$$

- Onde x e y são nós do grafo G , $P(x)$ indica a freqüência de x independente de filme e $con(x)$ indica o grau de x

Generalidade

Verbo	gen(x)	Número de ligações	Citações	Freq. Total (671)
tirar	0,8450	7	10	81
cortar	0,6781	7	7	65
dividir	0,4769	8	8	40
fazer	0,2504	7	7	24
separar	0,2399	7	7	23
partir	0,1610	4	5	27
abrir	0,1386	3	4	31

...

Verbo	gen(x)	Número de ligações	Citações	Freq. Total (671)
...				
desmembrar	0,0030	1	1	2
desmontar	0,0030	1	1	2
estilhaçar	0,0030	1	1	2
prensar	0,0030	1	1	2
achatar	0,0015	1	1	1
afinar	0,0015	1	1	1
danificar	0,0015	1	1	1
desagregar	0,0015	1	1	1
...				

$$P(x) = \frac{freq(x)}{\sum_{y \in G} freq(y)}$$

$$con(x) = \sum_{y \in G} \delta(x, y)$$

$$gen(x) = P(x) * con(x)$$

Generalidade

Verbo	gen(x)	Número de ligações	Citações	Freq. Total (671)
tirar	0,8450	7	10	81
cortar	0,6781	7	7	65
dividir	0,4769	8	8	40
fazer	0,2504	7	7	24
separar	0,2399	7	7	23
partir	0,1610	4	5	27
abrir	0,1386	3	4	31

...

Verbo	gen(x)	Número de ligações	Citações	Freq. Total (671)
...				
desmembrar	0,0030	1	1	2
desmontar	0,0030	1	1	2
estilhaçar	0,0030	1	1	2
prensar	0,0030	1	1	2
achatar	0,0015	1	1	1
afinar	0,0015	1	1	1
danificar	0,0015	1	1	1
desagregar	0,0015	1	1	1
...				

$$P(x) = \frac{freq(x)}{\sum_{y \in G} freq(y)}$$

$$con(x) = \sum_{y \in G} \delta(x, y)$$

$$gen(x) = P(x) * con(x)$$

Generalidade

Verbo	gen(x)	Número de ligações	Citações	Freq. Total (671)
tirar	0,8450	7	10	81
cortar	0,6781	7	7	65
dividir	0,4769	8	8	40
fazer	0,2504	7	7	24
separar	0,2399	7	7	23
partir	0,1610	4	5	27
abrir	0,1386	3	4	31

...

$$P(x) = \frac{freq(x)}{\sum_{y \in G} freq(y)}$$

$$con(x) = \sum_{y \in G} \delta(x, y)$$

$$gen(x) = P(x) * con(x)$$

Verbo	gen(x)	Número de ligações	Citações	Freq. Total (671)
...				
desmembrar	0,0030	1	1	2
desmontar	0,0030	1	1	2
estilhaçar	0,0030	1	1	2
prensar	0,0030	1	1	2
achatar	0,0015	1	1	1
afinar	0,0015	1	1	1
danificar	0,0015	1	1	1
desagregar	0,0015	1	1	1
...				

Convencionalidade

- Indica quanto uma palavra é preferida entre outras para descrever um determinado evento

Convencionalidade

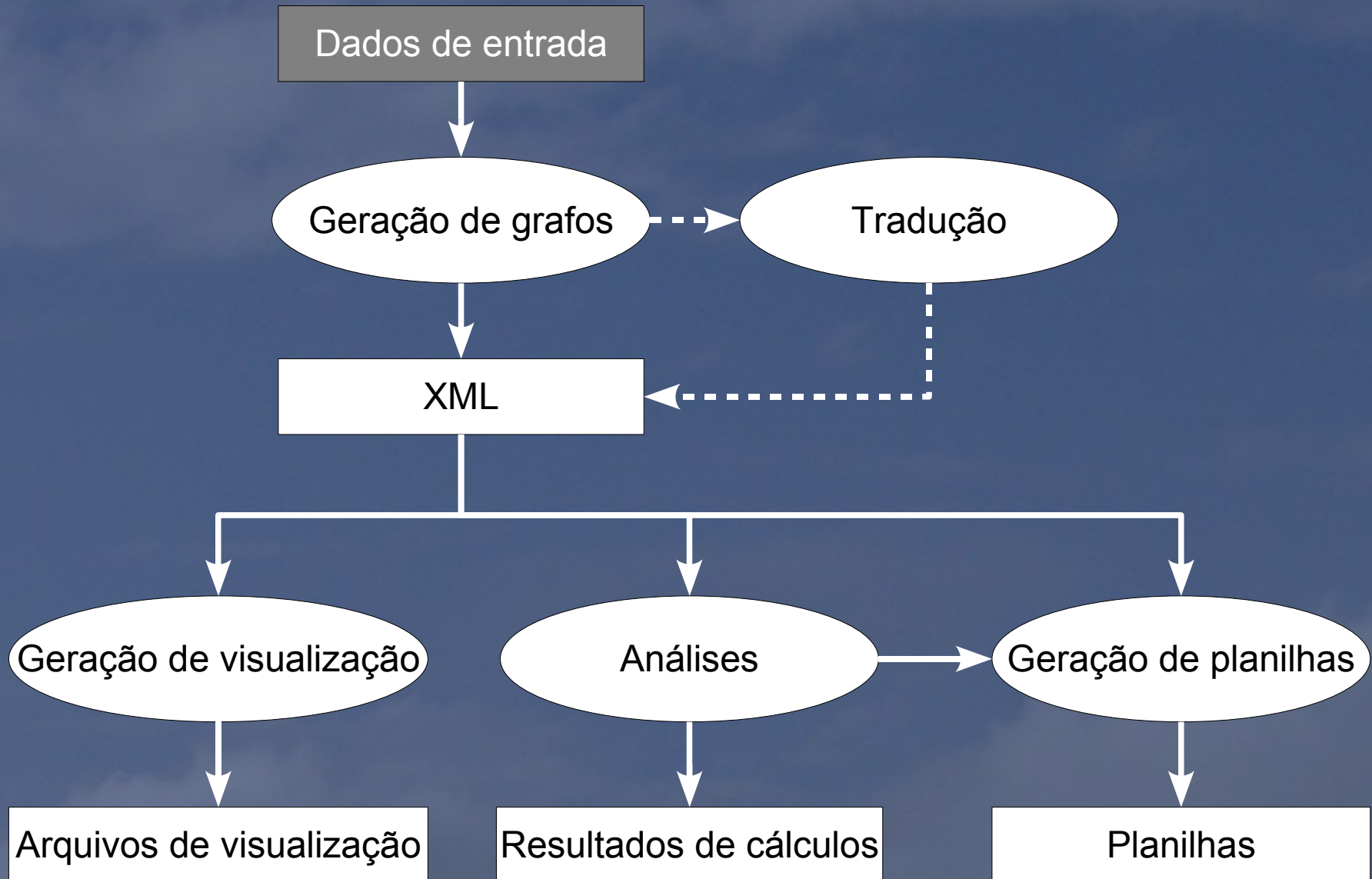
- Indica quanto uma palavra é preferida entre outras para descrever um determinado evento
- Sendo $x|f$ a palavra x considerando apenas suas ocorrências no filme f :

$$conv(x|f) = \frac{freq(x|f)}{\sum_{y \in G} freq(y|f)}$$

Verbo	Freq. legos (49)
Desmontar	35
Dividir	10
Desmanchar	2
Desencaixar	1
Tirar	1

Verbo	Freq. Chemise (39)
abrir	13
estragar	6
descosturar	6
rasgar	6
destruir	2
inutilizar	2
arrancar	1
romper	1
partir	1
fazer	1

Arquitetura Proposta



- O modelo possibilita realizar pesquisas sobre aquisição da linguagem

- O modelo possibilita realizar pesquisas sobre aquisição da linguagem
- Esta pesquisa permite um melhor entendimento dos mecanismos e algoritmos de aprendizado necessários para a aquisição da linguagem

- Os resultados preliminares até então obtidos sugerem que:

- Os resultados preliminares até então obtidos sugerem que:
 - Adultos utilizam muito mais palavras específicas que as crianças

- Os resultados preliminares até então obtidos sugerem que:
 - Adultos utilizam muito mais palavras específicas que as crianças
 - Os falantes de Mandarim possuem um léxico maior que os de Português

- Os resultados preliminares até então obtidos sugerem que:
 - Adultos utilizam muito mais palavras específicas que as crianças
 - Os falantes de Mandarim possuem um léxico maior que os de Português
 - Por essa razão as crianças chinesas também utilizam verbos mais específicos que as brasileiras.

Conclusões – Trabalhos futuros

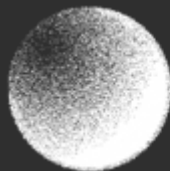
- Dar continuidade com dados coletados de crianças de ambas as comunidades dois anos após o estudo inicial

Conclusões – Trabalhos futuros

- Dar continuidade com dados coletados de crianças de ambas as comunidades dois anos após o estudo inicial
- Esse modelo deve capturar a mudança gradual da organização lexical

Obrigado!

Bruno Menegola
bmenegola@inf.ufrgs.br



Informática
UFRGS