MAKE IT SIMPLE WITH PARAPHRASES:
AUTOMATED PARAPHRASING FOR AUTHORING AIDS AND MACHINE TRANSLATION

Anabela Barreiro

PhD Defense in Linguistics
25 May 2009

Advisors:
Professor Belinda Mary Harper Sousa Maia
Faculdade de Letras da Universidade do Porto

Professor Adam Meyers
New York University
**General Motivation**

1. MT changed the world of translation  
   it can no longer be ignored or underestimated  

2. It is increasingly more used and useful  
   it often replaces human translation when the client requires gisting  

3. However, good quality MT is still an ambitious goal  
   after +50 years, MT still eludes researchers and developers who  
   are constantly challenged to provide better translations  

How can MT move forward?
**Machine Translation Challenges**

1. homography and common-noun nuance
2. anaphora + distant referential associations
3. ellipsis
4. extra-sentential and extra-textual information + extra-linguistic knowledge
5. lexical divergences, idioms, etc.
6. named entities
7. long sentences and wordiness
8. unusual word order
9. *multiword expressions* (including *support verb* constructions)
Support verb construction

= multiword or complex predicate consisting of a semantically weak verb (the support verb), and a predicate noun, a predicate adjective, or a predicate adverb.

Predicate nouns can be:

- morphologically related to a verb
  
  PT: fazer uma apresentação de N = apresentar N
  
  EN: pay a visit to N = to visit N

- autonomous
  
  PT: fazer um mestrado - *mestrar
  
  EN: have fun - *to fun
Why Support Verb Constructions?

- Abundant in language and indispensable to communicate

**Assessment Test**: Search pattern in corpora (COMPARA)

\[(\text{dar.V+inf} + \text{tomar.V+inf} + \text{pôr.V+inf} + \text{fazer.V+inf} + \text{ter.V+inf}) \langle \text{Modif}\rangle? \langle \text{N} \rangle\]

Selected the 1st 100 sentences for each verb and manually annotated the 500 sentences as SVC or NO SVC

<table>
<thead>
<tr>
<th>PT</th>
<th>dar</th>
<th>tomar</th>
<th>pôr</th>
<th>fazer</th>
<th>ter</th>
</tr>
</thead>
<tbody>
<tr>
<td>transl.</td>
<td>give</td>
<td>take</td>
<td>put</td>
<td>make/do</td>
<td>have</td>
</tr>
<tr>
<td>NO SVC</td>
<td>11%</td>
<td>12%</td>
<td>33%</td>
<td>53%</td>
<td>80%</td>
</tr>
<tr>
<td>SVC</td>
<td>89%</td>
<td>88%</td>
<td>77%</td>
<td>47%</td>
<td>20%</td>
</tr>
</tbody>
</table>

**Results**: These verbs occur very frequently in SVCs. Overall, in 64.2% of their occurrence, these verbs are support verbs.

- Extensively and systematically studied within the Lexicon-Grammar Theory

[M. Gross and followers], also in Portuguese [Ranchhod, 1983, 1990] [Baptista, 2005] [Chacoto, 2005] and in contrastive studies [Salkoff, 1990, 1999]

- Often can be replaced by stylistic variants - paraphrases

- Present several degrees of variability (variable - invariable - idiomatic)

- Semantically weak and ambiguous

- They cannot be translated literally
MT OF SUPPORT VERB CONSTRUCTIONS

**METRA3**

Tradução de “Eles querem uma vista de olhos ao museu”

E qual a melhor tradução na sua opinião?

- They are going to give a sight of eyes to the museum
- They want to give an eye sight to the museum
- They want to give a sight of eyes to the museum
- Nenhuma destas traduções

**MT SYSTEMS MISHANDLING SUPPORT VERB CONSTRUCTIONS**

**Ah... That’s better!**

We promptly verify the superior quality of the results in the second case!
**THESIS STATEMENT**

Linguistic knowledge of paraphrases can improve the quality of machine translation

- The conversion of **support verb constructions** into morphossyntactically and/or semantically related **verbs** produces a controlled language that is applicable both to general and domain specific contexts and makes MT more reliable

1. Reduces complexity (= paraphrasing by simplification)
2. Reduces lexical ambiguity (paraphrase conveys the correct interpretation)
3. Facilitates interpretation when the support verb construction is idiomatic
4. Reduces wordiness (sometimes, text quality)
5. Improves translation quality
OBJECTIVES

1. Build a body of lexical, syntactic and semantic knowledge around support verb constructions

2. Apply this linguistic knowledge to paraphrasing

3. Improve machine translation
THEORETICAL BACKGROUND - LINGUISTICS


- **FrameNet** [Fillmore et al., 2002, 2003] – records the information necessary for the representation of argument mapping relations between a support verb and a nominalization.

- **NomBank** [Meyers et al., 2004b, 2004b] - maps syntactic positions in nominalizations to verbal arguments and identify the allowed complements for a nominalization, relating the nominal complements to the arguments of the corresponding verb, including information about support verbs.

- **Lexicon-Grammar Theory** [M. Gross and followers] - theoretical framework adopted in this research
THEORETICAL FRAMEWORK


the basic unit of analysis and description of meaning is the simple or elementary sentence (predicate or head of the sentence and its compulsory arguments)

predicate nouns select both arguments and support verbs

transformations are non-oriented relations of equivalence between sentences

systematic and detailed description of predicates

formal representation of descriptions

suitable for automated text processing (NLP)

applicable to machine translation - LG studies contemplate many languages
RELATED WORK ON PARAPHRASING

- **question answering** – discovering paraphrased answers provides additional evidence that an answer is correct - [Ibrahim et al., 2003], [Paşca, 2003], [Duboué & Chu-Carroll, 2006]

- **information extraction** and **text mining** - paraphrases help text categorization tasks or mapping to texts with similar characteristics - [Ibrahim et al., 2003], [Shinyama et al., 2002] [Shinyama & Sekine, 2003], [Sekine, 2005] [Paşca, 2005], [Paşca & Dienes, 2005]

- **summarization** - the identification of paraphrases allows information across documents to be condensed and helps improve the quality of the generated summaries [McKeown et al., 2002], [Barzilay, 2001, 2003], [Hirao et al., 2004] [Zhou et al., 2006b]

- **natural language generation** – the generation of paraphrases allows the production of more varied and fluent text - [Iordanskaja et al. 1991]

- **machine translation** – paraphrases help create a more fluent translation and are valuable in the evaluation of MT results - [Zhou et al., 2006], [Callison-Burch et al., 2006a, 2006b, 2007 and 2008]
**RELATED WORK ON CONTROLLED WRITING AND STYLISTIC EDITORS/AUTHORING AIDS**

- **Controlled language** - application of human language techniques in an industrial environment. It is used to support technical writers in producing high quality technical documentation by checking spelling, grammar, style and terminology in technical documents (text optimisation). Automated rewriting for controlled language translation

- **Stylistic editing** - focuses on clarity and expression, re-ordering of paragraphs, sentences and words; elimination of verbiage and jargon; simplification of untangle complicated clauses; elimination of inconsistencies, use of stylistic paraphrases (reductions/extensions)

- **Authoring aids** - contemplate mostly spell checking, simple grammar checking (correction of text), writing and morphological variants and synonyms LREC 2008

Some available tools:

- **MULTIDOC** [Haller, 2000]
- **KANT CE Checker** [Mitamura and Nyberg, 2001] [Mitamura et al., 2003] [Rascu, 2006]
- **CLAT** [Schmidt-Wigger, 1998]; [Carl et al., 2002]; [Hernandez and Rascu, 2004]
- **CLOUT** - the Controlled Language Optimized for Machine Translation.
METHODOLOGY

Task 1
Analysis, extraction from **corpora** and formalization of support verb constructions by using information described in **dictionaries** and **grammars** – **NooJ linguistic environment** was the tool used

Task 2
**Paraphrasing** of predicate noun constructions

Phase 1
Development of **resources** and **pre-processing**
Construction of **simple grammars** to recognize and extract SVCs from texts
Definition of **properties in the dictionary** and establishment of **semantic links** between verbs and SVCs + establishment of classes of **paraphrases**
Application of **grammars** with dictionary information to texts
Refinement of **grammars**
**Finer recognition** of SVCs in texts -> **paraphrasing** + **translation**

Phase 2
**Evaluation** experiments
**FORMALIZATION IN THE DICTIONARIES**

Explicit marking of derivation and semantic verb associations

Adjective entries:
- Identification of derivational paradigms for *predicate adverbs - adverbializations* (annotation AVDRV)

```
literal,A+FLX=PRINCIPAL+IN+symb+EN=literal+DRV=AVDRV00:LITERALMENTE
```

Autonomous predicate nouns:
- Identification of *autonomous predicate nouns* (annotation Npred)
- Identification of a *semantically related verb* (annotation VRB)
- Link to the *support verb* (annotation VSUP)

```
curso,N+FLX=ANO+Npred+IN+inst+EN=course+VSUP=tirar+VRB=estudar+NPrep=de+Det=um
```
Explicit marking of derivation and support verb associations

Verb entries:
• Identification of derivational paradigms for **predicate nouns - nominalizations** (annotation \( NDRV \)) and **predicate adjectives** (annotation \( ADRV \))

• Link to the **support verbs** (annotation \( VSUP \)) occurring with the derived **predicate nouns** and **adjectives**

adaptar,V+FLX=FALAR+Aux=1+INOP57+Subset=132+EN=adapt+\( VSUP=fazer \)+\( DRV=NDRV00:CANÇÃO \)
azedar,V+FLX=LIMPAR+Aux=1+OBJTRundif98+Subset=740+EN=sour+\( VSUP=estar \)+\( DRV=ADRV00:ALTO \)
NEW RESOURCES

➢ Port4NooJ

• a set of open source Portuguese linguistic resources representing ontological relations, and paraphrasal relations. This resource integrates a bilingual extension for PT-EN MT.

<table>
<thead>
<tr>
<th>Semantic relations</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb - Predicate Nouns</td>
<td>8472</td>
</tr>
<tr>
<td>Predicate Adjectives - Adverbs</td>
<td>222</td>
</tr>
<tr>
<td><strong>Total Links</strong></td>
<td><strong>8694</strong></td>
</tr>
</tbody>
</table>

• Publicly available resources at:
  http://www.nooj4nlp.net
  http://www.linguateca.pt/Repositorio/Port4Nooj/

➢ DicTUM (Diccionário de Termos e Unidades Multipalavra)

• a Dictionary of Multiword Expressions
NEW TOOLS

- **ReWriter**
  - a monolingual standalone paraphraser to pre-edit texts, using paraphrasing capabilities

- **ParaMT**
  - a bilingual/multilingual paraphraser to be integrated in machine translation systems
USEFULNESS OF THE AFOREMENTIONED PARAPHRASING TOOLS

- Controlled writing & text pre-editing
- Machine translation
- Text production and stylistics (authoring aids)
**REWRITER: A STANDALONE PARAPHRASER**

**CONTROLLED WRITING**

**TEXT PRE-EDITING**

**ELEMENTARY SVC > LEXICAL VERB** – fazer uma amputação = amputar (to amputate)

**ELEMENTARY SVC > NON-ELEMENTARY SVC** - realizar/efectuar uma amputação (to perform an amputation)

---

**Submeter-se/ser submetido a uma operação (to undergo surgery)**

Ser operado (to be operated)
Foram encontradas 26 sugestões para o texto apresentado.

Também gostava de ir [dar aulas | leccionar | ensinar] na universidade, mas não posso dar-me ao luxo de ser colocado agora que estou casado e com um bebê a caminho.

Escute, os administrativos da MLA vão [dar uma festa | festejar] hoje à noite, na suite do último andar.

Resolvi [fazer a experiência | experimentar] e passei cuidadosamente um dedo pela face que se me apresentava;

Catherine correu porque se sentiu como se estivessem a [fazer troça | troçar] dela.

Era preso há dois meses e, embora continuasse fraco e em perigo de uma recada, tive de [fazer uma viagem de | viajar] algumas léguas para chegar à sede do tribunal.

Quanto a Catherine, nem sequer fingIU querer [fazer conversa | falar | conversar].

Disse que Frances nem sequer estava a [pôr a hipótese de | considerar | fazer um aborto | abortar], e que provavelmente ir ter o bebê em Espanha e depois a sua vez em Mercer.

William largou o rolo de cordel e avançou em passo estufado para o caminho de acesso, a acenar e a [fazer gestos | gesticular].

Disse que as escadas eram uma mação, que o senhorinho estava a [fazer planos para | planear] instalar um elevador, o que seria uma vantagem quando Frances fosse lá ficar com o bebê.

Ultimamente começei a [dar ainda mais valor | valorizar] à privacidade e ao anonimato desta casa.

Só tenho pena da minha mãe e do meu pai já não serem vivos e eu não poder levá-los a [dar uma volta | passear] nele.

Pede-me sempre descupa se uma ou outra reação particularmente fortes e saiu uma alugilha me fazer [dar um salto | saltar].

Dois espancamos no joelho, enquanto estive a escrever isto, um dos quais forte ao ponto de me fazer [dar un grito | gritar].

Não querendo [dar a impressão de | parecer] ser racista, engoli a reclamação.

Na verdade, Gordon Masters tinha decidido [dar apoio a | apoiar] Philip para o intercâmbio em Eufória.

Adrian, é perfeitamente capaz de [dar glória a | glorificar] Deus.

Deixou-me então ao objetivo de [pôr termo | terminar] ao racismo e à injustiça, objetivo que prossegui e prosseguirei enquanto viver.

Mas a criada que Miss Kenton deixava a [tomar conta | tratar | cuidar] dele levantou-se ao ver-me e começou a sacudir-lhe o ombro.

O movimento do braço, a [tomar balanço | balançar], desequilibrou-a.

Não há de que [ter medo | recear].

Saí da loja antes de ele [ter tempo de | poder] chamar a brigada de narcóticos.

Decidi tentar-me levantar dessa vez e deslizei os pés para o chão apalpando à procura dos chinelos, antes de [ter chance de | poder] mudar de ideia.

A noite parecia não [ter fim | acabar].

Quando se dirigiu ao gabinete exterior e estendeu o envelope a Jean, esta pareceu [ter dificuldade em | evitar] olhar o nos olhos.

Por agora não era possível olhar para trás e tudo o que podia fazer era seguir em frente e [ter esperança de | esperar] acabar antes da semana seguinte.
ParaMT: a bilingual/multilingual paraphraser for MT

Recognition of Portuguese SVC and translation into English verbs
**EVALUATION EXPERIMENT 1**

5 support verbs

500 baseline sentences - manually annotated

100 for each elementary support verb

<table>
<thead>
<tr>
<th></th>
<th>SVC Recognition Precision</th>
<th>SVC Recognition Recall</th>
<th>SVC Paraphrasing Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pôr</td>
<td>73/73 - 100%</td>
<td>73/100 – 73%</td>
<td>72/73 - 98.6%</td>
</tr>
<tr>
<td>Tomar</td>
<td>75/75 - 100%</td>
<td>75/100 – 75%</td>
<td>68/73 - 93.1%</td>
</tr>
<tr>
<td>Ter</td>
<td>65/65 - 100%</td>
<td>65/100 – 65%</td>
<td>59/65 - 90.7%</td>
</tr>
<tr>
<td>Dar</td>
<td>57/60 - 95%</td>
<td>57/100 – 57%</td>
<td>46/51 - 90.1%</td>
</tr>
<tr>
<td>Fazer</td>
<td>43/45 – 95.5%</td>
<td>43/100 – 43%</td>
<td>40/45 - 88.8%</td>
</tr>
<tr>
<td></td>
<td><strong>98.4%</strong></td>
<td><strong>62.6%</strong></td>
<td><strong>93.4%</strong></td>
</tr>
</tbody>
</table>
5 different support verbs

100 sentences with SVC (for each language PT and EN)

20 sentences for each support verb – manually assigned a paraphrase for each one of them

> 20 SVC sentences + 20 paraphrases of the original SVC sentences

TOTAL: 200 sentences for each language

100 with SVC + 100 with paraphrases of the SVC

10 testers translated and evaluated translations

students of the MA in Translation and Linguistic Services

<table>
<thead>
<tr>
<th></th>
<th>Equal quality result</th>
<th>SVC translation is better</th>
<th>Paraphrase translation is better</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN-PT MT</td>
<td>17%</td>
<td>26%</td>
<td>57%</td>
</tr>
<tr>
<td>PT-EN MT</td>
<td>19%</td>
<td>30%</td>
<td>51%</td>
</tr>
</tbody>
</table>

For both EN-PT and PT-EN, testers considered that the paraphrase translations presented a better quality. The evaluation results clearly confirm that these paraphrases helped the systems produce better MT.
MAIN CONCLUSION

- Linguistic knowledge formalized in this research when applied to a machine translation system improves its output quality.

When SVC were identified and replaced with semantically equivalent or similar verbal expressions as a pre-processing step to translating:

- a 21% improvement was observed in the evaluated quality of the results of PT-EN machine translation
- a 31% improvement in the results of EN-PT machine translation.
FUTURE RESEARCH SUGGESTED BY THIS THESIS

- Enlargement of syntactic-semantic relations between predicates, and other morphosyntactically and semantically related elements
  
  Exs: *apresentar* – *apresentação* – *apresentado*

  *literal* – *literalmente*

- Paraphrasing of various linguistic phenomena for ReWriter extensibility (comprising stylistic variance and controlled language)

- Development and enhancement of ParaMT

- Enhancement of the machine translation model
# Paraphrasing for ReWriter Extensibility

A formal linguistic study of paraphrases as these ones would represent a significant contribution to NLP in general, and to MT in particular.

<table>
<thead>
<tr>
<th>Linguistic Phenomenon</th>
<th>Expression or sentence</th>
<th>Paraphrase</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Adverbial</strong></td>
<td>à volta da órbita</td>
<td>periorbital</td>
</tr>
<tr>
<td></td>
<td><em>around the orbit of the eye</em></td>
<td>periorbital</td>
</tr>
<tr>
<td></td>
<td>de forma interactiva</td>
<td>interactivamente</td>
</tr>
<tr>
<td></td>
<td><em>in an interactive way</em></td>
<td>interactively</td>
</tr>
<tr>
<td><strong>Relative Clause</strong></td>
<td>N0 que têm sido escritos</td>
<td>N0 que foram escritos  &gt;  N0 escritos</td>
</tr>
<tr>
<td></td>
<td><em>N0 that have been written</em></td>
<td><em>N0 that were written</em>  &gt;  <em>N0 written</em></td>
</tr>
<tr>
<td></td>
<td>A velocidade a que se move a luz</td>
<td>A velocidade da luz</td>
</tr>
<tr>
<td></td>
<td><em>The speed to which light moves</em></td>
<td><em>The speed of light</em></td>
</tr>
<tr>
<td></td>
<td>O papel que a Europa tem</td>
<td>O papel da Europa</td>
</tr>
<tr>
<td></td>
<td><em>The role that Europe plays/has</em></td>
<td><em>The role of Europe</em>  =  <em>Europe’s role</em></td>
</tr>
<tr>
<td></td>
<td>As dificuldades que temos</td>
<td>As nossas dificuldades</td>
</tr>
<tr>
<td></td>
<td><em>The difficulties we have</em></td>
<td><em>Our difficulties</em></td>
</tr>
<tr>
<td><strong>If clause</strong></td>
<td>se for necessário</td>
<td>se necessário</td>
</tr>
<tr>
<td></td>
<td><em>if it is necessary</em></td>
<td><em>if necessary</em></td>
</tr>
<tr>
<td><strong>Named Entity</strong></td>
<td>A rainha de Inglaterra</td>
<td>A rainha inglesa</td>
</tr>
<tr>
<td></td>
<td><em>The queen of England</em></td>
<td><em>The British queen</em></td>
</tr>
<tr>
<td><strong>Noun Phrase</strong></td>
<td>O heróico povo português</td>
<td>Os heróicos portugueses</td>
</tr>
<tr>
<td></td>
<td><em>The heroic Portuguese people</em></td>
<td><em>The heroic Portuguese</em></td>
</tr>
</tbody>
</table>
RELEVANT PUBLICATIONS


  http://poloclup.linguateca.pt/Port4NooJ/ResourcesOverview.pdf

REFERENCES


REFERENCES


ACKNOWLEDGMENTS

This research work was funded by Fundação para a Ciência e a Tecnologia (doctoral scholarship SFRH/BD/14076/2003), co-financed by POSI and partly supported by Fundação para a Computação Científica Nacional.

In addition to the doctoral scholarship, I had a ½ month contract from Linguateca to help make Port4NooJ and ReEscreve public.
MAKE IT SIMPLE WITH PARAPHRASES:

AUTOMATED PARAPHRASING FOR AUTHORING AIDS AND MACHINE TRANSLATION

Anabela Barreiro

PHD DEFENSE IN LINGUISTICS
MAY, 25TH 2009

ADVISORS:

PROFESSOR BELINDA MARY HARPER SOUSA MAIA
FACULDADE DE LETRAS DA UNIVERSIDADE DO PORTO

PROFESSOR ADAM MEYERS
NEW YORK UNIVERSITY