ParaMT:

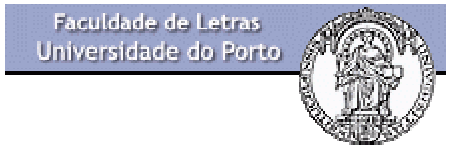
A Paraphraser for Machine Translation

Anabela Barreiro

barreiro_anabela@hotmail.com

FLUP & CLUP-Linguateca

New York University



Motivation

1. MT is a highly desired NLP application
2. MT has eluded researchers and developers for over 50 years
3. MT is increasingly more used and useful

BUT…

4. MT is still far from perfect!

Subject and Motivation

Support verb construction = predicate noun construction

is a multiword expression containing a verb with weak semantic value and a noun which is the predicate of the sentence.

Predicate nouns can be:

- **morphologically related to a verb**

  fazer uma apresentação de = apresentar

  to pay a visit to = to visit

- **autonomous**

  fazer um mestrado - *mestrar

  to have fun - *to fun

Subject and Motivation

Francisco Vieira adianta ainda que está a fazer um esforço no sentido de tomar uma decisão ainda esta semana, definindo se avança ou não para uma candidatura à RTLRS. [CdP]

Translation Engine	Translation Results
FreeTranslation	Francisco Scallop advances even if is it do an effort in the sense of ***take a decision*** still this week, defined advances or not for a candidacy to the RTLRS.
WorldLingo	advances despite he is to make an effort in the direction to still ***take a decision*** this week, defining if he advances or he does not stop a candidacy to the RTLRS.

I can't make a decision about anything these days. [Compara]

Translation Engine	Translation Results
Google	Eu não posso ***fazer a uma decisão*** sobre qualquer coisa estes dias.
Amikai	que eu não posso ***fazer para uma decisão*** sobre qualquer coisa estes dias.
FreeTranslation	Eu não posso ***tomar uma decisão*** sobre algo estes dias.
Babelfish	Eu não posso ***fazer a uma decisão*** sobre qualquer coisa estes dias.
WorldLingo	Eu no posso ***fazer a uma deciso*** sobre qualquer coisa estes dias.
E-Translation Server	Não posso ***tomar uma decisão*** sobre qualquer coisa estes dias.

Main 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

Outcome

- **Port4NooJ**
  - an open source, ontology driven Portuguese linguistic system, which integrates a bilingual extension for Portuguese-English machine translation

- **ReWriter**
  - a monolingual paraphraser to pre-edit texts

- **ParaMT**
  - a bilingual/multilingual paraphraser to be integrated in machine translation systems

Resources

- Port4NooJ - Publicly available at:

  http://www.nooj4nlp.net

  http://www.linguateca.pt/Repositorio/Port4Nooj/

  - Based on:
  - NooJ linguistic environment (http://www.nooj4nlp.net/)
  - OpenLogos English-Portuguese dictionary (http://logos-os.dfki.de/)

    OpenLogos is an open-source derivative of the Logos Machine Translation System

- Data Used
  - COMPARA (http://www.linguateca.pt/COMPARA)
  - METRA (http://www.linguateca.pt/metra)
  - Other corpora

Port4NooJ Dictionaries

Lemma | Part of Speech | Inflectional Paradigm | Syntactic-Semantic Attributes | English Transfer

mesa,N+FLX=CASA+CO+surf+EN=table
cair,V+FLX=ATRAIR+INMO+IntoType+EN=fall
holandês,A+FLX=INGLÊS+AN+lang+EN=Dutch
actualmente,ADV+FLX=FACILMENTE+TEMP+punc+pres+EN=nowadays
alguém,PRO+IMPERS+INDEF+EN=somebody
porque,RELINT+why+EN=why
e,CONJ+JOIN+EN=and
durante,PREP+TEMP+EN=during
cada,DET+IMPERS+INDEF+SG+EN=each
terceiro+NUM+ord+EN=one third

General dictionary sample representing all PoS, variable and invariable forms

a curto prazo,ADV+TEMP+EN=in the short run
a favor de,PREP+CAUS+EN=in favor of
cada um,PRO+INDEF+SG+EN=each one
de quem,INT+ThatType+EN=whose
quem quer que seja,REL+WhateverType+EN=whoever
além disso,CONJ+COOR+EN=besides
um quarto,NUM+frac+EN=one fourth

Sample of invariable compounds in the general dictionary

HIV,N+FLX=PORTUGAL-[redacted]EN=HIV
doença maníaco-depressiva,N+FLX=CASA-[redacted]EN=manic-depressive disorder
doença bipolar,N+FLX=CASA-[redacted]EN=bipolar disorder
asma,N+FLX=CASA-[redacted]EN=asthma

Sample of the dictionary of Biomedical Terms

adro da igreja,N+FLX=MENINO+PL+encl+EN=churchyard
cabo de vassoura,N+FLX=MENINO+COtool+EN=broomstick
bebida alcoólica,N+FLX=CASA+MA+liqu+EN=alcoholic drink+UNAMB
bebida alcoólica,N+FLX=CASA+MA+liqu+EN=booze+slang
cor de laranja,A+NAV+Apred+EN=orange
sul-americano,A+FLX=ALTO+AN+des+EN=South American
a curto prazo,ADV+LocTime+TEMP+EN=in the short run
fora de serviço,ADV+STAT+phr+EN=out of order
há muito tempo,ADV+LocTime+TEMP+puncpast+EN=a long time ago
isto é,CONJ+COOR+EN=i.e.
já não,CONJ+COOR+EN=no longer
mesmo assim,CONJ+SUB+EN=even so
juntamente com,PREP+ASSOC+EN=along[redacted]
à direita de,PREP+Loc+AT+EN=at the right of
em conformidade com,PREP+ALOG+EN=in congruence with

Sample of the dictionary of Terms and Multiword Expressions

DicTUM

Amesterdão,N+PL[redacted]EN=Amsterdam
Estados Unidos da América,N+P[redacted]EN=United States of America
África,N+PL[redacted]EN=Africa
Extremo Oriente,N+PL-[redacted]EN=Far East
Mediterrâneo,N+FLX=ANO+PL-[redacted]EN=Mediterranean
Alpes Peninos,N+FLX=ALPES+PL-[redacted]EN=Pennine Alps
ONU,N+AN[redacted]EN=UN

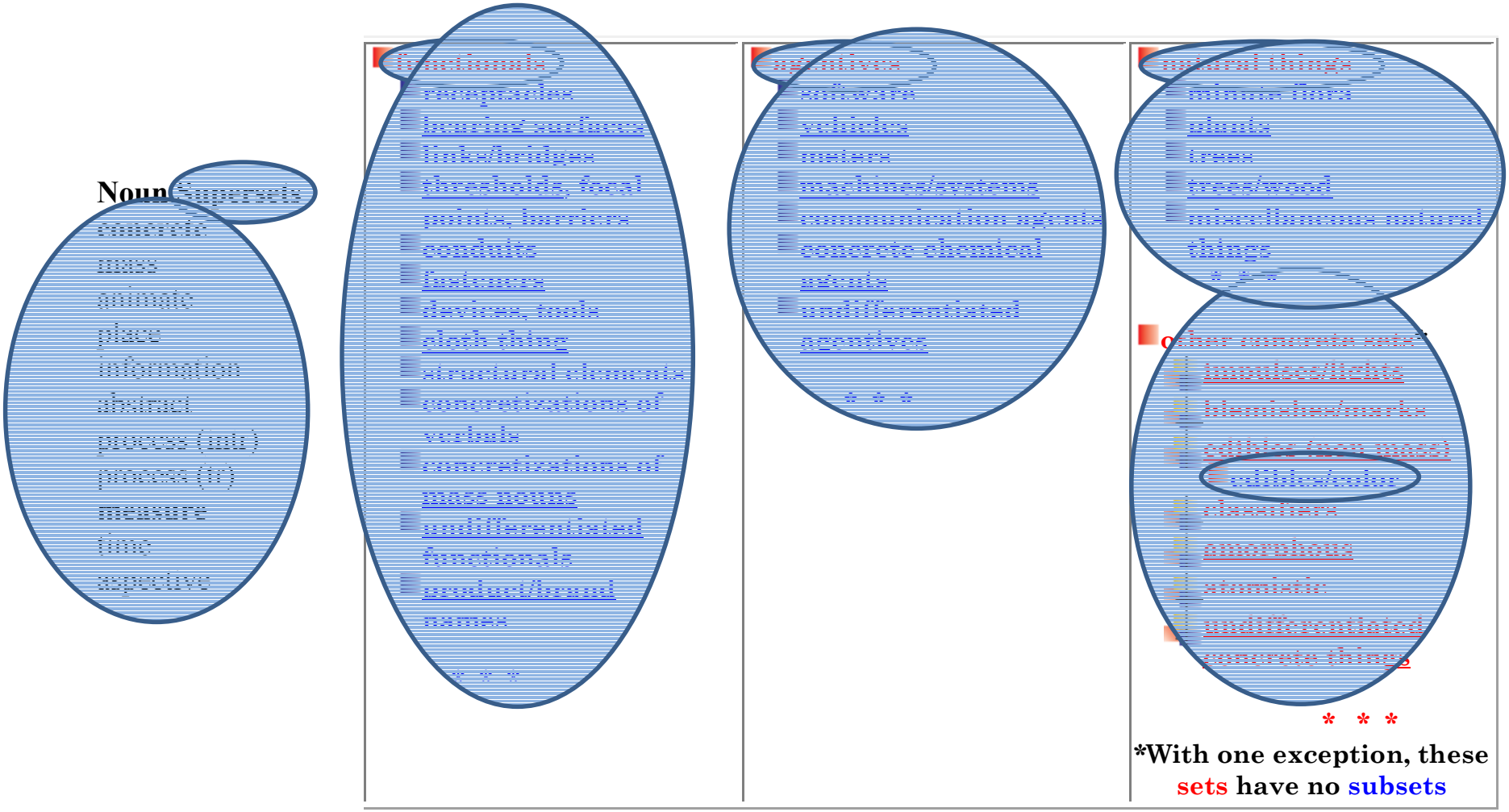
Sample of the dictionary of Proper Names

Syntactic-Semantic Ontology

- Representation abstract language
- Hierarchical taxonomy (sets, supersets and (sometimes) subsets)
- Based on Logos SAL ontology
- Integrated in the dictionary
- It represents both meaning (semantics), and structure (syntax)
- Over 1,000 categories

Syntactic-Semantic Ontology

Sets and Subsets of the CONCRETE Noun Superset
Click on CONCRETE Superset, sets and subsets for explanations



*With one exception, these sets have no subsets

Syntactic-Semantic Ontology

		Examples in English	Examples in Portuguese
agentives	CO+undagt	See subsets	See subsets
software	CO+soft	*routine*	*rotina, ficheiro*
concrete chemical agents	CO+chem	*catalyst, warhead*	*ácido sulfúrico*
machines/systems	CO+mach	*battery, camera*	*máquina fotográfica*
		truck, ship	*automóvel*
meters	CO+meter	*clock, gauge*	*manómetro*
communication agents	CO+comm	*radio, radar*	*rádio*
functionals	CO+undfunc	*trinket, ornament*	*ornamento*
devices/tools	CO+tool	*pliers*	*alicate*
fasteners	CO+fast	*nail, tendon*	*prego*
bearing surfaces	CO+surf	*table, shelf*	*mesa*
receptacles	CO+recp	*bottle, barrel*	*garrafa*
conduits	CO+cond	*chute, artery*	*artéria*
thresholds/focal points/barriers	CO+barr	*wall, door*	*porta*
links/bridges	CO+link	*circuit, nerve*	*circuito*
cloth things	CO+cloth	*shirt, blanket*	*camisola*
structural elements	CO+struc	*spar, bone*	*osso*
concretizations of verbals	CO+verb	*threading*	
concretizations of mass nouns	CO+mass	*acid lining*	
product/brand names	CO+brand	*Windows NT*	*Windows NT*
natural things	CO+nat	See subsets	See subsets
minute flora	CO+flora	*algae, spore*	*alga*
		rose, weed	*erva*
trees	CO+tree	*apple, willow*	*macieira*
trees/wood	CO+trwd	*oak, maple*	*carvalho*
misc. natural things	CO+mnat	*pebble, iceberg*	*iceberg*
edibles (non-mass)	CO+ednm	*pork chop*	*costoleta*
edibles/color	CO+edcol	*orange, cherry*	*laranja*
impulses/lights	Col+ight	*lamp, beam*	*lâmpada*
blemishes/marks	CO+blem	*scratch, freckle*	*sarda*
classifiers	CO+class	*element*	*elemento*
amorphous	CO+amor	*breeze, tide*	*brisa*
atomistic	CO+atom	*electron, atom*	*átomo*

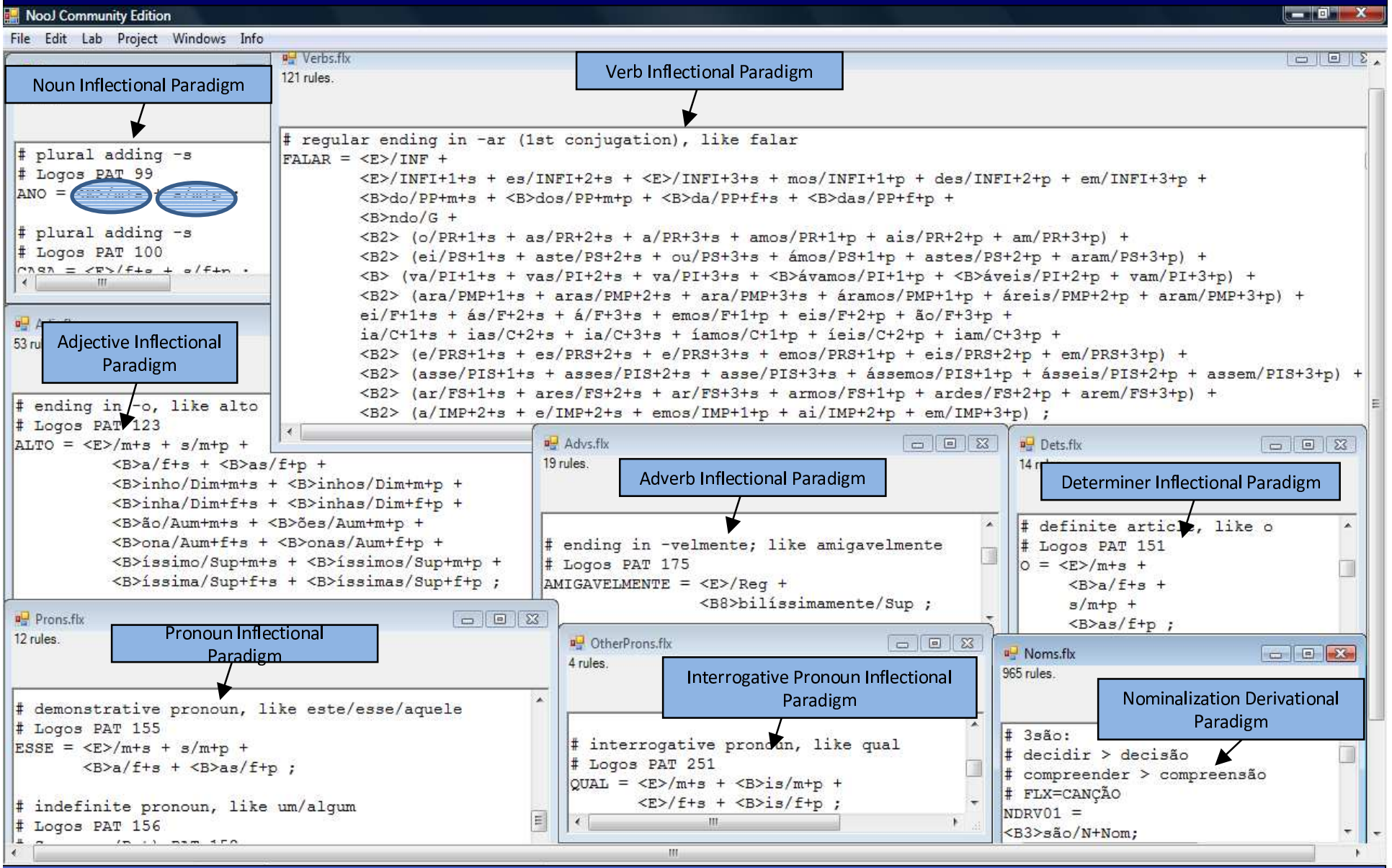
Categories of CONCRETE nouns

Syntactic-Semantic Ontology

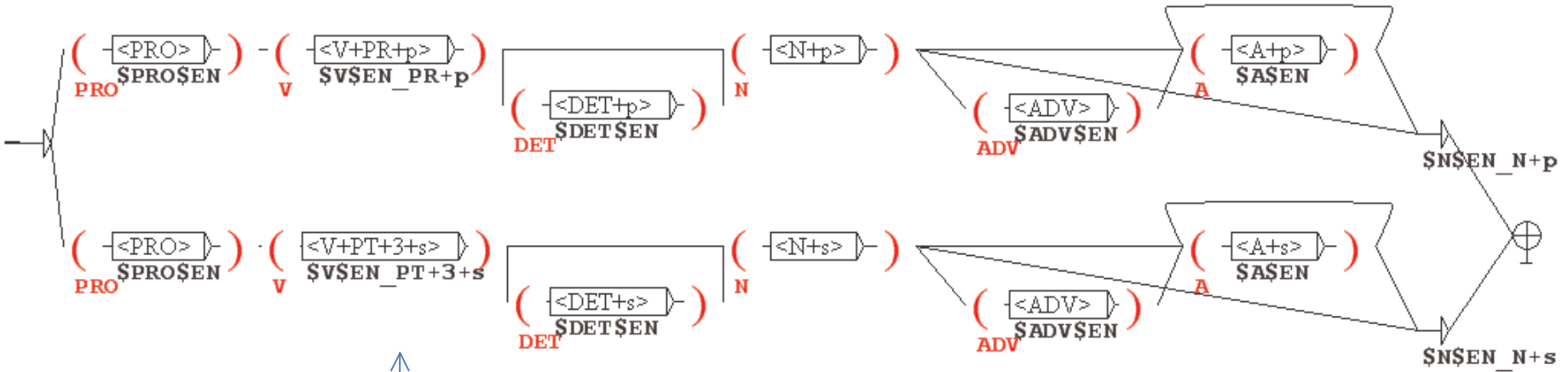
ME - MEASURE Noun Sets and Subsets		
Sets and Subsets	**Mnemonics (= SynSem)**	**Examples**
abstract concepts measured by unit	ME+abs	*humidity, length*
discrete measurable concepts	ME+dis	*sum, increment*
units of measure	ME+unit	See subsets
units of weight	ME+unit+wt	*ounce, pound*
units of velocity	ME+unit+vel	*mph, megahertz*
units of volume measure	ME+unit+vol	*gallon, liter*
units of temperature	ME+unit+temp	*degrees celsius*
units of energy/force	ME+unit+ener	*watt, horsepower*
measurement systems	ME+unit+sys	*fahrenheit, kelvin*
units of duration	ME+unit+dur	*hour, minute, year*
specialized units of measure	ME+unit+spec	*oersted, ohm, phon*
units of money/value	ME+unit+val	*dollar, euro, forint*
units of linear/area measure	ME+unit+lin	*inch, yard, mile*
general undifferentiated measure	ME+undif	*degree, gross, share*

Categories of MEASURE nouns

Inflectional and Derivational Description



Paraphrasing and Translation Grammars



Graph to translate simple sentences

Translation and bilingual paraphrasing of simple sentences

Concordance for Text PTSimpleSentences-3pl.not

Clear Concordance | 4 | characters | tokens | before, and | 5 | after. Display: Inputs | Outputs

Text	Before	Seq.	After
	Elas visitam cidades.	Eles são homen they ar en	. Elas são mulheres simpáticas. Eles
	Elas visitam cidades.	Eles são homens honestos/they are men	. Elas são mulheres simpáticas. Eles
	comida. Eles vendem sapatos.	Ela asas/the ouses	. Elas sonham acordados. Eles sonham
	comida. Eles vendem sapatos.	Elas compram casas/they purchase homes	. Elas sonham acordados. Eles sonham
	comida. Eles vendem sapatos.	Elas compran the	. Elas sonham acordados. Eles sonham
	comida. Eles vendem sapatos.	Elas compram casas/they buy	. Elas sonham acordados. Eles sonham

Query 6/29

Explicit Marking of Derivation and Support Verb

Verb entries:

- Identification of derivational paradigms for **nominalizations** (annotation *NDRV*) and **predicate adjectives** (annotation *ADRV*)
- Link to the derived noun's **support verbs** and to the adjective's **copula verbs** (annotation *VSUP* and annotation *VCOP*)

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+**VCOP=estar**+**DRV=ADRV00**:ALTO

Explicit Marking of Derivation and Semantic Verb Association

Adjective entries:

- Identification of derivational paradigms for **adverbializations** (annotation *AVDRV*)

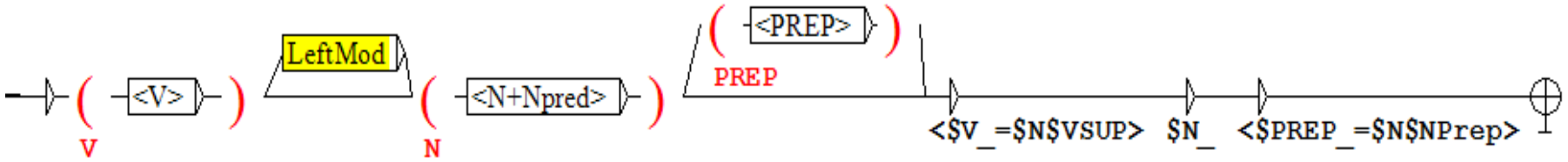
literal,A+FLX=PRINCIPAL+IN+symb+EN=literal+**DRV=AVDRV00**:LITERALMENTE

Autonomous predicate nouns:

- Identification of **autonomous predicate nouns** (annotation *Npred*)
- Identification of a semantically related verb

curso,N+FLX=ANO+**Npred**+IN+inst+EN=course+VSUP=tirar+**VRB=estudar**+NPrep=de+Det=um

ReWriter: a Monolingual Paraphraser



gosto de ver o comboio a	fazer corridas /correr	à velocidade máxima ao lon
o de cheque especial para	fazer doações /doar	às entidades que escolher. A
ores e, quando é preciso ir	fazer filmagens/filmar	fora do estúdio, às vezes fic
e queria trocar de pares e	fazer um jogo /jogar	ao melhor de três sets , mas
dra deu-me um papel para	fazer uma lista de/listar	todas as coisas boas que ex
res foram à caracterização	fazer uns retoques/retocar	, outros estão a descansar n

Recognition and monolingual paraphrasing
of support verb constructions
(support verb construction / morphologically related lexical verb)

ReWriter: Examples

nça, o cirurgião Faivre, ao fazer uma amputação

nça, o cirurgião Faivre, ao fazer uma amputação uma amputação

nça, o cirurgião Faivre, ao fazer uma amputação uma amputação

omista britânico, conseguiu fazer uma transfusão de sangu uma transfusão de sangue , de um cão para outro. A juiza

omista britânico, conseguiu fazer uma transfusão de sangu uma transfusão de sangue , de um cão para outro. A juiza

os pacientes que precisam fazer uma transfusão de sangue uma transfusão de sangue , colaborando também com a c

os pacientes que precisam fazer uma transfusão de sangue uma transfusão de sangue , colaborando também com a c

os pacientes que precisam fazer uma transfusão de sangu ma transfusão de sangue , colaborando também com a c

os pacientes que precisam fazer uma transfusão de sangue uma transfusão de sangue , colaborando também com a c

Elementary SVC > Lexical Verb

Elementary SVC > non-elementary SVC

realizar/efectuar

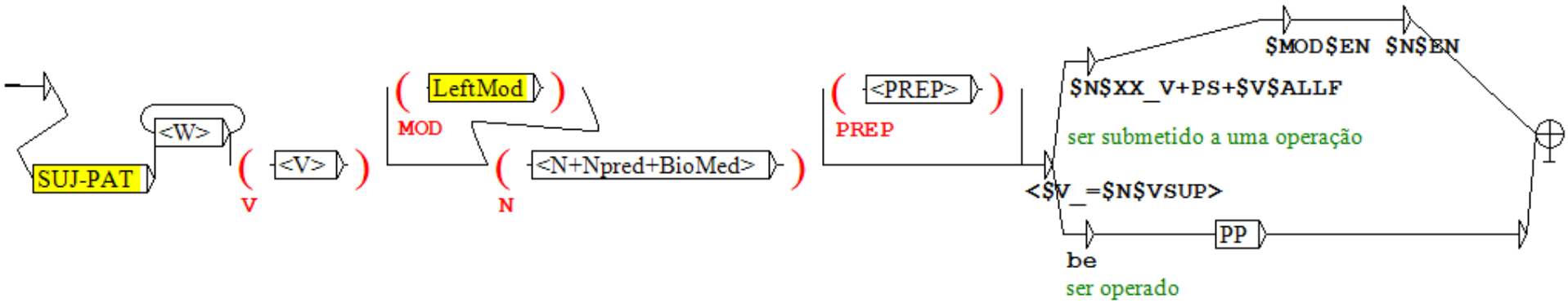
Elementary SVC > *sujeitar-se a*

submeter-se a

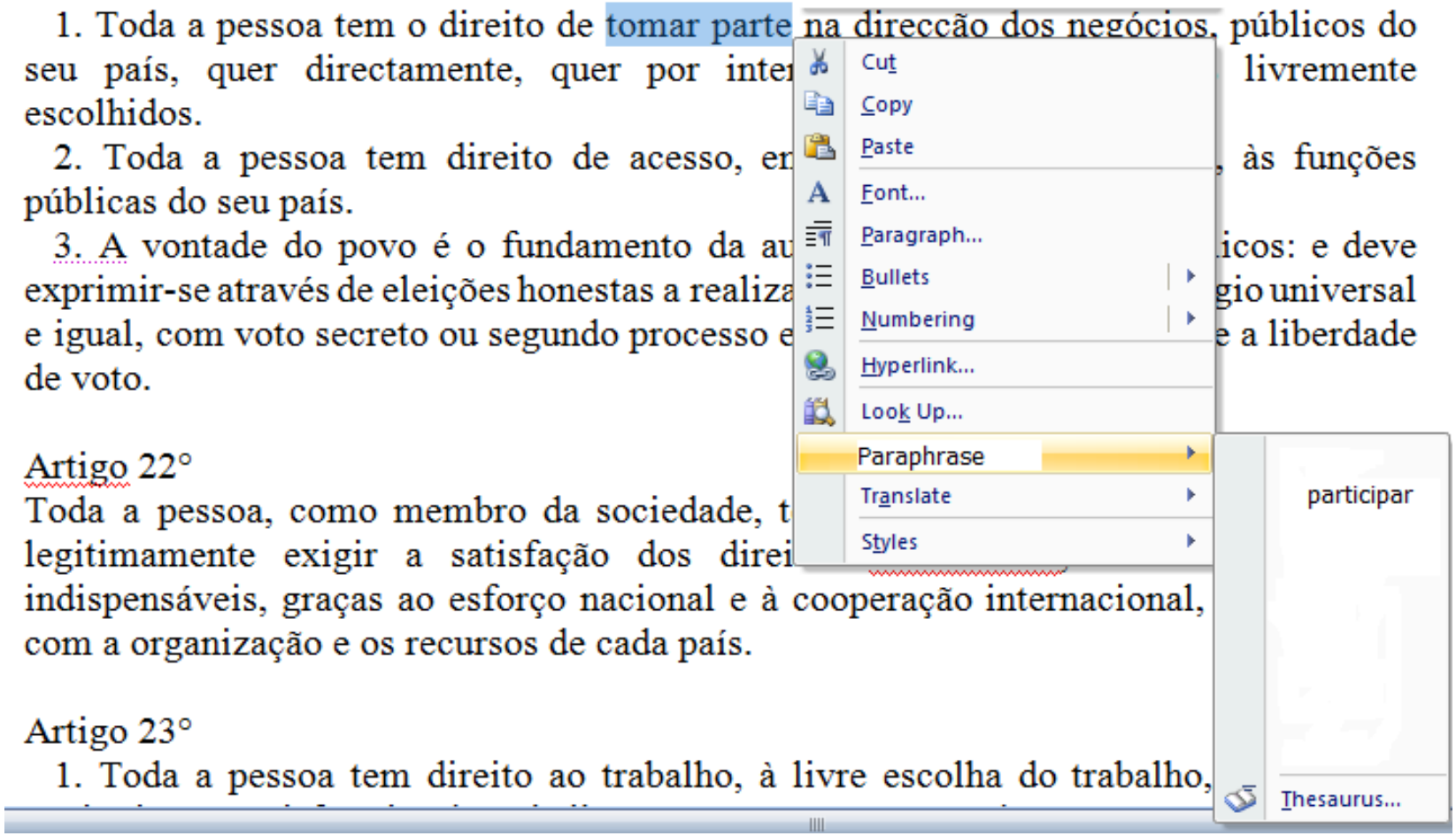
ONLY if the SUBJECT is a patient

Recognition and paraphrasing of elementary support verb constructions co-occurring with predicate nouns of the biomedical field

(support verb construction / lexical verb or stylistic variant / non-elementary support verb construction)



ReWriter: Example of Possible Application



Interactive ReWriter
for word processing applications
such as text editing

ParaMT: a Bilingual/Multilingual Paraphraser

a fazer um estágio para	dar aulas de/teach	religião, mas não se import
m -- os filhos -- juntos e	fizeram a mudança para/change	Johannesburg, e ensinaram
. Necessitava apenas de	ter a certeza de/know	que não escapara à sua
ente hipotética. -- Deves	ter alguma ideia/know	. Dorothy andava a fazer u
. não podemos deixar de	ter cautela/beware	. Pobre Caro, pensou Lync
ra dos chinelos, antes de	ter chance de/can	mudar de idéia. Como pos
ope a Jean, esta pareceu	ter dificuldade em/avoid	olhá-lo nos olhos. Deixou
ao Kiss dela. Apesar de	ter falta de/lack	amor-próprio, isso não sign
igos e imprensa estava a	ter lugar /occur	numa longa galeria com car
guiu ter filhos. -- Tens de	ter mão /control	nessa confusão toda. Sam
spondi, minha mãe deve	ter medo de/fear	cobras. Eu disse no Gabin
da loja antes de ele	ter tempo de/could	chamar a brigada de narcó
a triste aventura havia de	ter um fim/finish	.
Ela ouvira a tia Velma	ter uma discussão com/argue	Jack acerca de mostarda r
de olhos fechados para	ter uma ideia de/know	como seria ser cego e
ter paciência.» «Voltei a	ter uma imensa vontade de/want	viver. A conversa parecia

Recognition and bilingual paraphrasing of support verb constructions
(Portuguese support verb construction / corresponding English verb)

Preliminary Quantitative Results

500 sentences

100 for each elementary support verb

	SVC Recognition Precision	SVC Recognition Recall	SVC Paraphrasing Precision
Pôr	73/73 - 100%	73/100 – 73%	72/73 - 98.6%
Tomar	75/75 - 100%	75/100 – 75%	68/73 - 93.1%
Ter	65/65 - 100%	65/100 – 65%	59/65 - 90.7%
Dar	57/60 - 95%	57/100 – 57%	46/51 - 90.1%
Fazer	43/45 – 95.5%	43/100 – 43%	40/45 - 88.8%
Average	62.6/63.6 - **98.4%**	62.6/100 - **62.6%**	57/61 - **93.4%**

Evaluation of recognition and paraphrasing
of support verb constructions

Conclusions

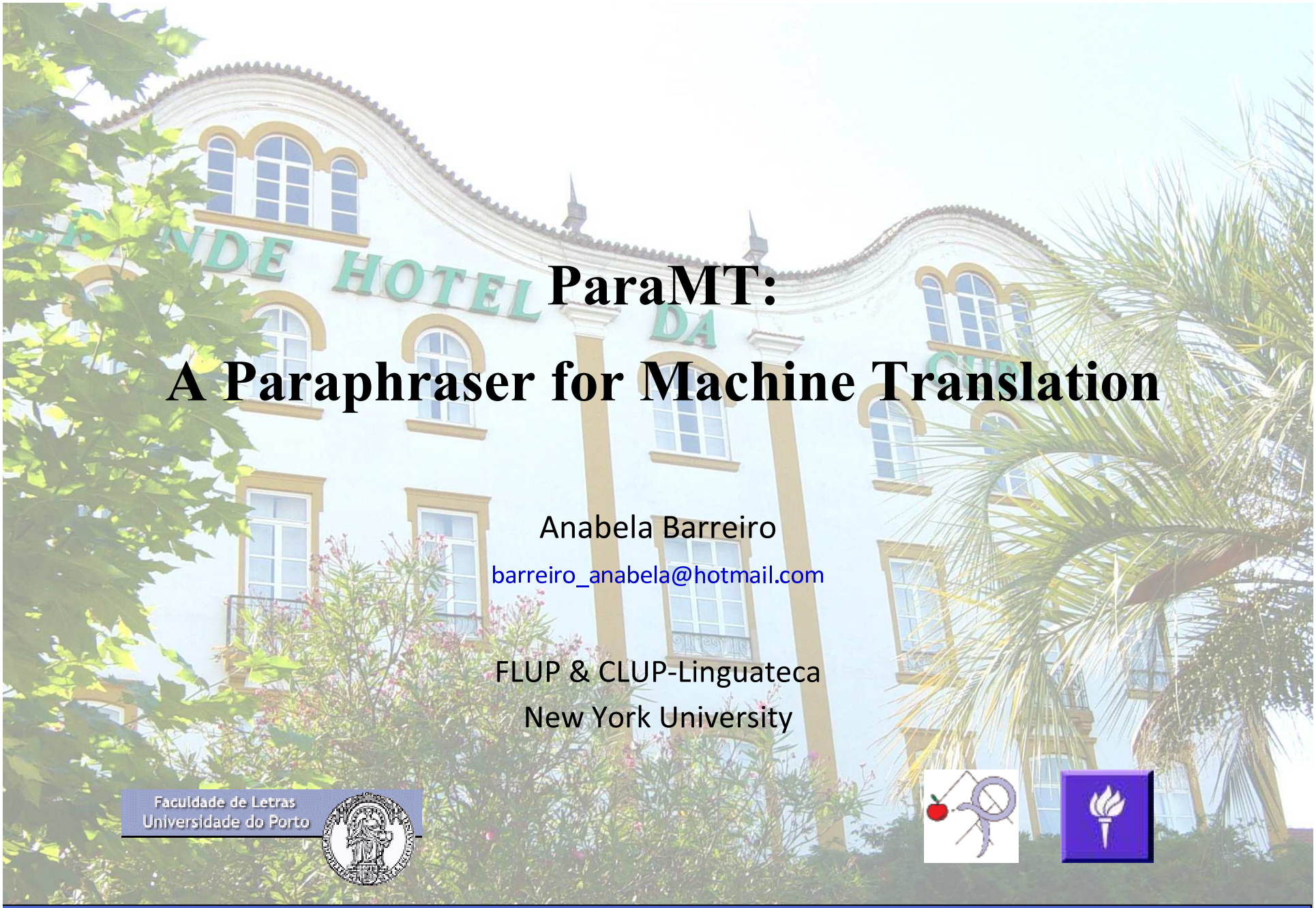
- Linguistic knowledge applied to a machine translation system improves its output quality.

- Effective results from linguistically based research on paraphrases can save substantial effort and resources employed by statistically based machine translation systems

Acknowledgements

Thank you for your attention!

This work was partly supported by grant SFRH/BD/14076/2003 from *Fundação para a Ciência e a Tecnologia*, co-financed by POSI and partly by *Fundação para a Computação Científica Nacional*.



ParaMT:
A Paraphraser for Machine Translation

Anabela Barreiro

barreiro_anabela@hotmail.com

FLUP & CLUP-Linguateca

New York University