

## Introdução ao Processamento de Línguas Naturais

SCC5908 *Introdução ao Processamento de Língua Natural*

Thiago A. S. Pardo

1

## Recapitulando...

- Abordagens superficiais vs. profundas
- Simbolismo vs. estatística
- Racionalismo (gerativismo) vs. empirismo (estruturalismo, distribucionalismo)
  - Dominância atual do empirismo, trabalhos com base em *cópus* e em evidência linguística
  - Análises e modelos estatísticos, frequências de fenômenos textuais

2

## Abordagens: PLN

- Exemplo: livros de Tom Sawyer (de Mark Twain)

Word	Freq.	Use
the	3332	determiner (article)
and	2972	conjunction
a	1775	determiner
to	1725	preposition, verbal infinitive marker
of	1440	preposition
was	1161	auxiliary verb
it	1027	(personal/expletive) pronoun
in	906	preposition
that	877	complementizer, demonstrative
he	877	(personal) pronoun
I	783	(personal) pronoun
his	772	(possessive) pronoun
you	686	(personal) pronoun
Tom	679	proper noun
with	642	preposition

Tokens = 71.370  
Types = 8.018 (poucas para um texto tão grande)  
→ para crianças

Taxa type/token = 0,11 (11%)

Em geral, quanto maior o *cópus*, menor a taxa

3

## Abordagens: PLN

- Distribuição de palavras
  - Lei de Zipf
    - *George Kingsley Zipf*
    - Baseada em trabalho de Estoup (1916)
    - Proveniente do "Princípio do Mínimo Esforço", publicado no livro *Human Behavior and the Principle of Least Effort* (1949)

4



**Abordagens: PLN**

- Curva de Zipf

freqüência

Palavras ordenadas por freqüência (da maior para a menor)

9

**Abordagens: PLN**

- Distribuição de palavras
  - Curva de Zipf e corte de Luhn (1958)
    - Busca por termos importantes

freqüência

Palavras ordenadas por freqüência (da maior para a menor)

Palavras ordenadas por freqüência (da maior para a menor)

Corte superior

Corte inferior

Pico das palavras relevantes

(I) Curva de Zipf

(II) Curva de Zipf e Cortes Luhn

Soares et al. (2008)

**Abordagens: PLN**

- Distribuição de palavras
  - Curva de Zipf e corte de Luhn (1958)
    - Busca por termos importantes

freqüência

Palavras ordenadas por freqüência (da maior para a menor)

preposições, conjunções, etc.

termos raros

Corte superior

Corte inferior

Pico das palavras relevantes

Palavras ordenadas por freqüência (da maior para a menor)

(I) Curva de Zipf

(II) Curva de Zipf e Cortes Luhn

Soares et al. (2008)

**Abordagens: PLN**

- Distribuição de palavras
  - Curva de Zipf e corte de Luhn (1958)
    - Busca por termos importantes

freqüência

Palavras ordenadas por freqüência (da maior para a menor)

preposições, conjunções, etc.

termos raros

Corte superior

Corte inferior

Pico das palavras relevantes

Palavras ordenadas por freqüência (da maior para a menor)

Pontos de cortes arbitrários, definidos empiricamente

(I) Curva de Zipf

(II) Curva de Zipf e Cortes Luhn

Soares et al. (2008)

## [ Abordagens: PLN ]

- Distribuição de palavras
  - Outra lei de Zipf
    - O número de significados de uma palavra é correlacionado com sua frequência
      - Palavra com 10.000 ocorrências → 2.1 significados
      - Palavra com 5.000 ocorrências → 3 significados
      - Palavra com 2.000 ocorrências → 4.6 significados

13

## [ Abordagens: PLN ]

- Distribuição de palavras
  - Ainda outras leis de Zipf
    - Uma palavra de conteúdo tende a ocorrer próxima a outra ocorrência sua
    - A frequência de uma palavra é inversamente proporcional ao seu tamanho
    - Quanto maior a frequência de uma palavra, mais "permutações" há (em seus componentes morfológicos)

14

## [ Abordagens: PLN ]

- Leis de Zipf
  - Exageradamente valorizadas
    - Não deveriam ser "leis", mas "observações" aproximadas
  - Até alguns eventos aleatórios obedecem essas leis
    - Forma de gerar os dados, de construir a curva

15

## [ PLN ]

- Classificação
  - Recursos
  - Ferramentas
  - Aplicações

## [ Recursos ]

- **Cópus**
  - Anotação: humana e/ou automática
    - XML, XCES, TEI, etc.
  - Paralelo, comparável, alinhado, etc.
- **Dicionários monolíngues e bilíngues**
  - *Machine readable* vs. *machine tractable*
- **Léxicos**
  - Vários paradigmas

17

## [ Ferramentas ]

- Segmentadores textuais: palavras (*tokenizador*), sentenças, parágrafos, tópicos
- Stemmers, lematizadores, nominalizadores
- Etiquetadores morfossintáticos (*taggers*)
- Analisadores sintáticos *shallow* (*chunkers*) e *deep* (*parsers*)
- Analisadores semânticos e discursivos
- Alinhadores textuais: lexicais, sentenciais, etc.
- Concordanceadores, *word counting*, etc.

18

## [ Aplicações ]

- Tradutores automáticos
- Revisores ortográficos e gramaticais
- Ferramentas de auxílio à escrita
- Sumarizadores automáticos
- Simplificadores textuais

19

## [ Recursos, ferramentas e aplicações ]

- **Atenção**
  - Classificação difusa, às vezes
  - Dependente do uso
    - Sumarizador como passo intermediário para recuperação da informação → ferramenta
    - Dicionário eletrônico para consulta → aplicação

20

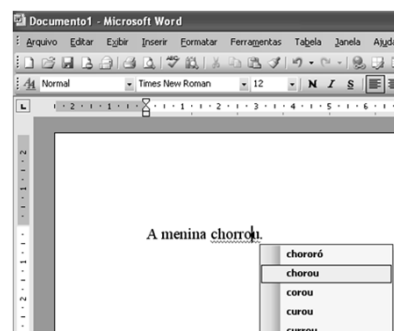
## PLN e áreas correlatas

- Limites entre PLN e outras áreas: como percebem isso?
  - Recuperação de informação
  - Extração de informação
  - Inteligência artificial
  - Banco de dados
  - Interação humano-computador
  - Tradução automática
  - Tradução
  - Mineração de textos
  - Linguística de cópús

21

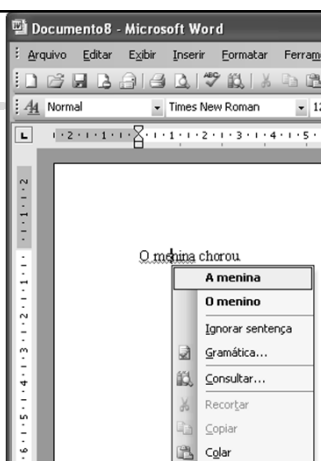
## Exemplos

- Revisão ortográfica
  - Tokenizador
  - Léxico
  - Regras para ordenar sugestões



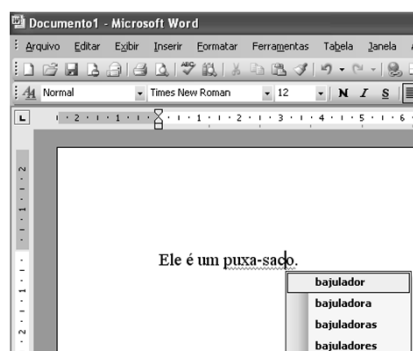
## Exemplos

- Revisão gramatical
  - Tokenizador
  - Segmentador sentencial
  - Etiquetador morfossintático
  - Analisador sintático
  - Léxico
  - Regras gramaticais



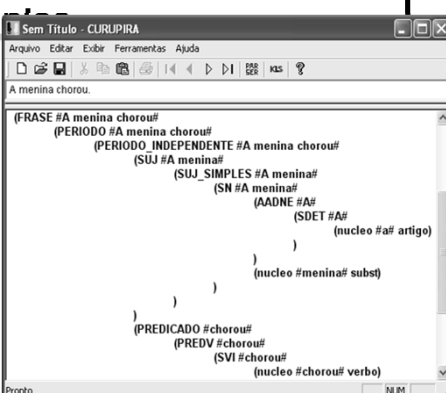
## Exemplos

- Revisão estilística
  - Tokenizador
  - Regras estilísticas
  - ...

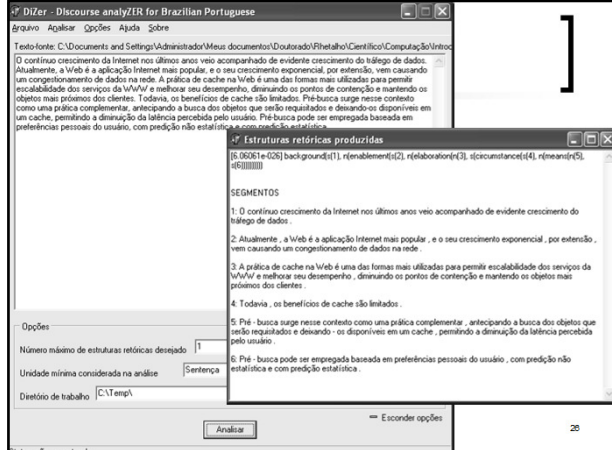


**Exemplos**

- Análise sintática
  - Léxico
  - Regras sintáticas
  - ...

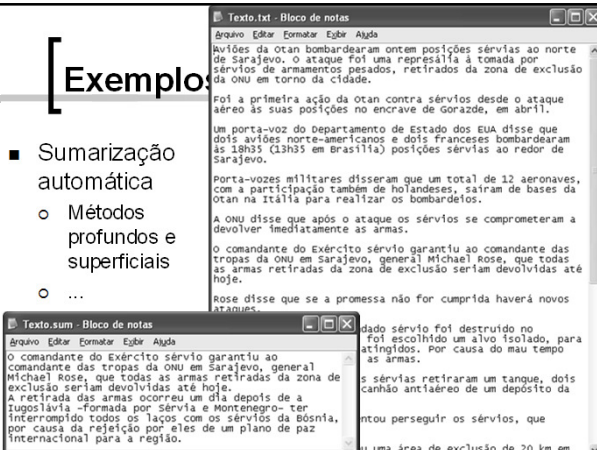


**Exemplos**



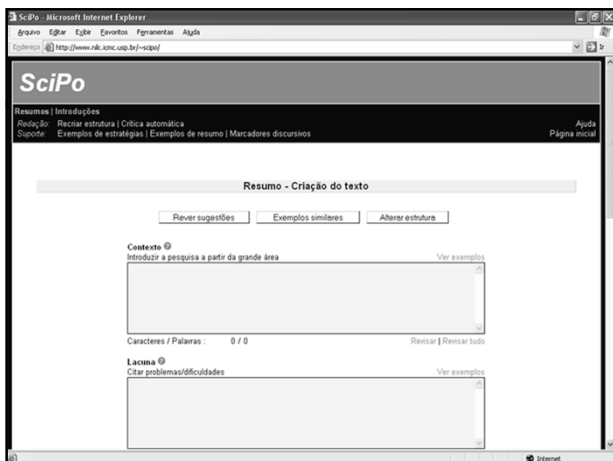
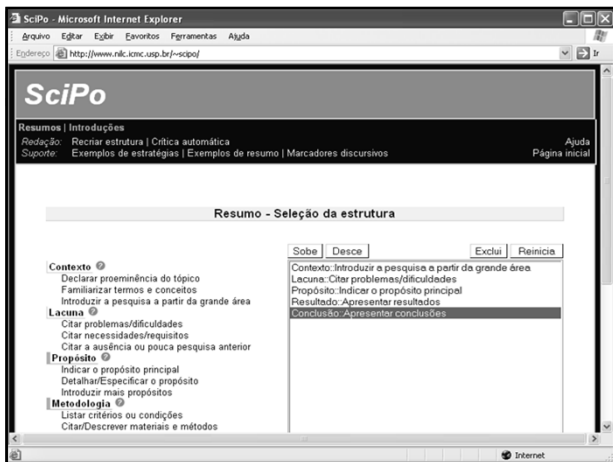
**Exemplos**

- Sumarização automática
  - Métodos profundos e superficiais
  - ...



**Exemplos**

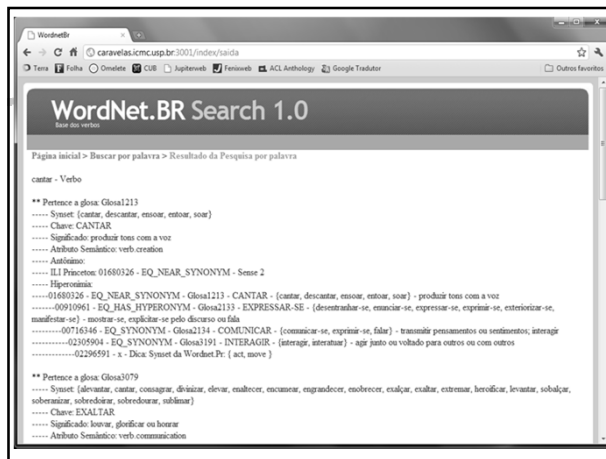
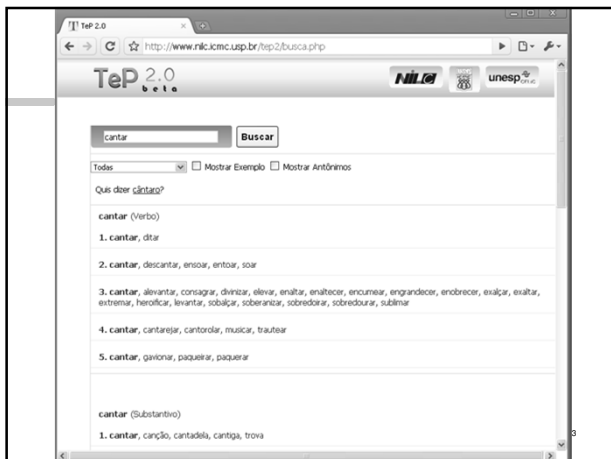
- Auxílio à escrita de textos científicos
  - Regras de estruturação textual
  - Exemplos da estruturas de outros textos
  - Crítica de cada parte do texto



## [ Exemplos ]

- WordNet
  - Base de dados lexicais e conceituais
  - Relações entre palavras
    - Sinonímia
    - Antonímia
    - Acarretamento
    - Etc.
  - Relações ontológicas

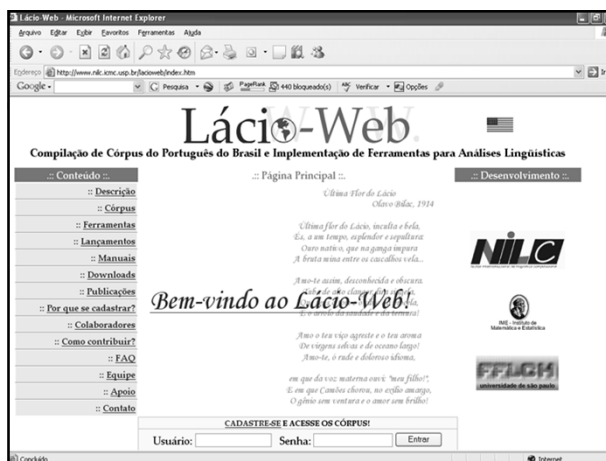




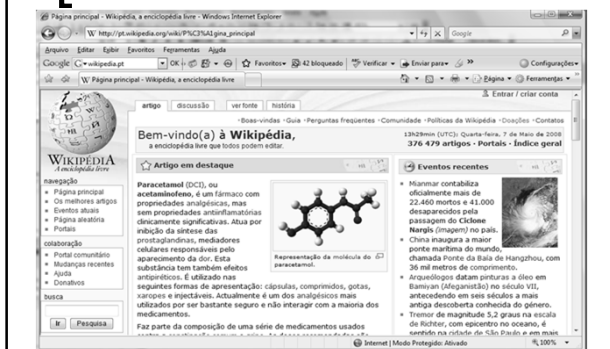
[ PLN ]

- Conhecimento linguístico é a base para muitos sistemas que manipulam língua natural
  - Extração de conhecimento de córpus
    - Regras gramaticais, sintáticas e discursivas
    - Estrutura textual
    - Regras de tradução
    - Critérios para resumir

35



## Conhecimento de mundo



The screenshot shows the Portuguese Wikipedia homepage. The main heading is "Bem-vindo(a) à Wikipédia," followed by the date "12h27min (UTC), Quarta-feira, 7 de maio de 2008" and the article count "376 479 artigos". Below this, there are sections for "Artigo em destaque" (highlighted article) and "Eventos recentes" (recent events). The highlighted article is about Paracetamol (acetaminofen), and the recent events list include the 2008 Sichuan earthquake and the 2008 Sichuan earthquake.

## Senso comum



The screenshot shows the Open Mind Common Sense website. The logo features a lightbulb and the text "PEN MIND common sense no Brasil". Below the logo, there is a message: "Ensinando ao computador as coisas que todos nós sabemos". The main content is a challenge titled "Resultado do Desafio OMC S Netat 2007" and "O que é Open Mind Common Sense? Junte-se a nossa comunidade no Orkut". There is a login form with fields for "Login" and "Senha" and a "Enviar" button. A link "Clique aqui para cadastrar-se no Open Mind Common Sense!" is also visible.

## PLN no Brasil

- Poucos grupos de pesquisa no país
  - São Carlos
  - Porto Alegre
  - Rio de Janeiro
  - Outros?

39

## Recentemente

- A área de PLN tem crescido no Brasil
  - Tecnologia da Informação
  - Google
  - Comissão especial da SBC
  - Eventos científicos próprios melhores e maiores a cada ano
    - Além dos eventos típicos de IA
  - Iniciativas internacionais importantes

40

## Comissão Especial de PLN

- Responsável pela condução da área e representação nacional
- [www.sbc.org.br/ce-pln](http://www.sbc.org.br/ce-pln)
  - Aproximadamente 200 membros na lista de discussão
  - Não precisa ser membro da SBC

41

Comissão Especial de Processamento de Linguagem Natural - Windows Internet Explorer

http://www.nllc.cnpq.br/cepln/

### Comissão Especial de Processamento de Linguagem Natural

**Principal**  
Comissão  
Regimento  
Eventos  
Periódicos  
Fóruns  
Novidades

A criação da Comissão Especial de Processamento de Linguagem Natural (CE-PLN) foi aprovada durante o XXVII Congresso da Sociedade Brasileira de Computação (realizado no Rio de Janeiro-RJ em Junho/Julho de 2007) por pedido da Profª. Dra. Maria das Graças V. Nunes (da Universidade de São Paulo - USP/São Carlos), Rosana Vieira (da Pontifícia Universidade Católica do Rio Grande do Sul - PUC-RS) e Vera L. Strube de Lima (da Pontifícia Universidade Católica do Rio Grande do Sul - PUC-RS), que representavam a comunidade de PLN. A comissão reúne associados com interesses comuns na área de PLN.

A área de Processamento da Linguagem Natural (PLN), também denominada Linguística Computacional ou, ainda, Processamento de Línguas Naturais, lida com problemas relacionados à automação da interpretação e da geração da língua humana em aplicações como Tradução Automática, Sumarização Automática de Textos, Ferramentas de Apoio à Escrita, Perguntas e Respostas, Categorização Textual, Recuperação e Extração de Informação, entre muitas outras, além das tarefas relacionadas de criação e disponibilização de dicionários léxicos e corpora eletrônicos, desenvolvimento de taxonomias e ontologias, investigações em linguística de corpus, desenvolvimento de sistemas de marcação e extração de conhecimento linguístico-computacional, resolução analítica, análise morfosintática automática, análise semântico-discursiva automática, etc.

Em seus processos, e no desenvolvimento de recursos, ferramentas e aplicações, a área tem uma forte interação interdisciplinar, principalmente com as áreas de Linguística e Ciência da Informação, e no Brasil tem suas raízes na área de Inteligência Artificial.

O cenário gerado com a Internet e a demanda por serviços e produtos de Tecnologia da Informação tem ampliado ainda mais o campo de atuação do pesquisador desta área e impulsionado o mercado de trabalho.

O objetivo da CE-PLN é promover e representar a área de PLN no Brasil, apoiando e realizando eventos científicos, propondo e organizando meios de publicação e divulgação para a área e gerenciando listas e fóruns de discussão, dentre outras medidas.

STIL 2011, 24-26 October ...

www.ufmt.br/stil2011/

### The 8th Brazilian Symposium in Information and Human Language Technology

October 24-26 | Mato Grosso, Brazil

**Welcome to STIL 2011!**

STIL (formerly known as TIL - Workshop on Information and Human Language Technology) is the bi-annual Language Technology event supported by the Brazilian Computer Society (SBC) and by the Brazilian Special Interest Group on Natural Language Processing. More details about the event and its history are available at [www.sbc.com.br/stil](http://www.sbc.com.br/stil)

In 2011 it will take place at Cuiabá, Mato Grosso, Brazil. The conference has a multidisciplinary nature and covers a broad spectrum of disciplines related to Human Language Technology, such as Linguistics, Computer Science, Psychology, Information Science, among others. It aims at bringing together both academic and industry participants that work on those areas.

**Topics of Interest**

The STIL 2011 welcomes research work in human language technology in general (and not only Portuguese) in various fields. Topics of interest include, but are not limited to:

- \* Natural Language Processing Applications.
- \* Natural Language Resources & Tools.
- \* User Studies and Evaluation Methods.
- \* Corpus Linguistics.
- \* Phonology/Morphology, Tagging and Chunking, Word Segmentation.
- \* Terminology, Lexicology and Lexicography.
- \* Lexical Semantics - Grammar Formalisms, Syntax and Parsing.
- \* Semantics, Semantic Representations and Semantic Parsing.

**Mapa para o Hotel Holiday Inn Cuiabá**

A: UFMT  
B: Holiday Inn Cuiabá

International Conference on Computational Processing of the Portuguese Language

**propor 2012**

17 TO 20 APRIL 2012  
COIMBRA, PORTUGAL

**Home**  
Program  
Call for papers  
Call for demos  
PhD & MSc/MA  
Dissertation contest  
Author information  
Important dates  
Committees  
Grants  
Venue  
Satellite events  
Registration  
Accommodation  
Contacts  
Propor 2014

**PROPOR 2012**

The International Conference on Computational Processing of Portuguese, former Workshop on Computational Processing of the Portuguese Language - PROPOR - is the main event in the area of Natural Language Processing that is focused on Portuguese and the theoretical and technological issues related to this specific language.

The meeting has been a very rich forum for the interchange of ideas and partnerships for the research communities dedicated to the automatic processing of the Portuguese language. PROPOR brings together research groups in the area, promoting the development of methodologies, linguistic resources and projects that can be shared among all researchers and practitioners in the field.

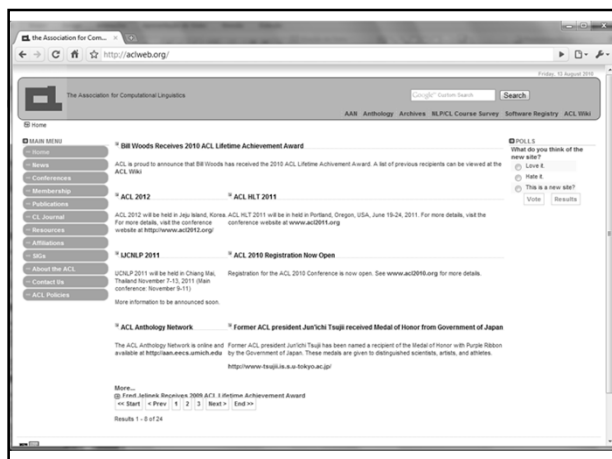
PROPOR, a bi- or tri-annual event, is hosted in Brazil and in Portugal. The meetings have been held in Lisbon, PT (1993); Curitiba, BR (1998); Porto Alegre, BR (1999); Évora, PT (1999); Oeiras, BR (2000); Faro, PT (2002); Bahia, BR (2005); Aveiro, PT (2008); www.propor.org.br (2010)

**Invited Speakers**

Robert Berwick (MIT)  
Paul Biscarra (U. Amsterdam)

**News**

22-March-2012  
Preliminary Program is now available.  
22-February-2012  
A Call for Dissertations is open for the PhD and MSc/MA Dissertation Contest.  
27-January-2012



## Outras iniciativas

- ACL ([aclweb.org](http://aclweb.org))
  - *ACL anthology*, listas de discussão, wiki
  - *Registry of Latin American Researchers in Natural Language Processing and Computational Linguistics*
- Linguateca ([www.linguateca.pt](http://www.linguateca.pt))
  - Oficialmente finalizado
- forum-lp
- Eventos correlatos
  - Escola Brasileira de Linguística Computacional
  - Encontro de Linguística de Corpus
  - Workshop de Descrição do Português
    - Junto ao STIL
- Toolkits
  - GATE, NLTK, Giza++ e Moses, AntMover, etc.

46