Sumário da
LIÇÃO

c) do nº 2 do art. 8º do dec.-lei 239/2007

António Manuel Horta Branco

Título Académico de Agregado
Ramo de Informática, Especialidade de Engenharia Informática

2016
Introdução

Finalidade

O presente documento acompanha o requerimento de instrução da candidatura de António Manuel Horta Branco ao título académico de agregado junto da Universidade de Lisboa no ramo de Informática, especialidade de Engenharia Informática, de acordo com o Decreto-Lei número 239/2007 de 19 de junho e corresponde ao sumário pormenorizado da lição a que se refere a alínea c) do número 2 do Artigo 8º.

A preparação da lição e deste seu sumário foi iniciada em 2014, durante a licença sabática do autor, e completada em 3 de maio de 2016.

Autor

nome António Manuel Horta Branco
afiliação Universidade de Lisboa
url www.di.fc.ul.pt/~ahb
email Antonio.Branco@di.fc.ul.pt
Âmbito e objeto

O âmbito da lição é delimitado pela alínea c) do Artigo 5º do Decreto-Lei número 239/2007:

"c) ... lição sobre um tema dentro do âmbito do ramo do conhecimento ou especialidade em que são prestadas as provas ..."

Tendo a lição por finalidade ser apreciada nas provas para o título académico de agregado, o seu conteúdo deve levar em consideração a natureza deste título, definida no número 1 do Artigo 3º do referido Decreto-Lei:

"1 — O título académico de agregado atesta:

a) A qualidade do currículo académico, profissional, científico e pedagógico;

b) A capacidade de investigação;

c) A aptidão para dirigir e realizar trabalho científico independente."

Da natureza do título académico, sobressai que a lição se destina a ajudar a demonstrar a capacidade do candidato para a investigação científica e para dirigir e realizar trabalho científico independente, devendo por conseguinte ter por objeto um tópico que sirva essa finalidade. Nesta medida, escolhi para tema desta lição o processamento profundo da linguagem (deep language processing no termo consagrado em língua inglesa). Este é um dos eixos que tem sido seguido com intensidade no meu trabalho investigação e no trabalho científico que tenho dirigido e cujos resultados são em parte apresentados nesta lição.

Para a listagem completa do meu trabalho de investigação e dos tópicos que este tem abordado, deve ser consultado o meu curriculum vitae que integra o conjunto dos documentos que acompanham a presente candidatura.

Ao abrigo do Artigo 18º do Decreto-Lei número 239/2007 e da prática da Universidade de Lisboa de aceitar a utilização de uma língua estrangeira nos documentos que acompanham a candidatura ao título académico de agregado,¹ o presente documento é escrito em língua inglesa nas partes que se seguem a esta introdução. Visa-se assim potenciar a sua circulação, apreciação, citação e reutilização por qualquer membro da comunidade científica internacional, que tem na língua inglesa a sua língua veicular. Esta opção está alinhada com a prática comum de se usar a língua inglesa como língua veicular para difundir os resultados da investigação científica também quando outras línguas que não o inglês são as línguas objeto de investigação por tal prática assegurar o muito importante benefício da mais ampla divulgação dessas línguas e do seu conhecimento científico junto de quem não as domina.

¹ http://www.ulisboa.pt/prov_acad_agre (consultada em 1 de abril de 2016)
Estrutura

A seguir à presente introdução deste documento, o mesmo encontra-se organizado em quatro partes principais.

Na **PARTE I—Bases** indico o enquadramento disciplinar e o âmbito científico da área do processamento profundo da linguagem.

A lição continua com a **PARTE II—Contributos** onde apresento a investigação realizada por mim ou sob a minha direção nesta área e os seus resultados.

Segue-se a **PARTE III—Extensão** em que descrevo a aplicação dos resultados obtidos na construção de soluções inovadoras para problemas ou necessidades sociais e de negócio.

Por fim, na **PARTE III—Perspetiva** pondero sobre linhas promissoras de investigação nesta área e sobre possível trabalho futuro.

Este documento é completado por outras partes auxiliares, com as **Referências bibliográficas** e com os **Reconhecimentos**, e ainda com os vários **Anexos** que agregam informação apresentada ao longo da lição sobre a forma de listagens sistemáticas.
Deep Language Processing
for the Science and Technology of Portuguese

António Branco

University of Lisbon
May 2016
## Table of contents

### INTRODUÇÃO
- Finalidade .......................................................................................................................... I
- Autor ......................................................................................................................................... i
- Âmbito e objeto ........................................................................................................................ ii
- Estrutura ...................................................................................................................................... iii

### TABLE OF CONTENTS ........................................................................................................... VII

### TABLE OF FIGURES ............................................................................................................... XI

1  A lecture on deep language processing .................................................................................. 1
   1.1  Goal, object and structure ................................................................................................. 1
   1.2  Deep human natural language computational processing ............................................... 2

### PART I — BACKGROUND .................................................................................................... 5

2  Science and technology of language with deep processing ................................................... 6
   2.1  Language, computation and cognition ............................................................................... 6
   2.2  Grammar and empirical generalizations ........................................................................... 8
   2.3  Parsing and computational complexity ............................................................................ 10
   2.4  Semantic representation and underspecification .............................................................. 11
   2.5  Ambiguity, novelty and robustness .................................................................................... 12
   2.6  Learning and linguistically annotated data sets ............................................................... 13
   2.7  Implementation and experimental validation ..................................................................... 15

3  Impacts ..................................................................................................................................... 17
   3.1  Natural language technology ........................................................................................... 17
   3.2  World languages and the new technological shock ........................................................... 18
   3.3  Portuguese language in the digital age .............................................................................. 21

### PART II — CONTRIBUTIONS ............................................................................................... 27

4  Deep language processing ....................................................................................................... 28
   4.1  LXGram, a computational grammar for Portuguese .......................................................... 28
   4.2  Exhaustive handling of nominal phrases ......................................................................... 32
   4.3  Flexible handling of language variants .......................................................................... 34
   4.4  Robust handling of out of vocabulary items .................................................................... 36
   4.5  Deep handling of temporal meaning .............................................................................. 39

5  Companion, derivative and auxiliary datasets and processing tools ...................................... 42
   5.1  CINTIL—International Corpus of Portuguese ................................................................. 45
   5.2  LX* collection of shallow processing tools .................................................................... 46
   5.3  CINTIL* collection of treebanks .................................................................................... 50
   5.4  LX* collection of deep processing tools .......................................................................... 55
   5.5  *PT collection of multilingual datasets .......................................................................... 57
6 Language processing tasks and applications ......................................................... 59
6.1 Annotation for the semantic web ................................................................. 59
6.2 Temporal extraction .................................................................................. 61
6.3 Summarization .......................................................................................... 66
6.4 Machine translation .................................................................................. 70

PART III — OUTREACH .................................................................................. 75

7 CLARIN research infrastructure ................................................................... 76

8 Innovative solutions ..................................................................................... 81
8.1 Business sector ........................................................................................ 81
8.2 Public sector ............................................................................................ 83

PART IV — OUTLOOK .................................................................................... 89

9 Promising research paths and future work .................................................... 90

REFERENCES ................................................................................................. 95

ACKNOWLEDGMENTS ................................................................................. 111

ANNEX A — CONTRIBUTIONS: PUBLICATIONS LIST ............................................ 113

ANNEX B — CONTRIBUTIONS: DATA SETS LIST .................................................... 125
  CINTIL-DeepBank ...................................................................................... 125
  CINTIL-Definitions ................................................................................... 125
  CINTIL-DependencyBank ........................................................................ 125
  CINTIL-DependencyBank PREMIUM ....................................................... 126
  CINTIL-International Corpus of Portuguese ............................................ 126
  CINTIL-LogicalFormBank ....................................................................... 126
  CINTIL-NamedEntities ............................................................................ 126
  CINTIL-PropBank ..................................................................................... 127
  CINTIL-QATreebank ................................................................ ............... 127
  CINTIL-TreeBank .................................................................................... 127
  CINTIL-WordSenses ................................................................................ 128
  DeepBankPT .............................................................................................. 128
  DependencyBankPT .................................................................................. 128
  LogicalFormBankPT ................................................................................ 128
  LX-Abbreviations ..................................................................................... 129
  LX-NominalInflections ........................................................................... 129
  LX-Stopwords ........................................................................................ 129
  LX-VerbalInflections ................................................................................ 129
  Nexing Corpus ......................................................................................... 129
  PropBankPT .............................................................................................. 130
  QTLeap Multilingual Parallel Corpora ..................................................... 130
  QTLeap WSD/NED Multilingual Corpora .............................................. 130
  TimeBankPT .............................................................................................. 130
  TreeBankPT ............................................................................................. 131
  MWNT.PT - Portuguese MultiWordNet .................................................... 131

ANNEX C — CONTRIBUTIONS: APPLICATIONS, PROCESSING TOOLS AND
ONLINE SERVICES LIST ............................................................................. 133
  CINTIL-Concordancer .............................................................................. 133
ANNEX D — CONTRIBUTIONS: PROJECTS LIST

CINTIL—Treebank Searcher ......................................................... 134
LXGram .............................................................................. 134
LX-CEFR ........................................................................... 134
LX-Centre .......................................................................... 136
LX-Chunker .......................................................................... 136
LX-Conjugator ...................................................................... 137
LX-Definitions ..................................................................... 139
LX-DepParser ....................................................................... 139
LX-Inflector .......................................................................... 140
LX-ListQuestions ................................................................ 141
LX-Lemmatizer ..................................................................... 142
LX-NED .................................................................................. 142
LX-NER .................................................................................. 143
LX-Parser .............................................................................. 144
LX-SenseAnnotator .............................................................. 144
LX-SRLabeler ........................................................................ 145
LX-Syllabifier ......................................................................... 146
LX-Tagger ............................................................................... 147
LX-TimeAnalyzer ................................................................... 148
LX-Tokenizer ........................................................................... 148
LX-Translator .......................................................................... 150
LX-WSD .................................................................................. 150
MWN.PT Searcher ................................................................ 151
SIMBA ................................................................. .................................................. 152
XisQuê ...................................................................................... 152

ASSET—Intelligent Assistance for Everyone Everywhere .................. 155
CEFR—Text Categorization for Language Learning Assessment ............. 155
CLARIN—Common Language Resources and Technology Infrastructure ... 156
DP4LT—Deep Language processing for Language Technology ............... 156
GramaXing—Computational Grammar for the Deep Processing of Portuguese ... 157
LTRC—Language Typology Resource Center ........................................ 157
LT4el—Language Technology for eLearning ..................................... 158
METANET4U—Enhancing the European Linguistic Infrastructure .......... 158
NEXING—Natural Negation Modeling and Processing .......................... 159
QTLeap—Quality Translation by Deep Language Processing Approaches ........ 159
QueXting—Answering Questions in the Portuguese Web ......................... 160
S4S—Smartphones for Seniors .................................................... 160
SemanticShare—Ferramentas and Recursos para o Processamento Semântico ..... 161
TagShare—Tools and Resources for Shallow Morpho-syntactic Tagging and Processing ... 161
Table of figures

Figure 1: Roman copy in marble of a Greek bronze bust of Aristotle (384-322 BC) by Lysippus, c. 330 BC (left) and picture of Alan Turing (1912-1954). ........................ 2

Figure 2: Portrait of Gottfried Leibniz (1646-1716) by Francke, c. 1700. (left) and picture of Gottlob Frege (1848-1925), c. 1879. ................................................................. 7

Figure 3: Pictures of Richard Montague (1930-1971), c. 1967 (left) and of Noam Chomsky (born 1928), in 2005. ................................................................. 9

Figure 4: The underspecified logical form of the sentence Every dog probably chases some white cat in the meaning representation formalism Minimal Recursion Semantics (MRS) followed by six formulas into which its underspecification can be resolved representing different readings of the sentence (Copestake et al., 2006, (25), (26)). ......................................................... 12

Figure 5: Pictures of Martin Kay (born 1936) (left) and of Eugene Charniak (born 1946). .............................................................................................................. 13

Figure 5: A sample from the CINTIL-Treebank with the sentence A triatleta portuguesa tem muitos admiradores em Roth ("The Portuguese triathlete has many admirers in Roth") and its pretty-printed constituency tree with the leaf nodes annotated with semantic type (for proper names), the lemma and the inflection features of the respective word. ............................................................. 14

Figure 7: Picture of the cover of the White Paper on The Portuguese Language in the Digital Age (Branco et al., 2012). ................................................................. 18

Figure 8: Number of references to data sets and processing tools per language at ten top scientific conferences between 2010 and the publication of these results in 2012, from (Mariani and Frankopoulo, 2012). ........................................................ 20

Figure 9: World map with countries and territories with Portuguese speaking population. .................................................................................................. 22

Figure 10: Image of the cover of Helder Coelho's (1979) doctoral dissertation, an inaugural landmark in the research on the technology for Portuguese. ....... 23

Figure 11: Picture of the members of NLX-Group in 2015. ........................................ 23

Figure 12: Image of the front page of the website of PROPOR2016-The 12th International Conference on the Computational Processing of Portuguese, available at http://propor2016.di.fc.ul.pt ................................................................. 25

Figure 13: Pictures of João Silva (left) and Francisco Costa, doctoral and post-doctoral students under my supervision. They are two major co-authors of the results described in the present chapter. .......................................................... 28
Figure 14: A 1 m x 1.4 m printout in font size 4 of the fully-fledged grammatical representation, in the AVM format of (Pollard and Sag, 1994), for the 6 word sentence *Todos os computadores têm um disco* (*"Every computer has a disk"*) produced by the LXGram. The arm and the pen are included to help appreciate the relative proportion of the whole representation.............................................. 30

Figure 15: A group photo of members of the European R&D network of excellence Multilingual European Technology Alliance (META-NET) in Berlin in 2011.. 42

Figure 16: Pictures of the NLX-Group in 2011 (left) and in 2015 with many of the co-authors of the results introduced in the present chapter................................. 44

Figure 17: A sample of the internal representation in the annotated CINTIL corpus with the example sentence *A sentença estava há muito tempo anunciada: a APDL não iria aceitar a proposta do único concorrente ao concurso de concessão dos terminais de contentores.* (*"The judgment was announced long time ago: the APDL would not go to accept the proposal from the only candidate to the call for the concession of the container terminals."*). ............ 45

Figure 18: An example with the sentence *Entre os sete presos, há cidadãos dos Estados Unidos, da China e da Formosa* (*"Among the seven persons arrested, there are citizens from the United States, from China and from Taiwan"*) and its pretty-printed constituency tree decorated with grammatical functions and semantic roles as the sentences annotated in the CINTIL-PropBank.......................... 51

Figure 19: The same example sentence of previous figure now annotated with its pretty-printed dependency graph, as the sentences in the CINTIL-DependencyBank................................................................. 52

Figure 20: An example of the pretty-printed attribute-value matrix (AVM) diagram with the MRS logical form of the sentence *Deverá ser apresentado aos tribunais no Outono.* (*"He should be presented to court in Autumn."*) and its logical form in the CINTIL-LogicalFormBank........................................ 52

Figure 21: Pictures of Sara Silveira (left) and João Rodrigues, doctoral students under my supervision. They are two major co-authors of the results described in the present chapter................................................................. 59

Figure 22: RDF/XML representation of the example sentence *Washington acompanhou os movimentos de Saddam desde a primeira hora.* (*"Washigton followed Saddam's movements since the first hour."*) output by the parser for the semantic web................................................................. 61

Figure 23: A sample of the internal representation of the annotated TimeBankPT with an excerpt with the example sentence *ABC.1830.0611 REPORTAGEM Em Washington, a Federal Aviation Administration publicou gravações do controlo de tráfego aéreo da noite em que o voo TWA800 caiu.* (*"ABC.1830.0611 NEWS STORY In Washington, the Federal Aviation Administration released air traffic control tapes from the night the TWA flight eight hundred went down."*)...... 64

Figure 24: Partial image of the interface with the online demo of the LX-TimeAnalyzer temporal extraction system................................................................. 66

Figure 25: Image of the interface with the online demo of the SIMBA multi-document automatic summarizer................................................................. 70
Figure 26: A picture of the LX-Translator mobile interface. ............................................. 74
Figure 27: A picture of the LX-Translator web interface....................................................... 74
Figure 28: The interface of the mobile version of XisQuê question answering systems prepared for the S4S project. .............................................................. 82
Figure 29: Image of the interface of the LX-CEFR online service at the Camões website................................................................................................................................. 85
Figure 30: Part of Figure 4 of (LeCun et al., 2015) with a two dimension representation of distributional semantics vectors for some phrases where one can observe that semantically similar expressions are mapped to nearby representations......................................................................................................................... 91
1 A lecture on deep language processing

In this first chapter, an introduction to the present lecture is provided.

In Section 1.1 Goal, object and structure, the purpose of the lecture, its subject matter and the parts along which it is structured are introduced.

In Section 1.2 Human natural language deep computational processing, a delimitation of deep language processing within the overall scope of scientific and technological enterprise is offered.

1.1 Goal, object and structure

This document is part of my application for the academic title Agregado. It is the written summary of the lição, which is a public oral lecture with typical class length and one of the three required elements of the academic examinations for the title.

This lecture is on deep language processing, also referred to as natural language understanding or deep linguistic processing in some quarters, and is structured along three major parts.

The first part, PART I—Background, offers an overview of the area of deep language processing, presenting its key foundations and major features.

The lecture continues with PART II—Contributions, where the research that has been undertaken by myself or under my supervision in this domain is presented.

The third part, PART III—Outreach, reports on the exploitation of the results obtained with that research in view of providing innovative solutions to societal and business challenges.

The fourth and final part, PART IV—Outlook, closes the lecture by considering promising research lines in this area and possible future work.

The present written summary also includes the parts containing References, listing all publications referred to in the lecture, and Acknowledgments.

For the sake of readers’ convenience, this written summary includes also Annexes with the consolidated lists of the publications, data sets, processing tools, applications, services and projects referred to throughout the lecture as results of research undertaken by myself or under my supervision.

---

2 The title of Agregado is the top academic title in Portugal. It is distinct from the professional position of Catedrático (full professor) at the top of the Portuguese academic career.

This lecture addresses part of my research activity and results, a complete rendering of which can be found in the curriculum vitae included in the present application.

1.2 Deep human natural language computational processing

From Aristotle's view of humans as animals endowed with *logos* (*Nicomachean Ethics* I. 13) to Turing's (1950) proposal of a language based game to ascertain intelligence twenty three centuries later, language has been consistently regarded in the western culture as a quintessential feature and a hallmark of human distinctiveness in the universe of known entities.

![Figure 1: Roman copy in marble of a Greek bronze bust of Aristotle (384-322 BC) by Lysippus, c. 330 BC (left) and picture of Alan Turing (1912-1954).](image)

The science and technology of human natural language aims at understanding this hugely complex symbolic system, form of communication, cognitive capacity and specific kind of behaviour and at developing solutions that both enhance its utilization by human and non-human agents alike and support language based interaction between them.

Within the realm of natural language science and technology, deep language processing is a domain addressing the mapping between discrete linguistic forms and the representation of their meaning.

The qualifier *deep* in the term deep language processing evokes thus the procedures of uncovering, making explicit and representing the underlying meaning of an input expression (analysis) and, vice-versa, of obtaining expressions that convey a given meaning captured in a semantic representation (generation).

---

4 This usage of *deep* may have drawn inspiration from the term deep structure coined in the 1950’s in the scope of Chomskyan grammar (Chomsky, 1955), even though in those initial proposals of generative grammars deep structure was set apart from the so called logical form. The qualifier deep in deep language processing has thus a different purpose and predates its usage in the term deep machine learning, which designates a recently trendy area that concerns an approach for machine learning (LeCun et al., 2015) and not necessarily the concrete task of representing and mapping to and from linguistic meaning.
This domain shares the abstraction that the mapping between analogic acoustic forms and discrete linguistic forms takes place in some other concomitant procedure. This other procedure is to be tackled by another research domain, termed speech processing, addressed as part of signal processing typically in electrical engineering departments.

Likewise, the processing of linguistic form and meaning can be abstracted away from the overall cognitive capacity, which is tackled by cognitive science and artificial intelligence at large or by some of their other subdomains such as knowledge representation, reasoning, etc.

Being related to speech processing on the one hand and knowledge processing on the other, deep language processing has thus been considered and addressed as a specific domain of research.
PART I — Background

In this first part of the lecture, a brief introduction to the area of deep language processing, presenting its foundations and major impacts, is provided.

A very brief introduction to the major aspects of deep language processing is the content of Chapter 2 Science and technology of language with deep processing.

Chapter 3 Impacts highlights its economic, societal and cultural consequences.
2 Science and technology of language with deep processing

This chapter is structured along the following eight sections.

In Section 2.1 Language, computation and cognition, how deep language processing relates to foundational aspects of Computer Science and Cognitive Science is discussed.

Section 2.2 Grammar and empirical generalizations is used to clarify the roots of deep language processing in regularities that are determined by means of empirical inquiry and are gathered in a body of knowledge usually designated as grammar.

Section 2.3 Parsing and computational complexity helps to present the key issues on making grammar regularities that were empirically elicited to be uncovered, represented and actionable by computational means.

In Section 2.4 Semantic representation and underspecification, a crucial dimension of deep grammatical representation, namely the one seeking to capture the meaning of the natural language expression at stake, is focused on.

Section 2.5 Ambiguity, novelty and robustness brings into consideration the methodology used to resolve ambiguity, i.e. given the context where a certain expression occurs, the approach used to associate a grammatical representation to it from the range of its admissible representations.

In Section 2.6 Learning and linguistically annotated data sets, the instrumental data sets used to support the implementation of disambiguation techniques are introduced.

Section 2.7 Implementation and experimental validation introduces platforms that support the actual development of deep language processing software.

2.1 Language, computation and cognition

For practitioners from an array of different disciplines, deep language processing exhibits a clear resemblance, and is thus confluent, to a number of issues addressed in those disciplines, such as Science and Technology of Programming and Computation Theory (Computer Science), Knowledge Representation and Description Logic (Artificial Intelligence), Philosophical Logic, Philosophy of Language and Philosophy of Mind (Philosophy), Grammar and Semantics (Linguistics), Cognitive Science, etc.
In the area of Science and Technology of Programming, a central topic is the design and deployment of language processors, a family of software programs with a range of functionalities that includes assemblers, compilers, interpreters, translators, etc. (Aho et al., 2006). In broad terms, a language processor can be described as reading a given text, checking whether that text is a valid sentence in a given language, and executing some action determined by the meaning of that sentence. In the particular case of compilers, they translate from high level programming languages (e.g. C++, Java, Python) into machine languages, or machine codes, that are specific to a family of processors or central processing units.

The closeness to deep language processing is forceful, where an input sentence in a given natural language (e.g. Portuguese, English, Chinese, etc.) is translated into a sentence of some formal language (e.g. Web Ontology Language, First Order Logic, Generalized Quantifier Logic, etc.) that should convey the semantics of the input sentence and thus be taken as a representation of its meaning.

As these semantic representations should be actionable representations of the meaning of natural language sentences, which should allow for the automation of knowledge handling and of reasoning, their design and operation is at the heart of Artificial Intelligence. To a large extent, deep language processing is thus confluent also with the areas of Knowledge Representation and of Description Logics, fields of Artificial Intelligence concerned with formalism that are both epistemologically and computationally adequate to express knowledge (Baader et al., 2007).

While sharing key commonalities with the areas of knowledge representation and language processors, deep language is distinct in that it is focused on natural languages, in contrast to the artificial formalisms and programming languages. This distinction spans further to the fact that natural languages are processed in the human mind-brains. This brings deep language processing somehow into confluence also with the fields of Philosophical Logic and Philosophy of Language (Leibniz, 1666; Frege, 1879, 1982), and also Philosophy of Mind, including the discussion around the hypothesis of a language of thought (Fodor, 1975).

This is further strengthened by a particular subarea of Linguistics whose purpose is to address the cognitive capacity of language by investigating an explicit representation of the implicit grammatical knowledge that the speakers of a given language master in order to cope with it. It is hypothesized that it is this knowledge that allows speakers to tell sentences and non sentences apart and
systematically associate the surface form of sentences with their meaning, a desideratum at the core of the modelling undertaken by deep language processing.

This computational modelling is also of utmost importance for the eventual discussion concerning the foundations of Cognitive Science at large. This discipline studies the mind from the perspective of a complex system of operations on information, supported by biological devices ensured by the brain and the body (Bechtel and Graham, 1999). Under an influential formulation, “The mind is what the brain does; specifically, the brain processes information, and thinking is a kind of computation” (Pinker, 1998, p.21).

A major outcome of Computation Theory is that all the formalisms independently proposed to capture the intuitive notion of computability by effective procedure, including Turing machines, turn out to provably capture the same class of functions. Hence, information processing can be transferred among processing devices, whose ultimate computational ability is not superior to any Turing machine.

Along this vein, the eventual empirical determination of the complexity of language processing within the hierarchy of computational complexity classes is of crucial importance to assess the tractability and ambition of deep language processing. Moreover, given that language is a high-level cognitive capacity — in fact, one of the human cognitive capacities par excellence —, this is also of utmost importance to the conditions of viability of the very epistemological foundations of Cognitive Science (Branco, 2005).

### 2.2 Grammar and empirical generalizations

While language is a complex form of symbolic communication, cognitive processing and social behaviour, its complexity is scientifically addressed by deep language processing from the vantage point of its symbolic dimension. In concrete, the starting points for the modelling of language are the empirical regularities that have been uncovered from its symbolic forms and that are the basis of generalizations gathered under a field of study that throughout the centuries has been commonly called Grammar.

In very broad terms, these regularities can be seen as manifestations of two overarching generalizations. One is that linguistic forms group themselves into sets, called grammatical categories, whose elements can be recursively and systematically combined with each other into larger linguistic forms along a finite — and actually small — number of admissible patterns. The other is that the meaning of a linguistic form is a function of the meanings of the forms that compose it and the way they happen to be combined. In an abbreviated designation, the latter has been known as semantic compositionality, and the former as syntactic productivity.

With its origins in western culture dating back to classical antiquity, the study of language, in general, and of grammar, in particular, underwent a decisive upsurge in the second half of the 20th century, in line with the overall scientific
developments that took place in computation, linguistics and cognitive science. The domain of deep language processing has its origins in these developments and emerges at the confluence of important lines of research in this area, including generative grammar, Montague grammar and unification grammar.

Generative grammar was pioneered by Noam Chomsky and one of its seminal contributions can be found in his Doctoral dissertation, from 1955, concluded when he was involved in a machine translation project at the MIT Research Laboratory of Electronics (Chomsky, 1955). At the time, the sub-discipline of Computation Theory concerned with formal languages and automata was burgeoning with new advances. Generative grammar emerged, among other things, as the result of the concepts and instruments of analysis from the area of formal languages and automata being applied to natural languages.

Pursuing further this perspective, and bringing it also into the realm of the semantics of natural languages, Richard Montague published in 1970 a paper with the title “English as a Formal Language” (Montague, 1970). This became the seminal work of what would be known as Montague Grammar, and which later evolved into areas that are known nowadays more broadly as formal and computational semantics. A key insight is that the procedure of translating from a sentence of natural language into a semantically equivalent sentence of a logical language can be fully formalized, and that this can be the basis for computing the representation of its meaning.

![Figure 3: Pictures of Richard Montague (1930-1971), c. 1967 (left) and of Noam Chomsky (born 1928), in 2005.](image)

A few years later, borrowing from the flourishing logic programming paradigm, Martin Kay publishes a paper with the title “Functional Unification Grammar: A Formalism for Machine Translation”, which was one of the triggers that set in motion what became known as unification grammar (Kay, 1984). Here, the key contribution was to help popularizing unification as the underlying operation that supports the computation of syntactic and semantic information about natural language sentences in an integrated fashion.
2.3 Parsing and computational complexity

While grammar gathers the regularities that happen to be empirically uncovered from the symbolic manifestation of language, its nature is scientifically addressed by deep language processing from the vantage point of computation. This allowed for a quantum leap in the study of language, in as much as it opened novel research lines into crucial language dimensions. These dimensions include the systematicity of the body of regularities gathered under a grammar, the conditions of possibility of the knowledge thus collected and the way this could be effectively actionable by agents whether they be human speakers or artificial devices.

Once one pursues the goal of computationally processing natural language, the key question of which computation models are better suited to this endeavour arises. Given that computation can support useful technological solutions for some subtasks within a fully-fledged interaction in natural language, at a more instrumental level this is a question whose answer may be addressed by engineering approaches. But given that human language is a naturally occurring object, at a more fundamental level this is a question whose answer requires basic empirical research.

As different computation models have different levels of inherent computational complexity, empirically determining the model that eventually underlies natural language is thus inextricably related to the twin research question concerning the computational complexity of natural language. In this respect, the search for firm results may therefore be only ever as successful as the sub-procedure under consideration is simpler and as the empirical evidence is more elementary and less controversial.

When the issue of complexity is addressed, research results have been reported in the literature by focusing on what is known in parsing as the recognition problem: given a string $s$ of lexical forms in a natural language $L$, how complex is the procedure that determines whether $s$ is or is not a sentence of $L$. It has been common practice to use a threefold complexity hierarchy proposed by Chomsky (1956) that groups grammars into regular, context-free and context-sensitive types. In such a hierarchy, all regular grammars are context-free grammars, and the set of all languages defined by the former are properly included in the set of all languages defined by the latter, with similar considerations holding with respect to context-free and context-sensitive languages, respectively.

On the basis of empirical evidence and the underpinnings of the theory of computation, different arguments have been put forward that reasonably fill the whole spectrum of hypothesis (Branco, 2005). In what concerns the recognition problem in natural languages, they range from the position that natural languages are regular (Roches and Schabes, 1997) to the view that they are context-sensitive (Berwick and Weinberg, 1982), including the claim that they are context-free (Gazdar and Pullum, 1985). Hence, this is still a productive research question, and as more suitable empirical data from more and better articulated sources of evidence (e.g. contrasts in grammatical judgements, linguistic performance and behavioural scores, brain records of activity, neurological findings, etc.) become
available, one should expect that the number of working hypotheses about language complexity will be narrowed down.

2.4 Semantic representation and underspecification

While the systematic encoding of regularities into grammars has been experimented with on the basis of different underlying computational models, its nature has been addressed by deep language processing from the vantage point of the mapping between linguistic forms and the representation of their meaning. This envisages grammar as an actionable device that extracts an explicit grammatical representation of the linguistic regularities that are implicit in a given input sentence that happens to instantiate them; and conversely that outputs a sentence that instantiates and satisfies the regularities and constraints encoded in a given input grammatical representation.

A crucial component of a grammatical representation of a sentence is the representation of its meaning, typically formulated in a dedicated semantic representation language, based on some formal or logic language. Hence, another interesting way to envisage a computational grammar has been to take it as a device that operates the automatic translation of sentences between two languages, namely between a natural and a logic one.

The underlying operation of the deep computational grammar usually supports the compositional determination of the semantic representation of a sentence from the representation of its syntax and of the semantic representation of its parts. This semantic representation is usually encoded under a so-called underspecification approach.

Natural language expressions are often ambiguous, thus allowing for many readings. While each such reading is suited to receive an appropriate individual semantic representation, these readings can nevertheless be encoded altogether into a single expression of the representation language. This is thus an underspecified representation of the meaning conveyed by the actual occurrence of the natural language expression, that may itself be disambiguated later on (an example is displayed in the figure below). This permits to alleviate the grammar from the task of full disambiguation, thus increasing its efficiency, and to outsource that task to some external module or agent if necessary or convenient.
Figure 4: The underspecified logical form of the sentence *Every dog probably chases some white cat* in the meaning representation formalism Minimal Recursion Semantics (MRS) followed by six formulas into which its underspecification can be resolved representing different readings of the sentence (Copestake et al., 2006, (25), (26)).

As the linguistic capacity is in close interaction with other cognitive capacities, and as the semantic representation provided by a computational grammar is crucial in that contact particularly in relation to automated reasoning, an appropriate definition of the representation language turns out to be an important question. Given that meaning representation can support useful technological solutions, this is a practical question whose answer may be addressed by engineering approaches. But at a more fundamental level, given that human language is a naturally occurring object, it is a question whose answer requires basic empirical research.

Different logic languages have different expressive power and different computability and tractability limitations. In a landmark article, Barwise and Cooper (1981) showed that the treatment of the determiner *most* requires a semantic representation that resorts to a generalized quantifiers logic more expressive than first order logic. So far, this appears to have been the most stringent requirement concerning an empirically justified semantic representation formalism for natural language.

### 2.5 Ambiguity, novelty and robustness

While the mapping between linguistic forms and the representation of their meanings has been experimented with on the basis of different semantic representation formalisms, its effective implementation has been addressed by deep language processing from the vantage point of automated methods for the resolution of linguistic ambiguity and linguistic novelty.

On the one hand, as natural language ambiguity is ultimately resolved by world knowledge elements that are very hard, if not unfeasible, to properly and completely factorize, computational grammars help to bridge between linguistic and other non-linguistic knowledge and cognitive capacities. On the other hand, and together with their key relevance for the advancement of the science of language, deep computational grammars play an important practical role in
supporting robust and efficient tools and solutions in the area of language technology.

In this connection, the introduction of statistical parsing to natural language processing, by means of which one of the possible readings of an ambiguous expression is determined, represented a crucial breakthrough. It had a distinguished pioneer in Eugene Charniak (1996) and his work on probabilistic context-free grammars. These are grammars whose production rules are associated to probabilities, typically computed by machine learning programs running over specifically annotated datasets, termed treebanks. In treebanks, each occurrence of a sentence is associated with a grammatical representation encoding the linguistic regularities and the reading that that sentence happens to instantiate in that contextualized occurrence.

![Figure 5: Pictures of Martin Kay (born 1936) (left) and of Eugene Charniak (born 1946).](image)

As ambiguity is pervasive in natural language, its resolution has been addressed not only for the syntactic and semantic representation of sentences but also along other grammatical dimensions as well, including part-of-speech tagging, named entity recognition, word sense disambiguation, and many others.

Concerning natural language novelty, in turn, deep language processing adopts approaches also based on variations of statistical inference to cope with expressions that are newly introduced into language as the result of the slow yet steady process of language change.

Likewise, it also deals with expressions whose eventual occurrence, simply by virtue of their very low frequency, has the same practical effect for the computational grammars as the result of some language change or novelty.

### 2.6 Learning and linguistically annotated data sets

While treebanks allow the relevant stochastic parameters for ambiguity resolution to be estimated, these linguistically interpreted datasets have been addressed and further explored in deep language processing from the vantage point of the implicit grammatical regularities they eventually encode.

A landmark in the development of treebanks is the dataset known as the Penn treebank (Marcus et al., 1993). This treebank was developed over a corpus of approximately 1 million tokens of English text that consists of excerpts from news articles of the Wall Street Journal. This treebank resulted from each sentence being associated to a representation of its syntactic constituency structure drawn by a
human annotator. These grammatical representations encode the syntactic regularities instantiated by the sentences in question and are hand made by annotators following a set of so-called annotation guidelines, which include a list of the linguistic regularities of interest that should be checked for in the sentences.

Once treebanks are in place, the range of options is enlarged with respect to the gathering of grammar rules and to the construction of a grammar.

As grammatical regularities and generalizations are made explicit by principled empirical linguistic research, these can be manually programmed into the grammar code, and then for the sake of ambiguity resolution be supplemented with relevant stochastic parameters obtained from a treebank.

Another option is to resort solely to the treebank and to use machine learning techniques to automatically extract a grammar, including its relevant body of rules (Magerman, 1995; Collins, 1996; Charniak, 2000). Here, the empirically uncovered regularities eventually inform the resulting grammar, although by means of a lengthier circuit. By investigating empirical data, grammatical regularities are uncovered and reported in research publications by linguists; from there they are selected and used in the annotation guidelines for some treebank by the designer of the treebank; from there they are instantiated in the syntactic representations of the sentences by the human annotators of the treebank; and from there they are uncovered again, this time automatically, by some machine learning program, which was manually coded by a software developer (an example sentence annotated with its syntactic constituency tree is displayed in the figure below).

![Diagram of a constituency tree]

**Figure 6**: A sample from the CINTIL-Treebank with the sentence *A triatleta portuguesa tem muitos admiradores em Roth* (“The Portuguese triathlete has many admirers in Roth”) and its pretty-printed constituency tree with the leaf nodes annotated with semantic type (for proper names), the lemma and the inflection features of the respective word.

While treebanks like the Penn treebank offer only a partial grammatical representation, namely of the syntactic constituency of the sentences, other treebanks have been developed that seek to associate each sentence with its fully-fledged deep grammatical representation, including the semantic representation of
its meaning. Here the relation between grammars and treebanks gains yet another twist.

Given the high level of complexity of the grammatical representation of a sentence, it is not practically feasible for that representation to be manually built in a piecemeal fashion by some human annotator in order to be associated to the respective sentence (see Figure 14 below). In order to proceed with such deep treebanking, this annotation process is assisted by the computational grammar, which provides the relevant grammatical representation. The role of the human annotator is thus to select, from the parse forest delivered by the grammar, the particular parse result that can be best used to annotate the sentence.

Accordingly, as the grammar is necessary to support the development of the companion deep treebank, and the deep treebank is useful to support the development of the companion grammar, these datasets — pioneered by (Oepen et al., 2002) — became known as dynamic treebanks. The concept, design and first concreteization of multilingual parallel dynamic deep treebanks to support deep machine translation and semantic based multilingual applications would be developed by another group of researchers where I am included and were reported in a publication of which I am a co-author (Flickinger et al., 2012).

### 2.7 Implementation and experimental validation

While interest in grammars and parsers for natural language has been at the core of natural language processing at large, their experimental validation has been thoroughly explored by deep language processing from the vantage point of the large-scale implementation and long-term development of computational grammars for particular languages, like English or Portuguese.

A number of so-called grammatical frameworks have been put forward with the objective of accommodating the development of computational grammars. These frameworks combine and offer different options for encoding empirical regularities, most notably those concerning syntactic constituency, grammatical dependency functions, long-distance relations, syntax-semantics interface, semantic description formality, etc., and different options concerning the underlying parsing mechanisms and their location in the Chomsky hierarchy of computational complexity.

The grammatical frameworks supporting computational grammars that have received more substantial attention in research include Combinatory Categorial Grammar (Steedman, 2000), Tree-adjointing Grammar (Joshi, 1985), Lexical-functional Grammar (LFG) (Bresnan, 2001), and Head-driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1994). From these frameworks, two of them stand out, namely LFG and HPSG, due to the much larger volume of work devoted to them and to consistent and long-term development of a number of computational grammars of these kinds for particular languages.

To encode linguistic regularities, both LFG and HPSG resort to variants of attribute-value description formalisms that are suited to operation by unification.
While LFG relies on different strata of information where different grammatical dimensions are encoded and on the mapping between them, HPSG is monostratal, relying on an integrated grammatical representation for each linguistic expression.

In order to provide a concise introduction to HPSG and to further disseminate the interest on this framework among Portuguese speaking students, I co-authored with Francisco Costa, my PhD student at the time, three chapters of the book *Abordagens Computacionais da Teoria da Gramática* ("Computational Approaches to the Theory of Grammar"), published in 2012 in Brazil (Branco and Costa, 2012a,b,c).

The development of LFG computational grammars gathers around the PARGRAM initiative, that involves a number of research groups currently working on the grammars of English, Chinese, French, German, Polish, and Norwegian, among others. These grammars are supported by the XLE development platform that was created by Xerox to support LFGs as far back as the 1980’s.

Research on HPSG, in turn, has a major focal point in DELPH-IN, a collaborative initiative counting on the contribution of seventeen research groups worldwide, including the research group I have set up and am coordinating. These groups resort to the LKB open source platform to develop a number of grammars for English, German, Japanese, Portuguese and Spanish, among others.

DELPH-IN started in 2005 and sprung out of the conclusion of the speech-to-speech machine translation project VERBMOBIL funded by the German government for a decade (Whalster, 2000). The DELPH-IN researchers gather once a year for the annual summit of this initiative. The inaugural summit was organized by myself in 2005 at the University of Lisbon. To celebrate its 10th edition, the DELPH-IN annual summit was organized again in Portugal by myself in 2014.
3 Impacts

This chapter addresses the economic, societal and cultural impacts of deep language processing and is structured along the following three sections:

Section 3.1 Natural language technology proceeds with the consideration of the natural language processing and applications supported by deep language processing.

Section 3.2 World languages and the new technological shock addresses the impact of the new language technology in the world system of languages.

Section 3.3 Portuguese language in the digital age closes this chapter with a reflection on the level of technological preparation of Portuguese for the digital age.

3.1 Natural language technology

While the development of large-scale computational grammars aiming at full linguistic coverage can be seen as a tangible outcome of research on deep language processing, this endeavour is ultimately driven not only by scientific but also by application objectives.

The technological importance of the advances on deep language processing can be characterized both in terms of its direct and indirect impacts. By direct impacts, we refer to the solutions permitted by the exploitation of the semantic representation that is output by deep grammars.

By indirect impacts, we refer to the application of advances from a range of sub-domains that sprung out of the core research on deep language processing, and to a large extent became self-contained topics of research, such as part of speech tagging, named entity recognition, processing of multi-word expressions, semantic role labelling, word sense disambiguation, dependency parsing, etc.

The results obtained in these sub-domains provide essential components that are embedded in every application area with commercial relevance in which natural language processing is involved. Thus, the indirect technological impacts from deep language processing can be seen in a vast array of technologies, including document retrieval, information extraction, document clustering, summarization, plagiarism detection, computer-aided translation, localization, sentiment analysis, dialogue systems, and many others, as explained at length in the White Paper on The Portuguese Language in the Digital Age (Branco et al., 2012: Chaps. 4.1-4.3) (image of its cover is depicted in the figure below).
The direct technological impacts of deep language processing, in turn, are to be found in the exploration of the meaning representation of sentences, which is the essential output of computational grammars. The level of maturity of the advances obtained so far in this domain has permitted their technological exploration mostly in the application domains of machine translation and, more recently, question answering.

As two notorious examples of the exploration of question answering with broad commercial or public impact, one can mention the Powerset company and the Watson system. Powerset was a company founded to explore the LFG deep computational grammars developed at Xerox PARC. These grammars were used to obtain the meaning representations of sentences of webpages, which were then stored to support subsequent question answering. This company was eventually acquired by Microsoft for $100 million. The Watson system, in turn, is a flagship endeavour of IBM, whose first incarnation was as the system that beat humans in the "Jeopardy" TV show, where a question answering game was played. It is now announced as being extended to commercial applications, including solutions for the business intelligence and health care areas.

### 3.2 World languages and the new technological shock

While progress in language processing is permitting innovative solutions for both old and new societal challenges and opening the way for new business opportunities, these advances bring along also new challenges for the world languages.

Natural languages are historical entities and as such they emerge as autonomous units at some point in human history, they change and evolve, they may get merged and combined with each other, and they eventually get extinct when no further native speakers remain. Technological innovations with impact on
how one can use natural language have represented important landmarks in the history of languages, where the advent of the writing systems (some six thousand years ago) or of the printing press (some six centuries ago) are just two most prominent examples of such technological shocks. Decline, and many times extinction, is the fate of languages — and of the cultures and nations they support — that eventually had no sufficient institutional and historical conditions to benefit from these innovations, at least not in the same decisive degree as some other languages competing with them might have had.

There exist approximately 7000 human languages in the world (Lewis, 2016) and the ongoing advent of language technology represents the next big technological shock for their ecosystem. Like in the past, languages that will not be sufficiently prepared to benefit from the new technology, and from the concomitant new conditions of language usage it brings along, will face decline. Nevertheless, unlike in the past, the present technological shock is expected to be much more impactful than previous ones, with unprecedented negative consequences for language diversity, given the increasingly globalized and digital world we are living in.

In this respect, it is worth noting a serendipitous analogy between the realms of biological studies and language studies. There is certainly very much to learn and gain from the exhaustive study of the biological system and species of say a given island or territory. When moving to other territories, the results thus obtained may help to have a good starting point to understand the biological systems of those other places. Many species may be common, some other species present only slight differences that can be easily understood in the light of what is already known, and even when addressing some new species, the methods of study previously developed can be reused or adapted with no much effort. However, sooner or later, entirely new species and sufficiently different biological systems will require the development of new methods of research and will permit the amassing of a wealth of totally new results and discoveries. Eventually, by encompassing more islands, territories or continents and their biological systems under the scientific inquiry, it is the whole understanding of the biological world that is enlarged and more accurately reshaped.

Likewise, there is certainly very much to learn from the fully in depth study of a given single language, which will help to uncover specific sophisticated grammatical phenomena and brings to light non trivial morphological, syntactic and semantics regularities, among many other sorts of generalizations and intricate interactions between them. The empirical generalizations and methodological advances thus obtained will certainly be helpful, to a certain degree, to study and understand the structure and function of expressions from some other languages. However, that cannot replace the specific in depth study of the other languages and will not permit, only by itself, an appropriate understanding of those other languages and of the whole human linguistic capacity.

Naturally, this analogy with biological studies lends itself to be easily stretched further into the area of technological applications of scientific knowledge. If the biological and agricultural system of a given territory is to be explored in a sustainable and self-sufficient way, it is imperative that that system be thoroughly
studied and known and that its inhabitants develop and master the solutions that best fit their place and its specific natural conditions and potential.

Likewise, if a given natural language is to strive in the digital age, it is imperative that it is thoroughly studied and technologically prepared so that their speakers are empowered to master, foster and explore advanced digital technologies that are most appropriate for their language.

A significant objective indicator of the level of technological preparation of any given language for the digital age is certainly provided by the volume of activity that takes that language as an object of scientific research and technological development. And a significant indicator of the volume of research activity concerning a certain language, in turn, is given by the number of citations of the data sets and of the processing tools and applications for that language that can be found in scientific publications.

Following this rationale, Mariani and Frankopoulo (2012) undertook the estimation of the volume of research activity concerning different languages by compiling the number of references to data sets or processing tools for those languages in the articles published in ten top scientific conferences between 2010 and 2012. The results obtained cover 32 languages and are compiled in a chart that is reproduced in the figure below.

As expected, the number of citations and thus the volume of research about languages should be roughly proportional to the level of economic and scientific development of the countries where those languages are spoken and hence proportional to the volume of funding devoted to science and technology. These figures in the chart indicate, for instance, that for the French, German, Spanish,
Mandarin, Japanese and Italian languages, there are between two and almost four times more citations than for the Portuguese language.

Two apparent outliers in the chart above, namely Arabic and English, deserve though specific notes to help understand their figures. As for Arabic, the number of articles devoted to this language certainly includes the publications of the results obtained by the large number of research projects and initiatives that are funded by and in the United States of America (USA) and are specifically devoted to this language after the 2001 attacks to the World Trade Centre in New York in order to support intelligence activities.

As for English, it comes with no surprise that the prime language of the USA, the country that is the world economic and scientific superpower, and also of Australia, Canada and United Kingdom, is by far the object of the largest number of citations and thus of the largest volume of research effort. What might be not so expected, at least at first blush, is the maybe too wide difference between the number of citations to this language and the number of citations to the other languages.

This can be understood when the dynamics that are specific of the research profession is taken into account. To strive either as a successful doctoral student or as an advanced professional researcher, one should get as much publications as possible and with as much citations to them and impact as possible. Hence, doing research on the language for which there are more and more advanced data sets and previously developed processing tools — that is doing research on English — offers by far better chances to attract more attention and citations and achieve these desiderata.

These sociological constraints that are inherent to the scientific activity are thus inducing that the research work on English acts like a vortex that attracts into the technological preparation of this language much of the advantage brought by the funding that is being granted by authorities other than the USA’s to their researchers who do not have English as their national languages. The very wide gap between the research efforts devoted to English and to any other language is thus certainly supported by the USA’s economic and scientific capacity but also, mostly unknowingly and unwillingly, by all the other countries that are seeking to support their own languages and counterbalance the rampant dominance of English in the digital world. To an even more increased advantage of English and aggravated detriment of the other languages, this is most likely just another instantiation of the well-known vicious circles of economic and scientific underdevelopment.

### 3.3 Portuguese language in the digital age

| Among the approximately 7 000 languages in the world (Lewis, 2016), Portuguese is the 4th language with the largest number of speakers (Camões IP), being the working language of 27 international organizations. It is an international language with global projection and a major asset in terms of our national cultural heritage, sovereignty and geo-political projection. |
Its approximately 260 million speakers in 4 continents represent 3.6% of the world population (world map with countries and territories with Portuguese speaking population in figure above) and are getting technologically prepared for the digital age: Portuguese is the 5th language with the largest number of speakers as internet users, namely 131.9 million, which represent 3.9% of the internet users in the world; and whereas internet has an overall rate of penetration in the world of 46.4%, it has an overall rate of penetration of 50.1% in the countries with Portuguese as official language (Internet World Stats, 2016; data from 30 November 2015).

While speakers of Portuguese are getting technologically prepared, the thorough assessment in (Branco et al., 2012), together with further statistics like those compiled in Figure 8 above, clearly indicate that the technological preparation of the Portuguese language is lagging behind in comparison with the level of technological preparation of other languages with far less speakers or global projection, and is definitely not aligned with its status in the pre-digital world of international language with global projection.

The research on the technology for Portuguese has an inaugural landmark in 1979 with the doctoral dissertation of Helder Coelho, a professor emeritus of our department (picture of this dissertation’s cover in the picture below). Though this area has undergone a steady progress since then, it took some time until sufficient “critical mass” had been gathered.
Twenty years later, by the late 1990’s, when my doctoral research was completed, the number of data sets and processing tools for Portuguese that were available from previous research work and that could be reused and built upon to support new research advances and applications was still non-existent in the vast majority of relevant subareas. At that time, to be able to conduct with success their dissertations and research, investigators from either Portugal or Brazil had to turn mostly to English as the object language of their work. This was seriously detrimental for the Portuguese language and firmly counteracting this state of affairs was as major motivation and programmatic driver for my subsequent scientific activity, and for the setting up in 2002 of the Natural Language and Speech Group (**NLX-Group**) of the Department of Informatics of the University of Lisbon, Faculty of Sciences, the research group that I have been coordinating since then (group picture in the figure below).

**Figure 10:** Image of the cover of **Helder Coelho**’s (1979) doctoral dissertation, an inaugural landmark in the research on the technology for Portuguese.

**Figure 11:** Picture of the members of **NLX-Group** in 2015.
At present, almost four decades after Helder Coelho’s inaugural dissertation on deep language processing, there is a mature international community of researchers, mostly from Brazil and Portugal, but also from other countries, whose focus of activity is the technology for the Portuguese language (Branco et al., 2012: Chaps. 4.4-4.5). This community’s major focal point of organization and identity is the series of PROPOR conferences, PROPOR being the acronym of the International Conference on the Computational Processing of Portuguese.

This is a biennial event whose venue alternates between Portugal and Brazil and that adheres to the highest international scientific standards, with anonymous submission and reviewing, with an acceptance rate for submissions in the range of 25-35%, and with some 25 selected full papers at every edition published in the ISI indexed Springer’s Lecture Notes in Artificial Intelligence series. It is a full-blown international scientific meeting and networking event, covering a main conference and also a student research workshop, demos session, tutorials, job shop and collocated workshops, completed with best dissertations and best papers awards. I happen to be the general chair of its next edition, the twelfth one, which takes place in Tomar, Portugal, in 13-15 July 2016 (the image and url address of the respective website are presented in the figure below).
Figure 12: Image of the front page of the website of PROPOR2016-The 12th International Conference on the Computational Processing of Portuguese, available at http://propor2016.di.fc.ul.pt
Notwithstanding this important and qualified advancement, the community is fully aware that the support to and the volume of research on Portuguese is critically suboptimal in view of its key strategic relevance and of the very strong dynamics of competing languages, as thoroughly assessed in the *White Paper on the Portuguese Language in the Digital Age* (Branco et al., 2012).

To survive and prosper in the digital era, the Portuguese language needs to be properly and timely studied and prepared from a scientific and technological standpoint. Only in this way we will be able to ensure that all people, services and goods will be accessible in and through the information society to Portuguese speakers and it will be possible to ensure full citizenship for ourselves and for our culture in the society of the future.

Research on natural language offers challenges and results at least as fascinating as the study of the galaxy in the most distant sky. It turns out however that the responsibility, the priority and the urgency with respect to our own language are immeasurably larger and closer.

This strategic perspective has to a great extent guided the motivation, the goals and the choices in my research activity, whose results are partly being presented in this lecture.
PART II — Contributions

This second part of the lecture presents the research that has been undertaken by myself and under my supervision in this domain and is organized along three chapters.

The work on issues related to the deep language processing proper is gathered in Chapter 4 Deep language processing. It centred around the achievements obtained in the ongoing development of the computational grammar LXGram.

The results obtained concerning the instrumental datasets and processing tools are presented in Chapter 5 Companion, derivative and auxiliary datasets and processing tools.

Finally, language processing tasks and applications built on the results described in the previous two chapters are presented in Chapter 6 Language processing tasks and applications.
4 Deep language processing

An important part of our research devoted to deep language processing has been allocated to the development of LXGram, a computational grammar for the deep processing of Portuguese.

This grammar, in turn, has been an important stepping stone for further research work, namely in deep language processing proper, in the development of companion datasets and derivative processing tools, in the enhancing of natural processing tasks and in the advancement of natural language applications, all of which will be described in the sections below.

Figure 13: Pictures of João Silva (left) and Francisco Costa, doctoral and post-doctoral students under my supervision. They are two major co-authors of the results described in the present chapter.

This chapter is organized along the following structure.

Section 4.1 LXGram, a computational grammar for Portuguese introduces the computational grammar for the Portuguese language.

In Section 4.2 Exhaustive handling of nominal phrases, the advancements in the exhaustive treatment of nominal phrases are described.

In Section 4.3 Flexible handling of language variants, effective solutions for permitting deep grammars to cope with several language variants are reported.

Section 4.4 Robust handling of out of vocabulary items describes the results of the work on enhancing the robustness and text coverage of deep grammars.

This chapter closes with Section 3.5 Handling temporal meaning, with the reports on the progress obtained by enhancing the handling of temporal meaning with deep processing.

4.1 LXGram, a computational grammar for Portuguese

A computational grammar for deep language processing, the LXGram has been developed to take a Portuguese sentence as input and to deliver a thorough and
principled linguistic analysis of it, including its formal semantic representation, whose complete description can be found in the respective documentation (Costa and Branco, 2014).

**LXGram** code is written in the Linguistic Knowledge Builder (LKB) system, an open-source integrated development environment for the development of constraint grammars that includes a graphical user interface, a built-in parser, a debugger and a special purpose declarative specification formalism over a context-free grammar backbone (Copestake, 2002). This grammar encompasses the hand coding of linguistic generalizations, whose outcome for a given input sentence is a parse forest with possible grammatical representations for that sentence.

This is supplemented with a stochastic model for the resolution of parsing ambiguity, which linearly ranks parses. At the end of an analysis, this allows for the top-ranked parse to be selected and a single result to be returned instead of a full parse forest. The disambiguation module relies on a maximum-entropy model that is able to integrate the results of a variety of user-defined feature functions that test for arbitrary structural properties of analyses (Zhang et al., 2007).

**LXGram** follows the grammatical framework of Head-Driven Phrase Structure Grammar (HPSG) (Pollard and Sag, 1994), one of the most prominent grammar theories from Linguistics that is used in natural language processing. This permits a straightforward implementation of linguistic analyses of grammatical phenomena that are well linguistically grounded and have undergone thorough scientific scrutiny. It also has a positive impact in reusability and extendibility, with theoretical linguists being able to read and understand the (computational) grammar.

Natural language grammars in the HPSG framework associate grammatical representations to natural language expressions, including the formal representation of their meaning. Like several other computational grammars, **LXGram** uses Minimal Recursion Semantics (MRS) (Copestake et al., 2006) for the representation of meaning. An MRS meaning representation supports scope underspecification as it is a single description of a set of possible logic formulas that differ only in the relative scope of the relations present in these formulas. Additionally, the MRS format of semantic representation is well defined in the sense that it is known how to map between MRS representations and second order logic formulas, for which there is a set-theoretic interpretation.

**LXGram** has been in active development and in its 5th and current version it covers a wide range of linguistic phenomena, such as long distance dependencies, coordination, subordination, modification, many subcategorization frames, etc. Its lexicon contains 25 000 entries, covering both open and closed classes. In order to automatically select the most likely analysis of a sentence when the grammar produces multiple solutions, a statistical disambiguation model was trained resorting to the datasets described below in the next chapter.
To the best of our knowledge, this the first large-scale general purpose computational grammar for the deep linguistic processing of Portuguese, distributed under an open-source license, that delivers a thorough and principled linguistic analysis of sentences, including their formal semantic representation (an example of this grammatical representation for an example sentence is depicted in the figure below).

Figure 14: A 1 m x 1.4 m printout in font size 4 of the fully-fledged grammatical representation, in the AVM format of (Pollard and Sag, 1994), for the 6 word sentence Todos os computadores têm um disco ("Every computer has a disk") produced by the LXGram. The arm and the pen are included to help appreciate the relative proportion of the whole representation.
LXGram has its root in the grammar developed in 1999 in my PhD dissertation and this early implementation was then steadily developed with Francisco Costa in the scope of the FCT funded project GramaXing (2005-2007). Its development has been continued since then in a number of other projects, most notably in the FCT funded follow up projects SemanticShare (2007-2011) and DP4LT (2013-2016), and in the European project QTLeap (2013-2016), of which I was the coordinator.

The development of this grammar has been undertaken in the scope and with the support of the DELPH-IN consortium, an initiative that provides a variety of open-source tools to help deep grammar development by bringing together grammars for several languages and developers from different research laboratories across the world, including leading teams from Stanford University, Cambridge University, the German Research Centre for Artificial Intelligence (DFKI) and many others.5

As described in detail in the following Sections, LXGram also benefited from cutting edge improvements resulting from the research undertaken with postgraduate students under my supervision.

The grammar and its documentation are distributed from the LX-Centre,6 and is thoroughly documented in (more recent first):

Costa, Francisco and António Branco, 2014, A Computational Grammar for Deep Linguistic Processing of Portuguese: LXGram, version 5, University of Lisbon, Faculty of Sciences, Department of Informatics.


Besides the implementation reports, the development of the grammar (excluding the research results specifically presented below in the next Sections of this chapter) was reported in the following publications, including its seed predecessor (more recent first):


5 http://www.delph-in.net
6 http://lxcenter.di.fc.ul.pt
Additionally, a set of book chapters were published aiming at the dissemination of the grammar and deep language processing (more recent first):


4.2 Exhaustive handling of nominal phrases

Because of the different level of syntactic aggregation they empirically exhibit, expressions in a sentence group together into different units also known as constituency phrases. In any sentence where they may occur, phrases that can replace each other without loss of the grammaticality of the sentence are considered as belonging to the same set of phrases, also known as a category.

Expressions like the book on the table, every intelligent student, many of the several senators that missed the votation are a few examples that belong to one such category, namely the noun phrase category. Like the other expressions in this category, they characteristically share the property of their constituents assembling around one key constituent, which in this case is the head noun. For
instance, the nouns book, student and senators are the heads in the example noun phrases above.

Interestingly, provided an appropriate internal structure and an appropriate surrounding context in its sentence, a noun phrase may occur with its head absent. For instance, Mary’s and a verde are two such examples, respectively from English and Portuguese, when they occur in the expressions Although John’s friends were late, [Mary’s - ] arrived on time, and a casa azul e [a - verde] (literally: “the house blue and the - green”), where ‘ ’ was placed to symbolize the absent head noun.

This syntactic phenomenon represents an interesting challenge to computational grammars. They need to be designed to appropriately cope with the internal syntax of noun phrases and with the computation of their semantics even in the face of incomplete information in the input, namely when the heads around which they are crucially structured are missing.

Previous approaches to this problem in the literature, notably inspired by proposals from theoretical linguistics, stipulate that phonetically null nouns exist in the lexicon and that actually these occur as heads in such elliptical noun phrases, like the ones in the examples above, only that because they are phonetically null one cannot notice them.

In order to cope with this issue, and breaking with this reification of supposedly non-observable entities, we propose a novel grammatical analysis of noun phrases with so called elided nouns that dispense with the positing of phonetically null items in the lexicon. While tackling this specific elliptical grammar phenomenon, this approach seeks to promote and is in line with an emerging and more encompassing trend that defends a leaner grammar design that does not resort to "ghost" expressions not realized in sentences, including those usually related to other grammatical phenomena such as long-distance dependencies, anaphora, etc.

As explained in detail in the publications listed below, this is achieved by exploring in a novel, insightful way the information already stored in the lexical items that happen to be realized around the missing head noun. This involved redesigning how that information is factorized in terms of the relevant underlying type hierarchy and how this can be used to adequately constrain syntactic and semantic composition for this kind of natural language expressions.

Concomitantly, by means of a constituency rule that is specific to noun phrases with missing nouns, it is possible to maintain the structural syntactic parallelism between noun phrases with expressed noun heads and noun phrases with missing noun heads. With this in place, it turns out to be enough to do without the positioning of a phonetically null lexeme.

This work on nominal phrases was undertaken within the scope of the FCT funded project GramaXing (2005-2007), of which I was the coordinator, with Francisco Costa, who integrated its result as a central contribution of his MA dissertation.

It is worth noting that the masters work of Francisco Costa under my supervision was instrumental in obtaining the results described above. On the one hand, he extended the LXGram to allow it to cope with the deep processing of non-elliptical noun phrases following established proposals from the literature, and this served as the basis to experiment with the approach developed for processing
noun phrases with missing head nouns. On the other hand, he gathered exhaustive test suites that allowed this proposed approach to be evaluated and eventually ensured that it fully copes with them.

The results of this research line are reported in detail in the following publication:


They were also integrated in the following dissertation:

Costa, Francisco, 2007, Deep Linguistic Processing of Portuguese Noun Phrases, Masters dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.

4.3 Flexible handling of language variants

Any language with a reasonable number of speakers has what is known as language variants, which among themselves share the vast majority of their lexicon and core grammatical regularities and may have inessential differences, as it is the case, for example, with British, American and Australian variants of English.

For the benefit of generalization and grammar development economy, it is desirable for a grammar to handle different language variants. However, without any additional measures, the more a grammar handles different variants the larger the search space for the parser will be and the more spurious the analyses that are eventually produced by it will be for a given input expression. That expression will receive the parses from the grammar in accordance with the variant it belongs to, and eventually other parses as if it would pertain to other variants as well. Mutatis mutandis, an analogous situation will occur when the grammar is used for generation.

To make this concrete, consider the following illustrative examples. A grammar that handles European and American Portuguese should have two lexical entries for the noun policial, one corresponding to the European variant (whose meaning is equivalent to the meaning of the English criminal novel) and the other corresponding to the American variant (equivalent to policeman). A given input sentence with that noun would thus receive at least two analyses, one of which will be spurious.

By the same token, such a grammar would be able to cope with sentences with clitic pronouns to the left of the verb (in so called proclisis) and to the right of it (in so called enclisis). However, if the grammar is in generation mode and the variant
of interest is the American one, only one of the options, viz. proclisis, should be aimed at. Naturally, many other illustrative examples, even more complex and intertwined and with more serious consequences for the efficiency of the grammar, can be found.

In order to cope with this issue, we designed a strategy for deep language grammars to appropriately handle language variants, thus addressing a challenge that, to the best of our knowledge, has remained untackled in the literature. Under this design, a grammar is prepared (i) to detect the language variant a given input belongs to and (ii) accordingly, restricted itself to which language variant it should operate with.

As explained in detail in the publications listed below, this is achieved by extending every sign in the grammar with a feature whose value adequately records information on the variant that that sign may belong to. This solution relies on the design of an appropriate type hierarchy for the possible values of that feature so that these values constrain the combination of signs by means of the unification of that feature value among all signs in the parse tree of a given input sentence.

Concerning the detection of the language variant, there is an alternative to the self-tuning approach, that consists of a pre-processing step resorting to the best statistic methods previously developed at length in the literature to ascertain the language variant that any given input expression may belong to.

These two approaches, self-tuning and pre-tuning, were evaluated under comparable terms, with our proposed solution performing competitively in line with the best result that the pre-tuning method can ensure.

Interestingly, as an additional contribution of our work on the preparation of the datasets for this evaluation, we gather insight into the quantitative extent to which the written variants, European and American, of Portuguese may differ, which to the best of our knowledge is unheard of in the literature. For instance, among several other results, we found that over 80% of the sentences in our dataset are common to both language variants, and that about 1/3 of the differences are merely orthographic and will disappear when the unified orthography is adopted.

This work on language variants was undertaken within the scope of the FCT funded project GramaXing (2005-2007), of which I was the coordinator, with Francisco Costa, who later on became my PhD student.

The results of this research are reported in detail in the following publications (more recent first):


4.4 Robust handling of out of vocabulary items

Deep processing grammars produce highly sophisticated, fully-fledged grammatical representations for input sentences entered in the form of raw strings of characters. This is made possible because the necessary information to bridge from one end to the other had been previously encoded in the rules and in the lexicon of the grammar. This information represents knowledge about grammatical regularities, which are available for selection based on the effect of the words in the input sentence and can be combined into its final grammatical representation with the help of an underlying parsing algorithm.

For a given input sentence to be mapped to its appropriate grammatical representation, it is thus necessary that the regularities it instantiates are represented in the grammar. It is also necessary that every input word is represented in the lexicon and associated with relevant complex grammatical information. In deep processing grammars like LXGram, this is ensured by and encoded in a (deep) type of the hierarchy of lexical types, which is part of the overall computational grammar. Failure to meet these requirements results in a failure by the grammar to produce an adequate output or even any output at all for the input sentence.

In order to have a grammar with full text coverage, i.e. that is able to cope with any syntactically well-formed input, it is therefore imperative that its lexicon is complete, by including all possible input lexical expressions. This is, however, a desideratum that ultimately is not practically feasible because of the characteristic lexical creativity that is intrinsic of natural languages: as part of the language change process, the lexicon of a natural language is in permanent change, with new expressions being added and with old expressions changing their syntactic and semantic properties.

This is one of the major reasons that deep processing grammars typically do not ensure full text coverage. For instance, the English Resource Grammar (Flickinger, 2002; 2010) has been under development for the last three decades and is deemed to be the deep processing grammar with the widest text coverage. It is reported to have some 85% as its rate of coverage when running on unrestricted English text.7

7 Personal communication by Dan Flickinger at the DELPH-IN Summit 2014.
In order to address this problem and increase the robustness of deep grammars, we developed an approach that resorts to an auxiliary pre-processing step in which, at runtime, any out of vocabulary word is tentatively assigned a deep grammatical type that is predicted using a machine learning based classifier trained over a treebank.

To a certain extent, this is in line with previous contributions to this problem in the literature, where this approach is placed alongside the well-studied task of POS tagging. In POS tagging the goal is to select, and assign, the appropriate POS tag (e.g. Noun, Verb, etc.) for each occurrence of a word from the possible tags that that word may bear in the lexicon. In the literature, an important difference between this type of POS tagging and the task of assigning a deep type is the size of the tag set. The cardinality of the tag set for POS tagging is around a few dozen tags, whereas for deep type assignment is several hundred tags, which has let it to being termed as supertagging.

Supertagging has primarily explored the techniques developed for POS tagging, and has resorted to the same kind of information for its operation, which consists of a very small window of context that typically encompasses up to two words before and two words after the word to be (super)tagged.

Our contribution crucially departs from previous approaches in not assuming that the key difference between POS tagging and supertagging with deep types has to do with the size of their respective tag sets. Rather, we take into account the very essence of the deep lexical information encoded in a deep type, as opposed to the shallow information registered in a POS tag. While the latter encodes the category of a word resulting from its syntactic distribution, deep types encode highly complex linguistic information that includes, among many other things, the subcategorization frame of a word that determines its ability to shape the wider sentential vicinity around it in terms of the complements it can take and modifiers that may be licensed.

Accordingly, to support the training and decision of the classifier, our approach departs from the linear n-gram context and resorts to the much more complex, and informatively adequate, structured grammatical dependency graph (with complements and modifiers represented) that springs out of a targeted word to be supertagged when it is the head of that graph.

An important challenge when assuming this novel perspective relies on the technical means used to implement it. As explained in detail in the publications listed below, we used a tree kernel to obtain a vectorial representation of the grammatical dependency graphs. We used also support-vector machines (SVM) that are trained on that representation and support the classifier that predicts which deep type should be assigned to the targeted word. As an SVM is a binary classifier, the problem must first be binarized and so we opted for a one-vs-one binarization scheme, where multiple classifiers are created, each being responsible for discriminating between a pair of classes such that the class with the overall largest number of votes is eventually chosen.

To assess this proposal, we designed and explored a thorough experimental space, taking into account grammars for different languages, with different coverage and tag sets, training datasets of different sizes and deep types of
different frequency, etc. This experimental space permitted to undertake both an intrinsic and an extrinsic evaluation of the new approach developed.

In the intrinsic evaluation, the performance of the underlying supertagger was comparatively assessed with respect to other state of the art supertagging techniques. Here, what matters is from those cases that there are to solve — that is from those lexemes that are there to receive a deep type —, how many got a correct solution. This permitted to find out that the new solution we developed consistently outperforms the state-of-the-art, with different competitive gains depending on the different experimental settings that may amount to over 2 percentage points, when the competing state of the art approaches are already scoring in the range of 88%-94% accuracy.

In the extrinsic evaluation, the performance of the underlying supertagger was assessed in order to clarify its impact in the parsing by the grammars where it may be integrated. As parsing by a deep grammar is influenced by a myriad of interacting factors related to its inner workings, it might happen that upon the integration of the supertagger in the grammar, its impact might be mitigated — or in the limit even cancelled off — in these interactions. This exercise of extrinsic evaluation permitted to clarify that this is not the case and that the solution developed for out-of-vocabulary words helped to improve the performance of deep grammars.

This work was undertaken with João Silva in the scope of the doctoral research he completed under my supervision, and in the scope of the FCT funded project DP4LT (2013-2015), which I coordinated.

The results of this research are reported in detail in the following publications (more recent first):

Silva, João and António Branco, in preparation, "Deep Language Processing of Out-of-Vocabulary Words".


They were also integrated in the following dissertation:

Silva, João, 2014, Robust Handling of Out-of-Vocabulary Words in Deep Language Processing, Doctoral dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.
4.5 Deep handling of temporal meaning

From an input sentence under the form of a linear sequence of words, e.g. *Every one of the men loved a woman*, deep processing grammars produce its semantic representation under the form of a graph-structured sequence of symbols from some logical language, for instance:

\[ \text{utterance}(e1) \land \text{every}(x, \text{man}(x), a(y, \text{woman}(y), \text{love}(e2, x, y))) \land \text{before}(e2, e1) \]

A somewhat intriguing angle from which to underline the challenge involved in this task is to note that several words in the input (e.g. *every one of the*) may actually be projected into one symbol in the output, and that several symbols in the output may represent one word in the input (e.g. *loved*).

Interestingly, as it is notoriously the case that the verbal inflection renders important information on temporal relations, a number of expressions in the semantic representation account for the meaning of just parts of each of the natural language words, as is the case with the inflection suffix *-ed* in the example above. In fact, for a highly inflective language like Portuguese, deep language processing faces a non-trivial challenge in the adequate handling of verb inflectional morphology.

Verb inflection morphemes convey two-level information. On the one hand, an inflectional suffix encodes a so called feature bundle, which is a collection of grammatical features concerning temporal relations, namely grammatical tense, aspect and mode. For instance, in *remaste* ("you rowed"), the suffix *-aste* stands for the bundle with the tense value "pretérito", the aspect value "perfeito", and the mode value "indicativo".

On the other hand, a feature bundle expresses a range of semantic relations among events and temporal entities. For instance, the grammatical feature bundle "pretérito perfeito indicativo" can convey the information that the event being described by the verb occurred, and was terminated, before the utterance time.

Concomitantly, verb inflection morphemes bear a double level ambiguity. On the one hand, the same inflectional suffix may stand for more than one grammatical inflection feature bundle. For instance, in *comemos* ("we ate/eat"), depending on the context, the suffix *-emos* stands either for the feature bundle "pretérito perfeito indicativo" or for the "presente indicativo".

On the other hand, the same grammatical inflection feature bundle may convey different semantic temporal relations. For instance, depending on the context, the "presente indicativo" can convey the information that the time of the event described overlaps with the current time of the utterance — as in *Treinamos agora com música / "We are training now with music" —, or it can convey that the event time is after the current time of the utterance — as in *Amanhã treinamos com música. / "Tomorrow we will train with music".*
In order to address these challenges in a principled way, and to enlarge the coverage of the grammar in order for it to adequately handle tense and aspect, we adopted a divide and conquer approach and treated each of the two levels mentioned above separately. The determination of the grammatical feature bundles is handled in a lemmatization module that is pipelined with the grammar proper, which in turn maps the lemma and the grammatical feature bundles received as input into temporal relations.

This eventually led to several contributions. By its nature, the lemmatization module turns out to be quite specific to Portuguese, and to the best of our knowledge, is the only work published in the literature on this topic, which is complex and of paramount importance for this language.

The solution designed for the semantic representation of tense and aspect, in turn, while making instrumental use of the Portuguese grammar for its deployment, is of general relevance for any language, and offers clear improvements with respect to the state of the art. These two contributions will be addressed in turn.

For the lemmatization task, we undertook a detailed study and characterization of the problem space and set up a two-step lemmatizer module, the \textbf{LX-Lemmatizer}. The first step proceeds with the parsing of the verb form on the basis of a set of regular expressions whose outcome is filtered down to retain only grammatically correct lemmas and feature bundles by means of the reverse operation of running a conjugator tool over that outcome. The second step implements a quite straightforward algorithm, centred on the most frequent feature bundle given a set of admissible inflection-driven feature bundles for the input verb form. This approach ensured accurate results (around 96.5\%) that outperform those obtained using more sophisticated machine learning approaches that had been used for lemmatization modules in other inflective languages (around 92\% accuracy).

Interestingly, as one of its aside results, this work allowed us to provide an original, detailed and quantitative characterization of the Portuguese verb inflection system that was previously missing in the literature.

Concerning the mapping between grammatical tense and aspect feature bundles and their semantic representation, the grammar was extended by judiciously engineering the combination of a number of more appropriate mainstream results and proposals in the literature on formal semantics of natural languages. We adhered to a Davidsonian representation for events and a Reichenbachian approach for the representation of the meaning of tense and aspect. Its evaluation is presented in detail in the Section 6.2 below and shows that it contributes to improve the state of the art.

This work was done with \textbf{Filipe Nunes}, in the scope of his masters research under my supervision, and with \textbf{Francisco Costa}, in the scope of his doctoral research under my supervision. It was undertaken with the support of the FCT funded projects \textbf{TagShare} (2003-2005), \textbf{SemanticShare} (2007-2011) and \textbf{DP4LT} (2013-2016), which I coordinated.

The results of this research are reported in detail in the following publications (more recent first):

\begin{itemize}
  \item [40] In order to address these challenges in a principled way, and to enlarge the coverage of the grammar in order for it to adequately handle tense and aspect, we adopted a divide and conquer approach and treated each of the two levels mentioned above separately. The determination of the grammatical feature bundles is handled in a lemmatization module that is pipelined with the grammar proper, which in turn maps the lemma and the grammatical feature bundles received as input into temporal relations.
  \item[40] This eventually led to several contributions. By its nature, the lemmatization module turns out to be quite specific to Portuguese, and to the best of our knowledge, is the only work published in the literature on this topic, which is complex and of paramount importance for this language.
  \item[40] The solution designed for the semantic representation of tense and aspect, in turn, while making instrumental use of the Portuguese grammar for its deployment, is of general relevance for any language, and offers clear improvements with respect to the state of the art. These two contributions will be addressed in turn.
  \item[40] For the lemmatization task, we undertook a detailed study and characterization of the problem space and set up a two-step lemmatizer module, the \textbf{LX-Lemmatizer}. The first step proceeds with the parsing of the verb form on the basis of a set of regular expressions whose outcome is filtered down to retain only grammatically correct lemmas and feature bundles by means of the reverse operation of running a conjugator tool over that outcome. The second step implements a quite straightforward algorithm, centred on the most frequent feature bundle given a set of admissible inflection-driven feature bundles for the input verb form. This approach ensured accurate results (around 96.5\%) that outperform those obtained using more sophisticated machine learning approaches that had been used for lemmatization modules in other inflective languages (around 92\% accuracy).
  \item[40] Interestingly, as one of its aside results, this work allowed us to provide an original, detailed and quantitative characterization of the Portuguese verb inflection system that was previously missing in the literature.
  \item[40] Concerning the mapping between grammatical tense and aspect feature bundles and their semantic representation, the grammar was extended by judiciously engineering the combination of a number of more appropriate mainstream results and proposals in the literature on formal semantics of natural languages. We adhered to a Davidsonian representation for events and a Reichenbachian approach for the representation of the meaning of tense and aspect. Its evaluation is presented in detail in the Section 6.2 below and shows that it contributes to improve the state of the art.
  \item[40] This work was done with \textbf{Filipe Nunes}, in the scope of his masters research under my supervision, and with \textbf{Francisco Costa}, in the scope of his doctoral research under my supervision. It was undertaken with the support of the FCT funded projects \textbf{TagShare} (2003-2005), \textbf{SemanticShare} (2007-2011) and \textbf{DP4LT} (2013-2016), which I coordinated.
  \item[40] The results of this research are reported in detail in the following publications (more recent first):
\end{itemize}


They were also integrated in the following dissertations:


5 Companion, derivative and auxiliary datasets and processing tools

Collections of textual materials, also know as corpora, whose expressions are explicitly represented in association with linguistic information they convey are of paramount importance for research and development in natural language processing. Also known as annotated corpora, these datasets are indispensable for the training of statistically based language processing tools and applications, the evaluation of both statistical and symbolic tools, the study of natural language properties and the discovery of linguistic generalizations.

Expressions in annotated corpora are marked with tags or constructs that make explicit the language regularities that the respective expressions happen to instantiate. These regularities may concern any linguistic dimension, from morphology to discourse, and expressions of any size, from word affixes to texts. The annotations can be accurate, as the result of being assigned manually by human experts, or can be approximate, as the result of being assigned by some suboptimal automatic procedure. Naturally, the more accurate the annotation is the higher the reliability of the annotated corpus and the more useful it will be, as well as the more expensive it will be to produce.

Figure 15: A group photo of members of the European R&D network of excellence Multilingual European Technology Alliance (META-NET) in Berlin in 2011.

The European R&D network of excellence Multilingual European Technology Alliance (META-NET) undertook the initiative of mapping the level of preparation of European languages for the digital age (a group picture of the members of this network is displayed in the figure above). The results of this assessment are published in the international publisher Springer as a series of white papers, each one devoted to a different language. This series includes a white paper on the Portuguese language, produced by a team of Portuguese and Brazilian language technology experts, of which I am the first author:
The results reported in this book supported a number of other related publications (more recent first):


This initiative enabled a comparative assessment to be gathered, where the Portuguese language lines up in the group of languages for which less research has previously been devoted to. As a consequence, when compared to the language that has been by far the most common focus of research effort worldwide, the English language, Portuguese is considered one of the less-resourced languages, with a comparatively inferior number of datasets and processing tools available for it.

While this represents a detrimental situation in terms of the technology for Portuguese and its preparation for the digital age, it is nevertheless a much-improved situation when compared to the state of affairs towards the end of the 1990's. At that time, when my doctoral dissertation was concluded, the number, size and sophistication of datasets and of their companion processing tools for the computational processing of Portuguese were virtually non-existent. Against this background, I assumed as one of the key priorities of my research activity to contribute to rectify this situation, and thus this became also one of the motivations to set up the research group I have been coordinating since then, given that the production of this kind of datasets requires a collective, multidisciplinary and long-lasting effort. An overview of that endeavour, as it stood in 2009, can be found in the following publication:

Branco, António, 2009, "Research and Development on Natural Language Technology at the University of Lisbon, Department of Informatics: The NLX Group", In Proceedings, I Iberian SLTech - I Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages, Porto Salvo, Portugal, 3-4 September 2009, pp.121-122.
In the present chapter, the results of the work on the datasets that I have been promoting, coordinating, developing and curating with my team will be introduced (pictures of the team are displayed in the figure below). Likewise, we will also introduce the results on the companion, derivative and auxiliary tools that have been put forward. The first sections below are grouped into two pairs, the first pair on shallower language processing, the second on deeper processing. In each pair, datasets and tools will be addressed. This chapter is completed with a fifth section on multilingual datasets.

In terms of its unparalleled variety, volume, depth, reliability, compliance with de facto standards and documentation of its elements, as well as the compatibility among them, this collection stands out, in our view, as the most prominent collection of datasets and processing tools currently available for the computational processing of the Portuguese language.

Additionally, to the best of our knowledge, it is one of the very few collections, if not the only one, for any language, whose datasets can be searched and processing tools can be experimented with through freely available online services, which are gathered in the LX-Centre.

In order to present this work, this chapter is organized in the following sections, each describing a collection of data sets or processing tools, except the first section that presents a data set, the CINTIL corpus, that was the stepping stone for the development of all the other data sets and companion tools:

Section 5.1 CINTIL-International Corpus of Portuguese
Section 5.2 LX collection of shallow processing tools
Section 5.3 CINTIL collection of treebanks
Section 5.4 LX collection of deep processing tools
Section 5.5 *PT collection in multilingual datasets

The LX collection of deep processing tools (section 5.4) was mostly trained and evaluated on the LX collection of treebanks (section 5.3). The LX collection of shallow processing tools (section 5.2), in turn, which is necessary for the preprocessing stages of the deep tools, was trained and evaluated mostly on the CINTIL corpus (section 5.1).

Figure 16: Pictures of the NLX-Group in 2011 (left) and in 2015 with many of the co-authors of the results introduced in the present chapter.
5.1 CINTIL-International Corpus of Portuguese

The **CINTIL-International Corpus of Portuguese** contains approximately 1 Million tokens, of which approximately 1/3 belong to transcribed spoken materials. The other 2/3 are from written sources belonging mostly to the news domain (approx. 60% of this subcorpus) and to novels (approx. 25%).

The **CINTIL** corpus is accurately manually annotated with several types of linguistic detail, including information on sentence and paragraph splitting, tokenization, part-of-speech, inflection features, lemma, multi-word expressions of closed categories, delimitation and semantic type of named entities (an example annotated sentence is displayed in the figure below). In our view, it is the largest, the most accurate, the most rich in linguistic information and the most reliable data set of its kind publicly available for Portuguese.

![Image of a sentence in the CINTIL corpus]

**Figure 17:** A sample of the internal representation in the annotated CINTIL corpus with the example sentence *A sentença estava há muito tempo anunciada: a APDL não iria aceitar a proposta do único concorrente ao concurso de concessão dos terminais de contentores.* ("The judgment was announced long time ago: the APDL would not go to accept the proposal from the only candidate to the call for the concession of the container terminals.").

It was developed in close cooperation with teams of linguistic experts from CLUL-Centro de Linguística da Universidade de Lisboa under the coordination of **Amália Mendes**.

A first version was finalized within the scope of the FCT project **TagShare** (2003-2005), of which I was the coordinator, and is available through ELRA-European Language Resources Association. It went through a process of revision and correction within the scope of the European project **METANET4U** (2011-
2013), of which I was also the coordinator, and the resulting second version is available through the META-SHARE platform.8

The corpus can be searched without downloading a copy of it, through the online service **CINTIL Concordancer**.9

The results of this work are reported in detail in the following publications (more recent first):


Barreto, Florbela, António Branco, Amália Mendes, Fernanda Bacelar Nascimento and João Silva, 2005, **CINTIL-Corpus Internacional do Português: Annotation Manual**, University of Lisbon, Faculty of Sciences, Department of Informatics, version 6.0, 13pp.


### 5.2 LX* collection of shallow processing tools

By supporting their training and/or evaluation, the CINTIL corpus allowed a range of language processing tools to be developed that address the major canonical procedures of shallow language processing. To the best of our knowledge, this is a top-performing collection of tools for the computational shallow processing for the Portuguese language publicly available.

We will briefly describe them in turn.

**LX-Chunker** identifies the boundaries of sentences and paragraphs in Portuguese text. It seeks to cope with the ambiguity and the ambivalence of symbols that in some occurrences are indicators of separations among sentences and in other contexts are not. It is a hybrid tool, based on regular expressions and hidden Markov models, with an f-score of 99.94% obtained when evaluated

8http://metashare.metanet4u.eu
against a 12,000 sentence test set accurately hand tagged with respect to sentence and paragraph boundaries.

**LX-Tokenizer** identifies the boundaries of relevant word-level tokens in Portuguese text. It seeks to cope with the ambiguity of strings that in some contexts are single-word tokens and in some other contexts are contractions, i.e. double-word tokens. This tool achieves an f-score of 99.72% when evaluated with a 300 Ktoken portion of CINTIL in a 10-fold cross evaluation.

These two tools were developed with Joao Silva, who at the time was a Bachelors student of mine on the Natural Language Processing course and a member of our research team.

**LX-Conjugator** is a verbal conjugator, taking a Portuguese infinitive verb form as input and delivering the corresponding conjugated forms. It is the only available tool for fully-fledged Portuguese verb conjugation, including the full range of pronominal conjugation forms. Its capacity includes the handling of pronominal conjugation, compound tenses, double forms of past participles, past participle forms inflected for number and gender, negative imperative forms, and courtesy forms for second person. Given that it is based in principled linguistic generalizations captured by regular expressions and the appropriate lexica of affixes, it is the only available conjugator to handle neologisms. Their eventual faults have been correct along the time as it has been put to use, and at present no defect is known.

An initial version of this tool was developed by Francisco Costa in the scope of an assignment in a masters course taught by myself, and then further developed to its fully-fledged capacity under my supervision and with the help of Filipe Nunes and Tiago Henriques, who at the time were masters students and members of our research team.

**LX-Lemmatizer** is a verbal lemmatizer, taking a Portuguese verb form as input and delivering a ranked list of the corresponding lemmata (infinitive forms) together with inflectional feature values. To the best of our knowledge, it is the only tool with published results handling the full-fledged range of verb inflectional dimensions, as described above for the conjugator. Also, given that it is based in principled linguistic generalizations captured by regular expressions and the appropriate lexica of affixes, it is the only available conjugator to handle neologisms.

As mentioned above in Section 4.5, this lemmatizer follows a two-step approach where in the first step the verb form is parsed on the basis of a set of regular expressions whose outcome is filtered down to retain only grammatically correct lemmas and feature bundles by means of the reverse operation of running a conjugator tool over that outcome. The second step is centred on the most frequent feature bundle given the set of admissible inflection-driven feature bundles for the input verb form. When evaluated against a 300 Ktoken portion of CINTIL, it performs with 96,5% accuracy.

This work was done with Filipe Nunes, in the scope of his masters research under my supervision, and with the participation of Francisco Costa.

**LX-Inflect** is an integrated language processing tool for nominal lemmatization and inflection, taking a Portuguese word form that follows the
nominal inflection paradigm and an inflection feature bundle, and delivering both the corresponding lemma and the indication of the feature bundle, and the resulting form that conveys the feature bundle entered. It is based in principled linguistic generalizations captured by regular expressions and the appropriate lexica of affixes, thus handling neologisms. The lemmatization function has 97.67% f-score when evaluated against a 300 ktoken portion of CINTIL. To the best of our knowledge, this is the only available tool for Portuguese with this functionality and coverage.

This work was initiated by Pedro Martins and further pursued by João Silva, in the scope of their masters research under my supervision, and with the participation of Catarina Ribeiro and Ricardo Santos.

LX-Tagger is a language processing tool such that, for each word occurring in a text and from the possible different morpho-syntactic categories that word may have in the lexicon, assigns a single tag to it that indicates the morpho-syntactic category that it bears in that occurrence in the text. This tagger scored 96.8% accuracy when trained and tested under a 10-fold cross evaluation over a 260 Ktokens dataset.

This work was done with João Silva, in the scope of his masters research under my supervision.

LX-NER takes a segment of Portuguese text, identifies the expressions for named entities it contains and classifies them according to their semantic type. Furthermore, each named entity receives a standard representation. It is a hybrid tool by which the number-based entities are handled by regular expressions and the name-based entities by hidden Markov models. When evaluated against a manually constructed test-suite including over 300 examples, the number-based component scored 85.19% precision and 85.91% recall. When trained over a manually annotated corpus of approximately 200 Ktoken and evaluated against an unseen portion with approximately 50 ktoken, from CINTIL, the name based component scored 86.53% precision and 84.94% recall.

This tool was developed in cooperation with my colleague João Balsa, and with my students Eduardo Ferreira and Sara Silveira, and with the help of João Silva.

Complementing the development of this collection of language processing tools, and in order to foster the interest and activities in the area of science and technology for the Portuguese language, we set up freely available online language processing services powered by each of them. It is thus possible to experiment these tools, without downloading and installing them, through the respective online services, gathered at the LX-Centre, and also available at the specific addresses indicated in Annex C.

This research work was undertaken within the scope of the FCT funded projects TagShare (2003-2005) and QueXting (2005-2008), both under my coordination, and the European project LT4eL (2005-2007).

The results of this research work are reported in detail in the following publications (more recent first):


Branco, António and João Silva, 2003, "Morpho-syntactic Tagging without Training Corpus or Lexicon: How far is it possible to get?", In Proceedings, XVIII
5.3 CINTIL* collection of treebanks

The CINTIL-DeepBank contains approximately 17,000 sentences, corresponding to over 166,000 tokens, belonging to the news domain, part of them coming from the CINTIL corpus mentioned above.

In this corpus, sentences are accurately annotated with their fully-fledged grammatical representations, which encompass linguistic information ranging from morphology to semantics and include the representation of their meaning in a logical form the same style of representations that are output by the deep language processing grammar or Portuguese LXGram described above. It is a highly advanced treebank, being one of the very few of this kind reported in the literature, and the only one of this kind for the Portuguese language.

The development of this treebank follows a methodology that ensures the highest level of consistency and reliability for annotated treebanks. For consistency, the representations assigned to sentences are produced by a computational grammar, and thus any given grammar phenomenon that happens to be instantiated in different sentences always receives the same formal representation.

For reliability, a given sentence is associated with the appropriate representation — one of the many possible representations output by the grammar — in a double blind annotation followed by adjudication approach. This approach consists of independent annotation decisions made by two annotators followed by adjudication by a third annotator, who decides which representation to fix in the cases where the two annotators happened to have made different choices.

In order to ascertain the eventual reliability level of annotated datasets, inter-annotator metrics have become a requirement in the literature. For deep linguistic annotation specifically, however, no such metric existed and it had to be developed from scratch. This was a major outcome of the masters dissertation of Sérgio Castro, completed under my supervision.

It was then possible to assess the reliability of the CINTIL-DeepBank, which scores Kappa=0.86 which is over the threshold of 0.8 widely taken in the literature to ascertain an annotated dataset as reliable.
It is also of note that this is a dynamic treebank. In the case of further development of the grammar, and the need to assign a revised representation to some grammatical phenomenon that is present in an already treebanked sentence, such a sentence can be easily singled out and reannotated with minor effort by the human annotators. This is made possible by the annotation workbench used.

Given the fully-fledged grammatical representations associated to sentences in the CINTIL-DeepBank (an example is depicted in Figure 14 above), it was both possible and convenient to obtain other treebanks from it, of types that are very common in this research area and correspond to the needs of various types of supporting processing tools. These treebanks contain the same sentences but are associated to only specific part of the information contained in their deep representation:

**CINTIL-Treebank** contains around 17 000 sentences accurately annotated with trees representing syntactic constituency;

**CINTIL-PropBank** gathers 17 000 sentences accurately annotated with trees representing syntactic constituency and decorated with grammatical functions and semantic roles (an example is depicted in the figure below);

**CINTIL-DependencyBank** is the collection of 17 000 sentences accurately associated to graphs representing grammatical dependencies, whose arcs are decorated with grammatical functions and semantic roles (an example is displayed in the figure below);

---

**Figure 18**: An example with the sentence *Entre os sete presos, há cidadãos dos Estados Unidos, da China e da Formosa* ("Among the seven persons arrested, there are citizens from the United States, from China and from Taiwan") and its pretty-printed constituency tree decorated with grammatical functions and semantic roles as the sentences annotated in the CINTIL-PropBank.
Figure 19: The same example sentence of previous figure now annotated with its pretty-printed dependency graph, as the sentences in the CINTIL-DependencyBank.

CINTIL-DependencyBank PREMIUM is a collection of 3 000 accurately annotated sentences, with some 79 000 word tokens, similar in design to the previous one and differing from it in the sentences that were treebanked and in the circumstance that the support tool to draw the graphs is not the LXGram but the full coverage LX-DependencyParser (presented in the next Section below).

CINTIL-LogicalFormBank is a corpus of 17 000 sentences accurately annotated with logical forms representing their meaning in MRS format (an example is presented in the figure below);

Figure 20: An example of the pretty-printed attribute-value matrix (AVM) diagram with the MRS logical form of the sentence Deverá ser apresentado aos tribunais no Outono. ("He should be presented to court in Autumn."), a sentence and its logical form in the CINTIL-LogicalFormBank.
As reported in the literature, treebanks of these different types go through separate processes of development. To the best of our knowledge, our work was the first to follow this new methodological approach just described above, with unheard of and outstanding advantages over previous practice. Instead of developing several different treebanks, we are able to develop one treebank, namely **CINTIL-DeepBank**, from which others were extracted as specific vistas of it — namely **CINTIL-TreeBank**, **CINTIL-PropBank**, **CINTIL-DependencyBank** and **CINTIL-LogicalFormBank**, thus hugely speeding up their construction and reducing development efforts.

This new methodology is also the first to ensure full consistency across the treebanks, where for example the grammatical dependencies assigned to a given sentence in one dataset are compatible with the constituency relations in another dataset, etc.

A first version of the treebanks in this collection was finalized in the scope of the FCT project **SemanticShare** (2007-2011), of which I was the coordinator, and is available through the ELRA-European Language Resources Association.

It went through a process of revision and correction in the scope of further projects, namely the FCT project **DP4LT** (2013-2015) and the European projects **METANET4U** (2011-2013) and **QTLeap** (2013-2016), of which I was the coordinator, and the resulting subsequent versions are available also through the META-SHARE platform.

The development of the datasets has involved the contribution of a large team that included João Silva, Francisco Costa, Sérgio Castro, Catarina Carvalheiro, Rita Carvalho, Andreia Querido, Nuno Rendeiro, Marisa Campos, Rita Pereira, Catarina Correia, Patricia Gomes, Sílvia Pereira, Mariana Avelãs, Clara Pinto, Diana Amaral, Cláudia Martins, Joana Ramos, Sara Silveira, Patricia Gonçalves and Daniel Pereira.

The constituency, dependency and proposition treebanks can be searched without downloading copies of them, through the online service **CINTIL Treebank Searcher**.

The results of this research are reported in detail in the following publications (more recent first):


Branco, António, João Silva, Andreia Querido and Rita de Carvalho, 2015, **CINTIL DependencyBank PREMIUM Handbook: Design options for the representation of grammatical dependencies**, University of Lisbon, Faculty of Sciences.

---


and Language Technologies for Iberian Languages, Porto Salvo, Portugal, 3-4 September 2009, pp.107-110.

They are also part of the following dissertation:

Castro, Sérgio, 2011, Developing Reliability Metrics and Validation Tools for Datasets with Deep Linguistic Information, Masters dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.

5.4 LX* collection of deep processing tools

By supporting their training and/or evaluation, the treebanks in the CINTIL* collection allowed a range of language processing tools to be developed, the LX* collection of tools, that address the major canonical procedures of deeper language processing. To the best of our knowledge, this is a top-performing collection tools of their kind for the computational processing of the Portuguese language publicly available.

We will briefly describe them in turn.

**LX-Parser** is a stochastic parser that performs the syntactic analysis of Portuguese sentences in terms of their constituency structure. It seeks to cope with the so-called structural ambiguity where expressions may have different contributions to the meaning of its sentence and thus different underlying syntactic constituency structures depending on the different contexts where they may occur. It was trained with the Stanford Parser (Klein and Manning, 2003) software running over the CINTIL-Treebank. It achieves an F-score of 88% under the Parseval metric, in a 10-fold cross evaluation procedure.

This work was done with João Silva and Patricia Gonçalves.

**LX-DepParser** is a stochastic parser for Portuguese that for each input sentence delivers a graph connecting its words and whose directed arcs represent grammatical dependencies and the labels at the said arcs representing the grammatical function of those dependencies. Like any parser, it seeks to resolve so-called structural ambiguity. It was trained by applying the MaltParser (Nivre et al., 2006) over the CINTIL-DependencyBank. As for the evaluation of its performance, it obtained 91.21% in terms of labelled attachment score (LAS) in a 10-fold cross evaluation.

This work was done with Rúben Reis in the scope of his masters research under my supervision, and with the contribution of João Silva.

**LX-SRLabeller** is a parser that undertakes the constituency analysis of Portuguese sentences and the assignment of semantic role labels to their relevant constituents. Like the other parsers, it seeks to handle structural ambiguity. It is supported by the other two parsers described above, as its outcome results from the appropriate combination of key elements in their respective outcomes. It
achieves an f-score of 82% in terms of Parseval metrics, in a 10-fold cross evaluation.

This work was done with João Alves and Sérgio Castro.

As with the shallower processing tools and complementing the development of this collection of tools, we set up a freely available online language processing service powered by each of them. It is thus possible to experiment these tools, without downloading and installing them, through their respective online services, gathered at the LX-Centre, and also available at the specific addresses indicated in Annex C.

This research work was undertaken within the scope of the FCT funded project SemanticShare (2007-2011) and the European project METANET4U (2011-2013), both under my coordination.

The results of this research are reported in detail in the following publications (more recent first):


They are also part of the following dissertation:
5.5 *PT collection of multilingual datasets

On a par with the CINTIL* collection of treebanks, a second *PT collection of treebanks is being developed. Except in what concerns the sentences treebanked, this *PT collection is similar in all aspects to the CINTIL* collection: it includes the same range of individual sorts of treebanks — DeepBankPT, LogicalFormBankPT, DependencyBankPT, PropBankPT and TreebankPT —, extracted as different yet mutually consistent vistas from one of them — the DeepBankPT, with 3 500 sentences, covering 45 000 words —, which is a dynamic treebank whose development is supported by the same deep computational grammar — LXGram —, and which was accurately annotated by human experts under the same reliability savvy methodology — double blind annotation followed by adjudication.

The difference between the *PT collection and the CINTIL* collection lies in the corpus of sentences included in the first: that corpus is the result of the translation of the English Penn Treebank corpus into Portuguese.

The Penn Treebank is a 1 Million token treebank whose sentences belong to the financial news domain, more specifically to excerpts from Wall Street Journal articles. This has been the reference dataset traditionally used in the literature to support and evaluate the development of stochastic parsers for English. Other teams in the DELPH-IN initiative have also translated this corpus into other languages besides English and are working on treebanking the resulting corpus (Flickinger et al., 2012).

Constructing the *PT collection from this translated corpus and following this design model brings advantages of paramount importance. On the one hand, it allows for language processing tools for Portuguese to be trained and evaluated on materials with the same linguistic characteristics as the tools that have been developed for English, thus permitting a certain level of comparability between the research on two languages that may help to better uncover the specifics of each one of them and eventually more easily transfer results across the investigations on them.

On the other hand, it allows for the construction of multilingual, parallel and aligned treebanks, where DeepBankPT is aligned to DeepBankENG and to DeepBanks* being developed for other languages by teams in the DELPH-IN initiative. This is of great importance: firstly, because multilingual parallel treebanks are key assets for the progress on multilingual processing, in general, and machine translation, in particular, and are still very scarce language resources; secondly, because together with other deep treebanks gathered in the multilingual ParDeepBank, the DeepBankPT represents a pioneering effort to develop multilingual parallel treebanks with deep grammatical representations and thus...
opens up a whole new path towards cutting edge yet untapped research results in language processing.

At present each of the treebanks in the *PT collection contains over 44 000 tokens, corresponding to over 3 400 sentences. They are distributed through the META-SHARE platform.

The treebanks in this collection were developed within the scope of the FCT project **DP4LT** (2013-2015) and the European projects **METANET4U** (2011-2013) and **QTLeap** (2013-2016), of which I was the coordinator.

The development of the datasets has involved the contribution of a large team that included João Silva, Francisco Costa, Sérgio Castro, Catarina Carvalheiro, Rita Carvalho, Andreia Querido, Nuno Rendeiro, Marisa Campos, Rita Pereira, Catarina Correia, Patricia Gomes, Sílvia Pereira, Mariana Avelãs, Clara Pinto, Diana Amaral, Cláudia Martins, Joana Ramos, Sara Silveira, Patricia Gonçalves and Daniel Pereira.

The results of this research work are reported in detail in the following publications:


6 Language processing tasks and applications

The datasets and processing tools for the increasingly deeper language processing described above are the building blocks for a wide range of language processing tasks and applications of many different sorts.

On the basis of the resources that we have developed, described above, we undertook research on a number of topics that will be presented below. We focus primarily on those that resorted to deeper language processing approaches.

![Figure 21: Pictures of Sara Silveira (left) and João Rodrigues, doctoral students under my supervision. They are two major co-authors of the results described in the present chapter.](image)

This chapter is structured along the following sections:

In **Section 6.1 Annotation for the semantic web**, the results obtained in exploring language processing tools to annotate web pages with a representation of their meaning are introduced.

**Section 6.2 Temporal extraction** reports on the advancements obtained in the task of temporal extraction by means of the hybridization between deep and shallow processing approaches.

In **Section 6.3 Summarization**, the results obtained with the engineered application of our processing tools to the task of automatic multi-document summarization are described.

The **Section 6.4 Machine translation** closes this chapter with a presentation of our work on deep machine translation.

6.1 Annotation for the semantic web

In order to further capitalize on the potential of the web and the massive number of pages and information available therein, the semantic web has been fostered as an extension of the web that allows data to be shared and reused across applications. Such an extension relies on the annotation of web pages with
representations of their semantics. While this annotation resorts to a dedicated semantic formalism, the Resource Description Framework (RDF), this is typically formulated along the syntax of Extensible Markup Language (XML).

The delivery of the promise of the semantic web faces a number of highly non-trivial challenges, whose exhaustive listing and scrutiny is out of the scope of the present document, but among which it is relevant to underline here the need to find automatic means to annotate web pages with their semantic representation in RDF/XML.

At its core, searching for an automated mapping of raw text (whose content in a web page is accessible to human readers) into a representation of its meaning (accessible to further handling by artificial agents) is at the centre of the deep language processing endeavour. Accordingly, we proceeded with an exercise of exploring and engineering the resources and tools available in our group to support this automated mapping and thus the automated semantic mark up of web pages written in Portuguese in support of the semantic web undertaking.

In order to pursue this goal, we resorted to and extended the technology we had used to develop some of our deep processing tools. Firstly, we resorted to the CINTIL-PropBank, whose syntactic constituency trees are decorated with semantic roles assigned to the relevant constituents. We designed a tool that allowed this treebank to be converted into another dataset in the style of the CINTIL-DependencyBank. The result was a treebank where each sentence is associated with a graph, not of dependencies in terms of grammatical relations (e.g. Subject, Specifier, etc.), but of dependencies encoding the semantic roles between its words (e.g. Temporal modifier, Argument 1, etc.).

Secondly, we used this dataset to develop a semantic parser that assigns to an input sentence a graph made of triples that relate words by means of some semantic relation.

Thirdly, we designed a tool that converts from the triples-based output format of the parser to RDF/XML (an example is displayed in the figure below).
This exercise was done with Rúben Reis, as part of his masters research under my supervision, and with the participation of João Silva. It was undertaken within the scope of the FCT funded project SemanticShare (2007-2011), which I coordinated.

The results of this research were integrated in the following dissertation:

Reis, Rúben, 2010, Marcação Semântica de Páginas Web Apoiada por Parsers de Dependências Gramaticais, Masters dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.

### 6.2 Temporal extraction

Intertwined with the content conveyed by a natural language text, there is an intrinsic layer of temporal information. On the side of the linguistic form, this information is expressed by a wide range of grammatical means — including verb affixes conveying grammatical tense and aspect, time denoting noun phrases,
temporal adverbials and conjunctions, etc. —, and extra grammatical means — world knowledge, causality relations, calendar systems, temporal inference, etc. On the side of the semantic representation, the meaning of this layer should be captured by a number of adequate relations among time and event referring symbols from the semantic representation formalism.

Given the practical relevance of getting hold of the temporal information conveyed by a text, research on the specific language processing task of temporal processing has received substantial attention, with its goal being to map between raw text and the semantic representation of its temporal information.

This processing task has been pursued along research paths that can be somehow seen as differentiating among themselves given the stronger emphasis they put on the source of temporal information explored. On the one hand, deep language processing is specially tuned to address the temporal information conveyed by grammatical means. On the other hand, a research line termed as temporal extraction has advanced the exploration of non-grammatical sources by intensively resorting to machine learning classifiers.

With the aim of contributing to the advancement of this language processing task, we undertook a range of research activities that sought to build on the complementarities between the above two approaches and on their hybridization. We have thus explored combinations of them aiming at amplifying their strengths and mitigating their weaknesses. While this is in line with other works that explore hybrid combinations of deep and shallow methods, to the best of our knowledge, our endeavour stands out in the literature as pioneering a thorough application of this hybridization to the area of temporal processing.

We explored this methodological option to temporal processing along two key paths of progression. On the one hand, we sought to foster each approach, temporal extraction and deep processing, in isolation by further exploring their own potential and the vantage points from which each excels. On the other hand, we adopted a novel way of merging their contributions that does not resort to the usual pipelining scheme.

We address each in turn.

A careful assessment of the literature on temporal extraction revealed that two information elements that could very likely help had been neglected, namely aspect type and the logical constraints of a set of temporal relations.

Regarding aspect type, we designed a data-mining procedure to extract information from the web on verbs that permitted to assign to them their relevant aspect type, viz. state, process, culminated process or culmination. These values were then used as extra features to train classifiers that had been shown in the literature to support state of the art extractors. The result was a temporal extraction tool that performs above the state of the art.

Regarding logical constraints, we designed a system of rules that produce the closure of the set of temporal relations extracted from text under the logical constraints holding for these sorts of relations. As with aspect types, this enlarged set of relations supported the training of classifiers that permitted to improve on the state of the art.
A careful assessment of the literature on temporal deep processing revealed that the process could be improved, namely by a more adequate way of coping with some of the grammatical phenomena that may be particularly relevant in this respect. The improvement obtained along this line of research is already introduced above in Section 4.5.

As well as having contributed to the improvement of temporal extraction and of temporal deep processing, our work pursued further progress by the hybridization of these two approaches.

When hybridization between shallow and deep approaches has been used for other grammatical dimensions or processing tasks in the literature, it has resorted to the typical scheme of pipelining the shallow module, in some pre-processing fashion, with the deep grammar. Here we departed from this scheme, and resorted rather to a merging scheme whereby the semantic representations output by the deep grammar and by the temporal extractor are merged, thus complementing and refining each other’s outcome. Interestingly, the resulting hybrid system shows performance results that increase the quality of the temporal meaning representation and are better than the performance of each of the two components in isolation.

As an additional contribution of our work on the development of the resulting temporal processing system LX-TimeAnalyzer, the dataset TimeBankPT for training and evaluating the system was developed, with around 70 000 words (an example excerpt is depicted in the figure below). To the best of our knowledge, this is the only dataset annotated with standard temporal representations for the Portuguese language, which represents a major contribution for other researchers wanting to further pursue research on this topic.
ABC<TIMEX3 tid="t52" type="DATE" value="1998-01-14"
temporalFunction="false"
functionInDocument="CREATION_TIME">19980114</TIMEX3>.183
0.1611
REPORTAGEM
<s>Em Washington, <TIMEX3 tid="t53" type="DATE"
value="1998-01-14" temporalFunction="true"
functionInDocument="NONE"
anchorTimeID="t52">hoje</TIMEX3>, a Federal Aviation
Administration <EVENT
eid="e1" class="OCCURRENCE" stem="publicar"
aspect="NONE" tense="PPI" polarity="POS"
pos="VERB">publicou</EVENT> gravações do controlo de tráfego aéreo da
<TIME3 tid="t54" type="TIME" value="1998-XX-XXTNI"
temporalFunction="true" functionInDocument="NONE"
anchorTimeID="t52">noite</TIME3> em que o voo TWA800
<EVENT eid="e2" class="OCCURRENCE" stem="cair"
aspect="NONE" tense="PPI" polarity="POS" pos="VERB">caiu</EVENT>
</s>.

...<TLINK lid="l1" relType="BEFORE" eventID="e2"
relatedToTime="t53" task="A"/>
<TLINK lid="l2" relType="OVERLAP" eventID="e2"
relatedToTime="t54" task="A"/>
<TLINK lid="l4" relType="BEFORE" eventID="e2"
relatedToTime="t52" task="B"/>
...
</TempEval>

Figure 23: A sample of the internal representation of the annotated TimeBankPT with an excerpt with the example sentence ABC.1830.0611
REPORTAGEM Em Washington, a Federal Aviation Administration publicou gravações do controlo de tráfego aéreo da noite em que o voo TWA800 caiu. ("ABC.1830.0611 NEWS STORY In Washington, the Federal Aviation Administration released air traffic control tapes from the night the TWA flight eight hundred went down.").

This research was pursued with Francisco Costa, as part of his doctoral research under my supervision. It was undertaken in the context of the FCT funded projects TagShare (2003-2005), SemanticShare (2007-2011) and DP4LT (2013-2016), which I coordinated.

The results of this research are reported in detail in the following publications (more recent first):

Costa, Francisco and António Branco, 2013, “Temporal Relation Classification Based on Temporal Reasoning”, In Proceedings, 10th International Conference on Computational Semantics (IWCS), Potsdam, Germany, 19-20 March 2013, pp.54-64.


They were integrated also in the following dissertation:

Costa, Francisco, 2012, Processing Temporal Information in Unstructured Documents, Doctoral dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.

The temporal processing tool LX-TimeAnalyzer can be experimented in the demo available online11 (an image of this demo is depicted in the figure below).

11 http://lxtimeanalyzer.di.fc.ul.pt
6.3 Summarization

As informative texts are produced by their authors, these texts seek to attend to possible requirements and/or constraints holding at the moment of their production and also to eventually address some estimation by the authors about the profile of their hypothesized readers. Given the endless diversity of eventual readers, however, either in terms of the volume of information each one needs or in terms of the time each one can afford to acquire it, an informative text is likely to miss a perfect, or even a good match with the problem it seeks to address.

In order to address this problem, research on automatic summarization has been given considerable attention, with the goal of producing from an input text, or a collection of texts, another text with a length specified by the user that typically is
much shorter than the length of the input. This summary results from the concatenation of sentences that were selected from the input text according to some heuristics that seeks to optimize its representativeness of the information conveyed by the overall text being summarized.

Interestingly, no matter the sentences selected to be concatenated into a summary — and as a matter of fact even if such sentences are selected by humans —, the eventual human reader tends to find a resulting sequence of such sentences hard to read.

With the aim of contributing to the advancement of automatic summarization, we undertook a line of research that seeks to enhance the readability of the produced summaries by resorting to the post-processing of the sentences extracted to be included in the summaries. This post-processing combines the tasks of paragraph creation, of sentence reduction and of insertion of discourse connectives, performed over the extracted sentences.

To the best of our knowledge, our work was pioneering both in pursuing the task of paragraph creation, and in seeking to improve the usefulness of summaries for human end users with a combined contribution of the three aforementioned types of processing tasks. The tasks of sentence reduction and insertion of discourse connectives were also offered as innovative solutions in their own right. As in other research lines of ours, we pursued a hybrid approach that combines rule-based and data-driven approaches.

We will go through each one of them in turn.

The aim of the sentence reduction procedure is to compress the sentences extracted for summary by reducing them to component expressions that convey their key content. The challenge here is to give away with expendable information, and thus generate simpler and easier to read text likely to have a higher key information density, while keeping the full grammaticality of sentences.

To pursue this goal, we adopted a hybrid approach of combining a statistical constituency parser with a set of regular expressions based rules over trees that identifies and removes targeted syntactic constructs, such as appositions, parentheticals, etc. All of the possible reduced variants of an input sentence are ranked according to a relevance metric, and the best scoring one is kept in case its score overcomes the relevance score of the input sentence itself.

The objective of the paragraph creation procedure, in turn, is to gather the extracted and concatenated sentences into different groups that when separated among themselves by a blank line are displayed to the human end user as different paragraphs. The challenge here is a clustering one, consisting of placing together in the same paragraph all and only the sentences that are closely related to each other.

This challenge was addressed by extracting from the input texts to be summarized a set of representative keywords, and by then placing in the same cluster those sentences that are closer to the keyword of that cluster than to the other keywords. This quite straightforward approach produced results as good if not better than more excruciatingly sophisticated approaches. The keywords are selected from common and proper nouns by resorting to their tf-idf word score, and the clustering follows an adapted version of the k-means algorithm.
Finally, with the insertion of connectives, the goal is to enhance the readability of the concatenation of sentences extracted from different places in a given input text, and from different input texts, by inserting among them discourse connectives, such as however, consequently, on the contrary, etc. The non trivial challenge here is to uncover the discourse semantic relation between two given sentences such that a discourse connective expressing that semantic relation can be inserted between them to make this relation lexically overt and explicit, and produce a summary that is easier to understand.

As part of this work, a corpus of over 3 Million pairs of sentences associated with the discourse connective holding between the sentences in the pair, was developed. This was used to train a classifier based on support vector machines and a number of text features that, after experimentation, were shown to produce the best results. On the basis of the discourse relation assigned, there is a procedure to select the discourse connective that takes into account, among other things, its frequency in the training corpus and the frequency with which it was already inserted in the processed summary so far.

When evaluated against relevant baselines, our approach combining these three procedures shows a substantial enhancement of the resulting summary in terms of the appreciation by human end users, with the first two procedures contributing the largest share for said improvement.

This research was pursued with Sara Silveira, as part of her doctoral research under my supervision. The results of this research on automatic summarization are reported in detail in the following publications (more recent first):


Silveira, Sara and António Branco 2012, ”Extracting Multi-document Summaries with a Double Clustering Approach”, In Proceedings, 7th International Conference on Applications of Natural Language Processing to Information


They were also integrated in the following dissertation:

Silveira, Sara, 2015, Enhancing Extractive Summarization with Automatic Post-processing, Doctoral dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.

The multi-document automatic summarizer SIMBA was one of the results of this doctoral research work. It can be experimented in the demo available online\(^{12}\) (an image of this demo is depicted in the figure below).
Figure 25: Image of the interface with the online demo of the SIMBA multi-document automatic summarizer.

6.4 Machine translation

Given its highly non-trivial nature, machine translation (MT) is seen as a quintessential application in the realm of natural language processing. It is a computational process that seeks to deliver accurate translation of sentences from one natural language to another.

In the last decade, mainstream research on machine translation has benefited mostly from the significant advances obtained through the exploitation of increasingly sophisticated statistical approaches. To a large extent, this incremental advancement has also been achieved by encompassing a host of subsidiary and increasingly fine-grained linguistic distinctions that add to the surface level alignment on which these approaches are ultimately anchored.

It has been suggested in recent years, both in leading academic and industry circles, that the incremental progress towards quality MT may be asymptotically reaching a ceiling. The claim relies on the observation that such a curve tends to be
more evident as more fine-grained distinctions are needed to produce better translations with fewer gains in terms of quality increase.

The deeper the processing of utterances, the less language-specific differences remain between the representation of the meaning of a given utterance and the meaning representation of its translation. Further chances of success can thus be explored by machine translation systems that are based on deeper semantic engineering approaches. We have thus undertaken a line of work that seeks to explore deep language processing and the potential of its recent progress in order to obtain a methodological advancement that opens the way to higher quality MT.

To explore these research directions of interest, we resorted to TectoMT (Žabokrtský et al., 2008), a framework for the development of MT that is particularly suited to our purposes, and experimented with the translation between Portuguese and English in both directions. This framework supports advanced hybrid transfer-based MT where the source sentence analysis into and target sentence synthesis from deep representations can be performed by modules with any methodological orientation — statistical and/or symbolic — and the transfer of the deep representation then performed on a tree-to-tree basis using maximum entropy context-sensitive lexical translation models combined with hidden Markov tree models. Using TectoMT brings the additional advantage that it includes the top performing tools for the analysis and synthesis of English that can be reused when developing translation pairs involving English.

In order to set up the MT modules between English and Portuguese, we undertook the development of the synthesis component into Portuguese on the basis of the rule-based machinery made available by the TectoMT framework. As for the analysis component, we proceeded with the engineering of a pipeline of our processing tools for Portuguese, from tokenization until the deep representation. The transfer component, in turn, was trained on the basis of parallel data treebanked by appropriate deep grammatical analysers we developed on top of our pre-existing parsers for Portuguese.

As statistical machine translation has become the mainstream approach to MT, the MOSES framework (Koehn et al., 2007) has been widely used to develop state of the art MT solutions, both when further research on statistical MT is undertaken and when industrial solutions are needed. We have thus used this framework to develop state of the art statistical MT modules for the translation between Portuguese and English by resorting to the largest amount of parallel data available involving these two languages. As we used these modules as baselines to assess the deep MT modules we developed, it turns out that the latter display a competitive advantage with respect these baselines both in terms of automatic metrics and of human evaluation.

The output of the deep MT system was ranked as better than the output of the statistical MT baseline in 46% of the cases on average, and ranked as equally good in 27% of the cases on average, when they were comparatively assessed on a set of 100 input sentences and their respective output was evaluated by at least 3 different human evaluators for each sentence. When resorting to the BLEU automatic evaluation metric, our deep MT system scored 15.51 points, which are 1.76 points over the score of 13.75 obtained by the statistical machine translation baseline.
As another innovative contribution to this area, we enhanced the performance of the deep MT machinery by means of the integration and exploitation of word sense disambiguation. In order to train the lemma-to-lemma translation model from the parallel corpora, we included the unique identifiers of concepts expressed by source words and by their respective grammatical governors (after their lexical sense disambiguation), as well as the identifiers of the superconcepts of those concepts, as additional contextual features. We obtained a gain of 0.25 BLEU points with respect the baseline (without these extra features) when run over a 1 000 sentence test data set.

This work was undertaken within the scope of the European project QTLeap (2013-2016), under my coordination, in cooperation with the colleagues from the other partner institutions in the project. In our team, it was done with the members of this project João Rodrigues, Nuno Rendeiro, Andreia Querido, Sanja Štajner, Steven Neale, Luís Gomes, Rosa Del Gaudio and João Silva.

The baseline system was completed with João Rodrigues, as part of his masters research under my supervision.

The results of this research are reported in detail in the following publications (more recent first):


Štajner, Sanja, Andreia Querido, Nuno Rendeiro, João Rodrigues and António Branco, accepted, "Use of Domain-Specific Language Resources in Machine


Part of these results were also integrated in the following dissertation:

Rodrigues, João, 2015, *Speech-to-Speech Translation to Support Medical Interviews*, Masters dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.

The machine translation application **LX-Translator** was one of the results of this research work. It translates between Portuguese and English with written and spoken input, and can be experimented in the demo available online13 (images of the interfaces of this application are displayed in the figures below).

Figure 26: A picture of the **LX-Translator** mobile interface.

Figure 27: A picture of the **LX-Translator** web interface.
PART III — Outreach

This third part of the lecture reports on the exploitation of the results presented in PART II above in view of providing innovative solutions to research, societal and business challenges.

The Chapter 7 CLARIN research infrastructure reports on the research infrastructure set up to support the science and technology of the Portuguese language, which has the data sets and processing tool presented in the previous chapters as some of its major foundations.

Technological solutions for the business and public sectors built on the results described in the previous chapters are presented in the Chapter 8 Innovative solutions.
7 CLARIN research infrastructure

The collections of datasets and tools described above are a large part of the overall collection of language datasets and processing developed at our NLX-Group. The later also contains the following datasets for Portuguese, very briefly introduced here.

**CINTIL-Definitions** is an accurately annotated corpus (POS tags and morphological information) with an additional layer of annotation marking definitions, containing over 250 000 words.

**CINTIL-NamedEntities** is an accurately annotated corpus (POS tags and morphological information) of around 30 000 sentences with an additional layer of annotation where some 26 000 named entities are disambiguated by means of their association with Portuguese Wikipedia entries.

**CINTIL-QATreebank** is a treebank of 111 factoid questions and imperatives.

**CINTIL-WordSenses** is an accurately annotated corpus (POS tags and morphological information) of around 24 000 sentences with 45 000 words with an additional layer of annotation where a word is associated with the MWN.PT wordnet concept it expresses in the context of its occurrence.

**LX-NominalInflections** is a lexical lists with around 27 000 items associated to information concerning their nominal lemmatization and inflection properties.

**LX-VerbalInflections** is a lexical lists with around 800 000 items associated to information concerning their verbal lemmatization and inflection properties.

**LX-Abbreviations** is a list of annotated abbreviations in Portuguese with around 200 items.

**LX-Stopwords** is a list of categorized expressions from closed morphosyntactic classes.

**Nexing Corpus** is a corpus with the transcriptions of syllogistic reasoning protocols.

**QTLeap Multilingual Paralell Corpora** is a set of multilingual corpora of question and answer pairs in computer and IT troubleshooting domain collected through a real-life support service via chat for eight languages.

**QTLeap WSD/NED multilingual corpora** is a set of multilingual corpora for six languages, with several million tokens each, automatically annotated with several lexical semantics dimmensions.

**MWN.PT-Portuguese Multiwordnet** is a hand made, accurate lexical semantic network shaped under the ontological model of wordnets, with over 17 200 concepts covering 21 000 tokens and over 16 000 types of the European and American variants of Portuguese, aligned with wordnets of other languages.
The datasets listed above can be obtained from the META-SHARE platform and/or the **NLX-Group**.

They were developed with members of my research group and other colleagues mentioned in the respective publications and documentation.

The collection of processing tools and applications for Portuguese developed in the **NLX-Group**, in turn, further contains the following elements, very briefly introduced here.

**LX-Syllabifier** is a tool that marks the boundaries between syllables.

**LX-WSD** is a disambiguation tool that annotates the occurrence of an input word with the MWN.PT wordnet concept it expresses in its context.

**LX-NED** is a disambiguation tool that annotates the occurrence of an expression with the Wikipedia entry it refers to in its context.

**LX-Definitions** is a tool that identifies definitions in the input text.

**LX-CEFR** is a tool that for an input text, obtains scores for a range of linguistic metrics of that input and classifies it according the proficiency level of learners of Portuguese as second language in the standard scale of such proficiency levels.

**XisQuê** is a open-domain web-based question answering application for factoid questions.

**LX-ListQuestion** is a open-domain web-based question answering application for list questions.

These are complemented with the following research support tools:

**CINTIL-Treebank Searcher** is an online system to search for language patterns in the collections of syntactic trees of the CINTIL-Treebank.

**CINTIL-Concordancer** is an online system to search language patterns in the CINTIL corpus.

**LX-SenseAnnotator** is an offline system to support the accurate annotation of words with the concepts they express.

**MWN.PT Searcher** is an online system to search in the lexical semantic network MWN.PT.

Pointers for the tools listed above can be obtained from the Annex C. Most of them can be experimented, without downloading and installing copies of them, via de free online services they support, available at the **LX-Centre**.

Like for their companion datasets described above, these tools were developed with members of my research group and eventually other colleagues mentioned in the respective publications and documentation.

In this respect, the datasets **CINTIL-Definitions** and **CINTIL-QATreebank**, and the tools **LX-Definitions** and **LX-ListQuestion** deserve a special mention as they were developed in the scope of the doctoral research work, respectively, of Rosa Del Gaudio and Patricia Gonçalves, both under my supervision.

This work is reported in detail in the following publications (more recent first):

Gaudio, Rosa Del and António Branco, 2009, "Language Independent System for Definition Extraction: First results using learning algorithms", In
Proceedings, RANLP2009 Workshop on Definition Extraction (wDE2009), Recent Advances in Natural Language Processing (RANLP2009), Borovets, Bulgaria, 18 September 2009, pp.33-39.


They were also integrated in the following dissertations (more recent first):

Gonçalves, Patricia, 2015, Open-Domain Web-based Multiple Document Question Answering for List Questions with Support for Temporal Restrictors, Doctoral dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.

Gaudio, Rosa Del, 2013, Automatic Extraction of Definitions, Doctoral dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.

* The datasets and tools for deep language processing presented above in the previous chapters, together with the further elements of the overall collection of datasets and tools for the computational processing of Portuguese just listed in this Section, are a most relevant pillar of the CLARIN national research infrastructure, of whose proposal I was the coordinator, and of which I am currently the director general.

CLARIN is an infrastructure in the first Portuguese Roadmap of Research Infrastructures of Strategic Interest and is the Portuguese national node of the European Common Language Resources and Technology Infrastructure (CLARIN eric). CLARIN counts on the association of 21 research centres and institutions, from Portugal and Brazil, including Camões - Instituto da Cooperação e da Língua I. P., and makes resources and technology available and useful to scholars and experts from all disciplines whose topics of research, development or innovation concern human language, with special relevance to the humanities and social sciences, and to the cognitive and computation sciences. It also provides its users with top level technical support to fully exploit such scientific assets.

The collection of datasets and tools for Portuguese developed at the NLX-Group under my coordination, as well as all their companion free online services, acted as the compelling draft of the future infrastructure and were a key asset in the justification for the construction of CLARIN as the national infrastructure to support the science and technology of the Portuguese language.

The preparatory work for the infrastructure was pursued within the scope of the European project CLARIN (2008-2011), of which I was the coordinator of the Portuguese network, which was preceded by the European project LTRC (2002-2004).

Part of this work is reported in the following publication:

Regarding the development of the resources mentioned above, besides the projects already indicated in the previous chapters, that work was also undertaken as part of the FCT funded project NEXING (2001-2003).
8 Innovative solutions

The previous chapter reported on how the results and resources presented in PART II of this lecture were built on to support the research sector and the leveraging of further research and development activities by means of their reuse.

In the present chapter, the reporting on the exploitation of these results proceeds by focusing now on their innovative reuse, combination and application in the business and public sectors, along these next two sections:

8.1 Business sector
8.2 Public sector

8.1 Business sector

In a near future, modern societies are expected to enter an age of so called demographic depression when most of their population will be senior citizens. This represents a societal challenge at different levels, including in terms of the access and usage of information and communication technologies.

Like other societal challenges, this trend is bringing both new problems and opportunities and in order to address them, a consortium led by Microsoft was gathered under the project S4S (2013-2015).

This was an innovation project funded by the National Reference Strategic Framework (QREN-Quadro de Referência Estatégico Nacional) and undertaken by a consortium formed by seven partners, which included four companies, one technology transfer organization and three research groups. Among the companies, besides Microsoft, there was Optimus, a major national telecom operator. Fraunhofer was the transfer organization. And the NLX-Group I coordinate was one of the three research groups contributing to the project.

Question answering

A major goal of the project was to address the cognitive and other eventual specific constraints involved in the utilization of smartphones by senior users.

Web search with common search engines, which upon receiving input keywords deliver a list of pointers to documents where further "manual" peruse is needed in order to look for the necessary information, offers itself to convenient improvement in this usage context and with these target users. Alongside this search option, question answering engines offer the alternative to have a shortcut through the "manual" peruse on the retrieved documents phase and to deliver a direct answer to a question entered in full natural language.
As reported in PART II above, we developed ground technology for real-time open-domain question answering on the web of Portuguese documents in our \textbf{NLX-Group} in a previous FCT research project, \textbf{QueXting} (2005-2007), under my coordination.

The \textbf{S4S} project offered the chance to build on that technology and to apply it in a real usage scenario. This innovation project offered also the chance to entertain closer cooperation with a second research group contributing to the project, the Human-Computer Interaction and Multimedia Research Team (HCIM) of the Large-Scale Informatics Systems Laboratory (LASIGE), in my department, the Department of Informatics of Faculty of Sciences of the University of Lisbon (DIFCUL), namely with \textbf{Luís Carriço} and his doctoral student \textbf{Luís Duarte}.

The immediate outcome of this collaboration and the contribution to \textbf{S4S} project was the prototype of an app for Windows based smartphones that permitted question answering in the Portuguese language over the web (the interface of this app is shown in the picture below). The undertaking of the \textbf{S4S} project by Microsoft was confluent with the company activities aimed at fostering the development of the personal assistant Cortana.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{mobile_interface.png}
\caption{The interface of the mobile version of XisQuê question answering systems prepared for the \textbf{S4S} project.}
\end{figure}

\begin{itemize}
\item
\end{itemize}

In the current market of technological devices and advanced remote services, high-quality support that helps end-users to resolve difficulties by themselves in using such products represents a decisive factor for economic agents in the industry.

This business opportunity has been commercially explored in the Portuguese market by Higher Functions Lda. (HF), a SME with an online service for troubleshooting via a chat channel. This helpdesk service supports customers in self-repairing their equipments, devices and services, in the area of Information
Technology, that were purchased to some third party that contracted HF to ensure that helpdesk service.

**Machine translation**

As the company growth is pushing it to expand its business to other international markets, a key challenge to be faced is the barrier represented by multilingualism. In the current helpdesk, the gains in terms of efficiency are obtained by resorting to a database where all previous questions entered by costumers and respective answers are stored so that a new incoming question, if matching a previously stored question, is automatically answered with the respective previously stored answer, thus avoiding resorting to a human operator, which is a more costly mode of operation. As the answers are entered and stored in the users’ native language, moving to other markets with different natural languages implies to start with an empty database an in a greatly diminished competitive vantage point.

In order to address this challenge posed by the internationalization of this type of business, HF is teaming up with **NLX-Group** and is leading the innovation project **ASSET** (2016-2019). This is a project funded by the Portugal 2020 program (a national initiative fostering innovation for the period from 2014 to 2020) and is supporting the on going expansion of the helpdesk to new markets.

As reported in PART II above, we had developed ground technology for machine translation in a previous European research project, **QTLeap**, that I coordinated and in which HF also participated.

This **ASSET** project offers the chance to build on this machine translation technology and exploit it in a real usage scenario. The HF database of questions and answers is stored in a natural language that acts as a pivot language. When a new question is entered to the helpdesk service, in a different language, it is machine translated into that pivot language and the result is used to undertake the search for a matching with a possibly previously stored question. If that matching succeeds, the retrieved answer is machine translated back from the pivot language into the user's language.

This permits not only that the technology and the business be extended to other international markets, but also that the database can be continuously fed and improved with all users’ interactions no matter what are their languages.

This project offers also the chance to continue the cooperation with members of the Human-Computer Interaction and Multimedia Research Team (HCIM) of my department, headed by **Luís Carriço**.

### 8.2 Public sector

Camões - Instituto da Cooperação e da Língua I. P. (Camões) is part of the Portuguese Ministry of Foreign Affairs and is the public body responsible for implementing the national policy concerning the Portuguese language. One of its
activities is to ensure the teaching of Portuguese as a second, or foreign, language across the world at the different delegations of Camões or at its local partners.

Among the tasks related to the teaching of Portuguese as a foreign language, Camões is responsible to run certification exams. These exams are the instrument to assign official diplomas certifying that the successful exam takers reached a certain level of proficiency in their mastering of the Portuguese language. These exams are run across the world and adhere to the Common European Framework of Reference of Languages, thus being structured in accordance to the cross language levels of proficiency A1, A2, B1, B2, C1 and C2.

In order to design one of these exam sheets with its questions, it is necessary to resort to excerpts of text to be exhibited to exam takers and around which questions, completion tasks, etc. in the exam sheet can be developed and pedagogically anchored. This apparently simple endeavour turns out to be a significant practical challenge given the combination of constraints imposed by the context. Excerpts have to be natural occurring ones, i.e. they cannot be written by the writer of the exam, or someone else, for that specific purpose but need to be a quotation of some pre-existing and published text. Excerpts must be used only once and cannot be reused in different exams in order to ensure fairness across exam takers at different occasions. Excerpts need to be obtained in a considerable large number given the high number of exam takers, of exam locations, and of different exam dates along the year, etc. And more critically, an excerpt should correspond to the proficiency level that is being certified by the exam sheet where it is integrated.

Language analytics

While the first constraints above are relevant given the number of exam sheets necessary per academic year and thus add to the overall volume of work that is necessary, the latter brings a further, somehow subtle and yet more stringent, difficulty that cannot be addressed by brute force, that is by just adding more work force to the team responsible to come out with the exam sheets. A given excerpt that is intended to correspond to a certain level of proficiency has many different linguistic dimensions and desirably each one of them should be at its own difficulty score that is the most appropriate to that level of proficiency. As reported by exam writers, this calibration exercise is a most difficult endeavour to ensure in view of keeping a consistent classification and judgements across different exam writers and even across different occasions for the same exam writer.

Against this background, I was addressed by Camões in their search for help with some possible contribution for this problem by some innovative application of language technology. As reported in PART II above, we had already developed ground language technology for eLearning in a previous European research project, LT4eL (2005-2008), in our NLX-Group.

In order to address this challenge, together with the Camões’ teaching services, the project CEFR was designed and commissioned to the NLX-Group, and was run until its successful completion, from January 2014 to June 2015, under my coordination, and with the contribution of João Rodrigues.
The outcome of this project is an online service, LX-CEFR, that is now freely provided by Camões. This service takes an excerpt of Portuguese text as input and delivers a number of relevant statistics concerning the linguistic make up of that excerpt (the interface of this service is depicted in the figures below14). Concomitantly, on the basis of those analytic elements, the service delivers the scores of the input excerpt in a radar diagram of increasing levels of language proficiency (from A1 to C2) along a number of linguistic dimensions that are most relevant for its classification. On the basis of this information, a decision on the proficiency level a given excerpt belongs to can be made with higher certainty and much improved consistency.

To the best of our knowledge, this is the first service of this kind set up for any language.

![Image of the interface of the LX-CEFR online service at the Camões website.](http://cvc.instituto-camoes.pt/tecnologias-da-lingua/classificador-lx-cefr.html (consulted on 15 March 2016)).

When they were inquired a few months after the service came into production, Camões’ services indicated that given the high level of satisfaction of the professionals in their team of exam writers with this new service, these

professionals have now fully integrated this application in their workflow and are not designing new exam sheets any longer without resorting to its support.

Interestingly, as this text classification service was freely made available online, its usefulness evolved beyond the initially intended usage purpose.

Besides the language professionals designing exam sheets, students of Portuguese as a foreign language across the world are reportedly using this service as a new pedagogical tool that help them to check the language level of texts of their interest and thus to support their autonomous language learning activities in an innovative way.

Part of this work is reported in the following publications:


* The results introduced in the present chapter, and specially this CEFR project, are most eloquent examples of the importance of fostering the science and technology of the Portuguese language and its preparation for the digital age.

The implementation of a service like LX-CEFR would never be feasible if there would not exist a wealth of previously developed and publicly available results in terms of knowledge, tools and data sets for the processing of Portuguese. These building blocks are indispensable to foster innovative solutions whose progressive development, enhancement and curation requires decades of consistent efforts. They need to be ready to be used and exploited because their putative dedicated production “on demand” as a specific part of a project like CEFR (say by “adaptation” from functionally similar building blocks for other language, like English) would simply not be economically and practically viable given the prohibitive unbalance between costs and benefits for the stake holders of the project.

This CEFR project is also another eloquent example that is not necessary to have English as the object language to achieve innovative results advancing the frontier of science and technology of natural language.

*
As a concluding note of PART II and PART III of this lecture, it is worth noting that with the exception of this CEFR project, which was commissioned by Camões, all projects mentioned in this lecture, national or international, thus including the 2 European and the 8 national projects under my scientific coordination since my doctoral graduation, were approved in open and competitive calls for project proposals submission after their proposals having been evaluated and selected by independent experts.

In the last 15 years, since the creation of the NLX Group I coordinate, with the exception of the 4 Keuro corresponding to the CEFR project, the rest of the over 9,5 Meuro of external funding were thus obtained in open and competitive calls for project proposals submission.
This fourth and final part closes the lecture by considering promising inquiry lines and possible future research undertakings.

It includes the final Chapter 9 Promising research paths and future work.
9 Promising research paths and future work

Deep language processing has its stepping stone in linguistically principled approaches and, by embedding data driven solutions, has been in a steady progress and has evolved towards providing ever more ample support to natural language applications.

In terms of efficiency, it has benefited from semantic underspecification techniques and from advances in parsing technology that have improved its speed. In terms of text coverage, it has benefited from research on the automatic expansion of the lexicon. In terms of ambiguity resolution, word level processing has been outsourced to shallow pre-processing modules and language models have been coupled with grammars to rank their parses. In terms of robustness, techniques have being developed to cope with multi-word expressions and to handle out of vocabulary words. And in terms of language coverage and multilingualism, deep grammars for more languages and more monolingual and parallel deepbanks have been developed. Our results described above are in line with and represent contributions to this advancement of natural language processing, in general, and deep processing, in particular, along several of these progress paths.

The hallmark of deep language processing ultimately lies in the mapping between the linguistic form of sentences and their fully-fledged grammatical representations, which crucially include an actionable representation of their meaning that support the interface with other high level cognitive functions, such as reasoning, and with knowledge representation at large. This semantic representation is encoded in terms of some logical language whose formal properties make that meaning representation suitable to render explicit, and actionable by artificial agents, the key semantic properties of sentences. While this has allowed crucial sentential semantic relations such as entailment or synonymy, among others, to be captured, it should not go unnoticed that it has offered less traction to address sub-sentential semantic relations, and in particular lexical semantics.

In this connection, and very importantly, as deep linguistic representation is tightly anchored to the representation of meaning, the progress of deep language processing can be further supported by building on the recent and very promising advances of lexical semantic processing, which have been strengthened by enhanced distributional semantics techniques. This is the key insight that will be proactively guiding our upcoming research undertakings.

In distributional semantics, the meaning of a word (in a corpus) is taken as being captured by a real valued high dimensional vectorial representation of the contexts in which that word occurs (in that corpus), such that each component of the vector ultimately relates somehow to the weighted frequency of the co-occurrence of that word with a given context, typically consisting of some small window around that word encompassing a few words before and after it. The
cosine between the vectors of two words is typically used to base the measurement of the distance between those two representations. The interpretative leap here is to assume that a smaller distance is indicative of higher semantic relatedness, and thus the cosine between the vectors is supporting a measure of semantic distance between words.

Interestingly, these vectorial representations have shown quite straightforward yet intriguing predictive capacity (Erk, 2012; Mikolov et al., 2013). Just to give an example, among others, the vector of the word *king* subtracted by the vector of *man* and added to the vector of *woman* produces a vector very close to the vector of *queen*.

![Figure 30](image.png)

**Figure 30:** Part of Figure 4 of (LeCun et al., 2015) with a two dimension representation of distributional semantics vectors for some phrases where one can observe that semantically similar expressions are mapped to nearby representations.

Like other solutions in artificial intelligence, in general, and in natural language processing, in particular, and as has repeatedly been the case in the past with other research topics, this upsurge of interest on distributional semantics has little to do with its conceptual novelty. The key insight and justification underlying distributional semantics is anchored on the approach to modelling meaning based on semantic spaces (as opposed to alternative approaches such as semantic networks or feature-based models), which dates back to many decades ago, from the very inception of Artificial Intelligence and Formal Linguistics around the 1950’s. It was only with the latest breakthroughs in terms of the increasingly lower costs of data storage, of speed of processing, and of the availability of big data sets, that the promise of this approach could be effectively explored.

The lexical semantic representation obtained via the distributional methodology has thus been the focus of much recent interest ranging from the optimization of the inner workings of that methodology, to research on its use to support high level natural language processing applications.
There have emerged two main model families for distributional vectors (also referred to as word embedding in some literature) with different advantages and drawbacks, according to (Pennington et al., 2014) and (Levy et al., 2015). On the one hand, the so called "count"-based or global matrix factorization models, as advocated by Deerwester et al. (1990), among others, and on the other hand, the "prediction"-based or local context window methods, as assumed by Mikolov and colleagues (2013c).

In recent years, tools have been made freely available that support the training of vectorial models, such as Glove tool from the Stanford team (Pennington et al., 2014), and the word2vec tool from the Google team (Mikolov et al., 2013).

A competition orthogonal to this difference in the formal underpinnings of distributional models concerns the difference in the computational procedures used to train them. Interestingly, in this case, systematic comparative assessment (Levy et al., 2015) has shown that there is no substantial difference in performance between neural-network and non neural-network based vectorial models, despite all the recent media spin around so-called deep learning (LeCun et al., 2015).

While the (lower) distance between vectors has been attributed to (higher) semantic relatedness, it is worth noting that it is not entirely clear what principled semantic relations (e.g. synonymy, hypernymy, meronymy, etc.) are actually covered by this representation and can be disentangled from them, which is particularly acute for the case of the semantic relation of antonymy (Chen et al., 2015).

The eventual correspondence, if any, between vectorial meaning representations and the representation of meaning through semantic networks, including the highly popular Wordnet lexical ontology, has also attracted an increasing interest (Agirre et al., 2009). This interest eventually spreads to the more fine grained question of whether there can be training of distributional models whereby a vector could capture an individual word sense rather than a cluster of word senses associated to a lexically ambiguous word (Li and Jurafsky, 2015).

Further issues related to distributional semantics have also received special attention. For instance, it is important to understand if and how vectorial representations of subexpressions can be compositionally combined into a vectorial representation of their larger expressions (Bride et al., 2015), eventually guiding this composition along the internal linguistic structure of the relevant expressions (Tai et al., 2015).

Ultimately, it is of utmost relevance to discover whether and how one can obtain some measure of semantic relatedness among sentences and even among several sentence texts on the basis of the combination of more atomic vectorial representations (Le and Mikolov, 2014; Agirre et al., 2015).

Interestingly, and even without all of the more foundational issues touched upon above having been fully resolved, the exploration of distributional semantics has delivered very promising results in a number of different natural language processing tasks and applications, including POS tagging, named entity recognition or sentiment analysis (Li and Jurafsky, 2015), parsing (Socher et al., 2013), or discourse analysis (Ji and Eisenstein, 2014), among others. Recently, even if with
modest first results, the exploration of distributional semantics was used for very high level tasks such as machine translation (Sutskever et al., 2014; Bahdanau et al., 2014; Jean et al., 2014) or reasoning (Bowman et al., 2015).

We think that there are very cogent reasons to believe that this line of progress concerning distributional semantics will be maturing to a level such that its hybridization with previously seasoned methods and solutions will drive a game-changing approach to deep language processing, in particular, and eventually to natural language processing, in general.

In our view, this is the seed for the next epistemological shift in the science and technology of natural language. This is the strategic insight driving our research goals concerning deep language processing for the near future.
References

For the sake of readers' convenience, references in this lecture to the publications concerning activities that have been undertaken by myself or under my supervision are prefixed with "*".


* Branco, António, 2009, "Research and Development on Natural Language Technology at the University of Lisbon, Department of Informatics: The NLX Group", In Proceedings, I Iberian SLTech - I Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages, Porto Salvo, Portugal, 3-4 September 2009, pp.121-122.


Computacionais da Teoria da Gramática, Campinas, Brazil, Mercado de Letras, Chapter 8, pp.269-301.


* Costa, Francisco and António Branco, 2013, "Temporal Relation Classification Based on Temporal Reasoning", In Proceedings, 10th International Conference on Computational Semantics (IWCS), Potsdam, Germany, 19-20 March 2013, pp.54-64.

* Costa, Francisco and António Branco, 2014, A Computational Grammar for Deep Linguistic Processing of Portuguese: LXGram, version 5, University of Lisbon, Faculty of Sciences, Department of Informatics.


* Gaudio, Rosa Del, 2013, Automatic Extraction of Definitions, Doctoral dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.


* Gonçalves, Patricia, 2015, *Open-Domain Web-based Multiple Document Question Answering for List Questions with Support for Temporal Restrictors*, Doctoral Dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.


* Gonçalves, Patricia and António Branco, 2014, "Open-Domain Web-Based List Question Answering with LX-ListQuestion", In Proceedings, ACM 4th International Conference on Web Intelligence, Mining and Semantics


Leibniz, Gottfried, 1666, Dissertatio de Arte Combinatoria, In Sämtliche Schriften und Briefe, Berlin: Akademie Verlag, 1923, A VI 1, p. 163.


* Nunes, Filipe, 2007, Verbal Lemmatization and Featurization of Portuguese with Ambiguity Resolution in Context, Masters dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.


* Reis, Rúben, 2010, Marcação Semântica de Páginas Web apoiada por Parsers de Dependências Gramaticais, Masters dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.


* Rodrigues, João, 2015, Speech-to-Speech Translation to Support Medical Interviews, Masters dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.


* Rodrigues, João, Luís Gomes, Steven Neale, Andreia Querido, Nuno Rendeiro, Sanja Štajner, João Silva and António Branco, accepted, "Domain-Specific


* Silva, João and António Branco, in preparation, "Deep Language Processing Out-of-Vocabulary Words ".


* Silveira, Sara, 2015, *Enhancing Extractive Summarization with Automatic Post-processing*, Doctoral dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.


Acknowledgments

Many of the results and publications mentioned above were the result of joint work with post-graduation students of mine, whose research was undertaken under my supervision. The first acknowledgment in this section, and a very special word of thanks, go to the them: at the postdoc level, Francisco Costa, João Ricardo Silva, Sanja Štajner and Steven Neale; at the doctoral level, Francisco Costa, João Ricardo Silva, João Rodrigues, Patricia Gonçalves, Rosa Del Gaudio and Sara Silveira; at the masters level: Francisco Costa, João Rodrigues, João Ricardo Silva, Pedro Martins, Rúben Reis, Sérgio Castro and Filipe Nunes.

For the complete list of my postgraduation students, the interested reader should consult my curriculum vitae in the set of the documents supporting the present application.

Some of the activities and results presented above in the lecture were undertaken also in collaboration with my colleagues from my Department of Informatics of the Faculty of Sciences of the University of Lisbon (DIFCUL) João Balsa, Luís Carriço and Luís Duarte. They were undertaken also with Adam Przepiórkowski, Aljosa Burchardt, Amália Mendes, Andreia Querido, Andrejs Vasiljevs, Arantxa Otegi, Asunción Moreno, Audronė Bielevičienė, Bernardo Magnini, Bolette Sandförd Pedersen, Carmen García-Mateo, Catarina Carvalheiro, Clara Pinto, , Cláudia Martins, Cvetana Krstev, Dan Flickinger, Dan Tufiş, Eduardo Ferreira, Eiríkur Rögnvaldsson, Eneko Agirre, Florbela Barreto, Gerhard Budin, Gustavo Batista, Hans Uszkoreit, Hugo Meinedo, Inguna Skadiņa, Inma Hernáez, Isabel Trancoso, Jan Hajič, Jan Odijk, Joana Ramos, João Graça, John Judge, John McNaught, Jolanta Zabarkskaite, Josef Van Genabith, Joseph Mariani, Kadri Vider, Kiril Simov, Koenraad De Smedt, Krister Lindén, Lars Borin, Luís Gomes, Maciej Ogrodniczuk, Maite Melero, Maria Fernanda Nascimento, Mariana Avelãs, Marisa Campos, Marko Grobelnik, Marko Tadić, Martin Popel, Mike Rosner, Monica Monachini, Nicoletta Calzolari, Nora Aranberri, Nuno Rendeiro, Núria Bel, Oier Lopez de Lacalle, Paul Thompson, Paulo Henriques, Paulo Quaresmas, Petya Osenov, Piotr Pežik, Radovan Garabík, Rehm, Georg, Rita Carvalho, Rita Pereira, Rita Santos, Rui Vaz, Silvia Pereira, Simon Krek, Sophia Ananiadou, Stelios Piperidis, Svetla Koeva, Tamás Váradi, Thomas Pellegrini, Tiago Henriques, Valia Kordoni, Vera Lúcia Strube de Lima, Walter Daelemans and Yi Zhang.

For the complete list of my co-authors, the interested readers should consult my curriculum vitae in the set of the documents supporting the present application.

The activities and results reported above were partly supported by the Portuguese Republic through FCT-Fundação para a Ciência e Tecnologia, ADI-Agência de Inovação and Camões-Instituto da Cooperação e da Língua I. P. under the contracts: ASSET, ADI:P2020/3279/2016; CEFR, Camões:2014; DP4LT, FCT:PTDC/EEI-SII/1940/2012; GramaXing, FCT:PLUS/PLP/50301/2003; NEXING, FCT:PLP/34076/2001; QueXting, FCT:PLUS/PLP/61490/2004; S4S,

Annex A — Contributions: publications list

For the sake of readers’ convenience, this annex gathers in a single consolidated list the publications that were referred to in the lecture concerning activities that have been undertaken by myself or under my supervision.

For the complete list of this kind, the interested reader is referred to my curriculum vitae in the set of the documents supporting the present application.


Branco, António, 1999, Reference Processing and its Universal Constraints, Doctoral Dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.


Branco, António, 2009, "Research and Development on Natural Language Technology at the University of Lisbon, Department of Informatics: The NLX Group", In Proceedings, I Iberian SLTech - I Joint SIG-IL/Microsoft Workshop
on Speech and Language Technologies for Iberian Languages, Porto Salvo, Portugal, 3-4 September 2009, pp.121-122.


Costa, Francisco and António Branco, 2013, "Temporal Relation Classification Based on Temporal Reasoning", In Proceedings, 10th International Conference on Computational Semantics (IWCS), Potsdam, Germany, 19-20 March 2013, pp.54-64.


Frege, Gotlob, 1879, Begriffsschrift, eine der Arithmetischen Nachgebildete Formelsprache des Reinen Denkens, Halle.


Gaudio, Rosa Del, 2013, Automatic Extraction of Definitions, Doctoral dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.


Gonçalves, Patricia, 2015, Open-Domain Web-based Multiple Document Question Answering for List Questions with Support for Temporal Restrictors, Doctoral
Dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.


Nunes, Filipe, 2007, Verbal Lemmatization and Featurization of Portuguese with Ambiguity Resolution in Context, Masters dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.


Reis, Rúben, 2010, Marcação Semântica de Páginas Web apoiada por Parsers de Dependências Gramaticais, Masters dissertation, University of Lisbon, Faculty of Informatics.

Rodrigues, João, 2015, Speech-to-Speech Translation to Support Medical Interviews, Masters dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.


Silva, João, 2014, Robust Handling of Out-of-Vocabulary Words in Deep Language Processing, Doctoral dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.


Silva, João and António Branco, in preparation, "Deep Language Processing Out-of-Vocabulary Words".


Silveira, Sara, 2015, Enhancing Extractive Summarization with Automatic Post-processing, Doctoral dissertation, University of Lisbon, Faculty of Sciences, Department of Informatics.


Session, 10th International Conference on Computational Processing of Portuguese (PROPOR2012), Coimbra, Portugal, 17-20 April 2012.


Annex B — Contributions: data sets list

For the sake of readers' convenience, this annex gathers a single consolidated and lexicographically sorted list with extended descriptions of the data sets that were referred to in the lecture and that concern research that has been undertaken by myself or under my supervision.

CINTIL-DeepBank

Set of text materials to support the evaluation and training of tools for the processing of Portuguese, including language models for deep linguistic processing grammars. This corpus contains around 10 000 sentences (approximately 100 000 words) manually annotated by experts in natural language science and technology. Each sentence is associated to exhaustive characterization of its grammatical features in lexical, morphological, syntactic and semantic terms. It is distributed through META-SHARE http://metashare.metanet4u.eu.

CINTIL-Definitions

Set of text materials to support the evaluation and training of tools for the processing of Portuguese, with special relevance for definition extractors. This corpus contains over 250 000 words manually annotated by experts in natural language science and technology. Each definition in the corpus is delimited and annotated concerning its components. Each word is associated to its linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions. It is distributed through META-SHARE http://metashare.metanet4u.eu.

CINTIL-DependencyBank

Set of text materials to support the evaluation and training of tools for the processing of Portuguese, including grammatical dependencies parsers. This corpus contains around 10 000 sentences (approximately 100 000 words) manually annotated by experts in natural language science and technology. Each sentence is associated to the graph that represents the grammatical functions
holding between its words. Each word is associated to linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions. It is distributed through ELDA-Evaluations and Language Resources Distribution Agency and META-SHARE http://metashare.metanet4u.eu.

CINTIL-DependencyBank PREMIUM

Set of text materials similar in design to the previous one and differing from it in the sentences that were treebanked and in the circumstance that the support tool to draw the grammatical dependency graphs is not the LXGram but the full text coverage LX-DependencyParser. It contains 3 000 sentences (approximately 79 000 words). It is distributed through META-SHARE http://metashare.metanet4u.eu.

CINTIL-International Corpus of Portuguese

Set of text materials to support the evaluation and training of tools for the processing of Portuguese, including morphological analyzers, POS taggers and named entity recognizers. This corpus contains 1 million words manually annotated by experts in natural language science and technology. Each word is associated to linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions. It was developed in cooperation with CLUL-Center of Linguistics of the University of Lisbon. It is distributed through ELDA-Evaluations and Language Resources Distribution Agency and META-SHARE http://metashare.metanet4u.eu.

CINTIL-LogicalFormBank

Set of text materials to support the evaluation and training of tools for the semantic processing of Portuguese. This corpus contains around 10 000 sentences (approximately 100 000 words) manually annotated by experts in natural language science and technology. Each sentence is associated to the logical form that represents its meaning in a logical language for semantic description. It is distributed through META-SHARE http://metashare.metanet4u.eu.

CINTIL-NamedEntities

Set of text materials to support the evaluation and training of named entity disambiguators. This corpus contains around 30 000 sentences with 26 000 named
entities that are manually annotated by experts in natural language science and technology with identifiers of the corresponding entities in the DBpedia ontology. Additionally, each word is associated to the linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions. It is distributed through META-SHARE http://metashare.metanet4u.eu.

**CINTIL-PropBank**

Set of text materials to support the evaluation and training of tools for the processing of Portuguese, including semantic role labellers. This corpus contains around 10 000 sentences (approximately 100 000 words) manually annotated by experts in natural language science and technology. The syntactic constituents of sentences are associated to linguistic information about its semantic role. Each word is associated to linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions. It is distributed through META-SHARE http://metashare.metanet4u.eu.

**CINTIL-QATreebank**

Set of text materials to support the evaluation and training of processing tools relevant for automatic question answering. This treebank includes 111 imperative and interrogative sentences associated to their syntactic constituency trees obtained through the transformation of their declarative counterparts and their trees by experts in natural language science and technology. It is distributed through META-SHARE http://metashare.metanet4u.eu.

**CINTIL-TreeBank**

Set of text materials to support the evaluation and training of tools for the processing of Portuguese, including constituency parsers. This treebank contains around 10 000 sentences (100 000 words) manually annotated by experts in natural language science and technology. Each sentence is associated to linguistic information about its syntactic constituency tree tagged with phrase categories. Each word is associated to linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions. It is distributed through ELDA-Evaluations and Language Resources Distribution Agency and META-SHARE http://metashare.metanet4u.eu.
CINTIL-WordSenses

Set of text materials to support the evaluation and training of word sense disambiguators. This corpus contains around 24 000 sentences with 45 000 words that are manually annotated by experts in natural language science and technology with the identifiers of concepts (synsets) that they convey in terms of the lexical semantic network MWN.PT. Additionally, each word is associated to the linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions. It is distributed through META-SHARE http://metashare.metanet4u.eu.

DeepBankPT

Set of text materials translated into Portuguese from the Penn Treebank, to support the evaluation and training of tools for the processing of Portuguese, including language models for deep linguistic processing grammars. This corpus contains around 3 500 sentences (approximately 45 000 words) manually annotated by experts in natural language science and technology. Each sentence is associated to exhaustive characterization of its grammatical features in lexical, morphological, syntactic and semantic terms. It is distributed through META-SHARE http://metashare.metanet4u.eu.

DependencyBankPT

Set of text materials translated into Portuguese from the Penn Treebank to support the evaluation and training of tools for the processing of Portuguese, including grammatical dependencies parsers. This treebank contains around 3 500 sentences (approximately 45 000 words) manually annotated by experts in natural language science and technology. Each sentence is associated to the graph that represents the grammatical functions holding between its words. Each word is associated to linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions. It is distributed through META-SHARE http://metashare.metanet4u.eu.

LogicalFormBankPT

Set of text materials translated into Portuguese from the Penn Treebank, to support the evaluation and training of tools for the semantic processing of Portuguese. This corpus contains around 3 500 sentences (ca. 45 000 words) manually annotated by experts in natural language science and technology. Each
sentence is associated to the logical form that represents its meaning in a logical language for semantic description. It is distributed through META-SHARE http://metashare.metanet4u.eu.

**LX-Abbreviations**

List of abbreviations in Portuguese with around 200 entries. It is distributed through META-SHARE http://metashare.metanet4u.eu.

**LX-NominalInflections**

Lexical lists with around 27 000 items associated to information concerning their nominal lemmatization and inflection properties. It is distributed through s of DI/FCUL.

**LX-Stopwords**

List of stop words with over 2 500 items. It contains a list of Portuguese words belonging to closed classes, grouped by POS and associated to all their admissible categories. It is distributed through META-SHARE http://metashare.metanet4u.eu.

**LX-VerbalInflections**

Lexical lists with around 800 000 items associated to information concerning their verbal lemmatization and inflection properties. It is distributed through NLX-Group of DI/FCUL.

**Nexing Corpus**

Set of texts resulting from the transcription of the audio recordings of psycholinguistic experimental sessions for eliciting data concerning human performance in the syllogistic reasoning task. It is available at http://www.di.fc.ul.pt/~ahb/nexingcorpus.htm
PropBankPT

Set of text materials translated into Portuguese from the Penn Treebank, to support the evaluation and training of tools for the processing of Portuguese, including semantic role labellers. This corpus contains around 3 500 sentences (approximately 45 000 words) manually annotated by experts in natural language science and technology. Each sentence is associated to its syntactic constituency tree decorated with semantic roles. Each word is associated to linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions. It is distributed through META-SHARE http://metashare.metanet4u.eu.

QTLeap Multilingual Parallel Corpora

Set of 4 000 question and answer pairs in the domain of computer and IT troubleshooting for both hardware and software. This textual material was collected using a commercial support service via chat, in Portuguese, and the corpus is thus composed by naturally occurring utterances produced by users while interacting with that service. Each question answer pair is translated into seven languages, other than Portuguese, namely Czech, Basque, Bulgarian, Dutch, English, German and Spanish. It is distributed through META-SHARE http://metashare.metanet4u.eu.

QTLeap WSD/NED Multilingual Corpora

Set of text materials comprising the QTLeap Multilingual Parallel Corpora and the Europarl multilingual corpora for the Czech (9.2 Million tokens), Basque (5.2 Million), Bulgarian (4.9 M), English (53 M), Portuguese (5.7 M) and Spanish (57.1 M) languages, automatically annotated at multiple semantic levels by processing tools for tokenization, lemmatization, part-of-speech tagging, named-entity recognition and classification, named-entity disambiguation, word sense disambiguation and coreference resolution. It is distributed through META-SHARE http://metashare.metanet4u.eu.

TimeBankPT

Set of text materials with around 70 000 words for the evaluation and training of temporal processing and extraction tools. These are texts translated into Portuguese from materials of the first TempEval challenge and annotated in accordance to the TimeML annotation conventions with temporal information.
concerning temporal expressions and relations. It is distributed through META-SHARE http://metashare.metanet4u.eu.

**TreeBankPT**

Set of text materials translated into Portuguese from the Penn Treebank, to support the evaluation and training of tools for the processing of Portuguese, including constituency parsers. This treebank contains around 3 500 sentences (approximately 45 000 words) manually annotated by experts in natural language science and technology. Each sentence is associated to linguistic information about its syntactic constituency tree. Each word is associated to linguistic information about nominal and verbal inflection, lemma, POS and about closed classes multi-word expressions. It is distributed through META-SHARE http://metashare.metanet4u.eu.

**MWNT.PT - Portuguese MultiWordNet**

Lexical ontology with over 17 200 concepts (synsets) manually validated and connected among each other through the relations of hipo-/hipernymy. These concepts cover more than 21 000 tokens and over 16 000 types of the European and American variants of Portuguese, being aligned with equivalent concepts of the Princeton wordnet for English, and transitively with the concepts of the MultiWordnets of Italian, Spanish, Hebrew, Romanian and Latin. It is distributed through ELDA-Evaluations and Language Resources Distribution Agency and META-SHARE http://metashare.metanet4u.eu.
Annex C — Contributions: applications, processing tools and online services list

For the sake of readers' convenience, this annex gathers in a single consolidated and lexicographically sorted list the applications, the processing tools, the language research support tools, and the respective online services, referred to in this lecture and that concern research that has been undertaken by myself or under my supervision.

CINTIL-Concordancer

This is an online system to search language patterns in the CINTIL corpus, available at http://cintil.ul.pt:
CINTIL-Treebank Searcher

This is an online system to search for language patterns in the collections of syntactic trees of the CINTIL-Treebank, with an online service at http://cintilsearcher.di.fc.ul.pt:

LXGram

This is a deep language processing grammar for the Portuguese grammar.

LX-CEFR

This is an application for text analysis and classification according to the proficiency levels of the Common European Framework of Reference for Languages. For its online service and further details and pointers, go to http://lxcefr.di.fc.ul.pt:
LX-Centre

This is the entry point to the collection of linguistic resources, services and tools developed in all or part by the NLX-Natural Language and Speech Group at the University of Lisbon, Department of Informatics. These resources are freely available online on a best-effort basis to help foster the education, research and development in the realm of natural language science and technology. They are meant also to showcase some of our research results in natural language processing.

LX-Chunker

This is an identifier of paragraphs and sentences for Portuguese. It seeks to cope with the ambiguity and ambivalence of symbols that in some occurrences are indicators of separations among sentences and in other contexts are not. It is a hybrid tool, based on regular expressions and hidden Markov models, with an f-score of 99.9%.

For its online service and further details and pointers, go to http://lxsuite.di.fc.ul.pt (picture of its interface can be found above in the section of the LX-Tagger).
LX-Conjugator

This is a verbal conjugator for the Portuguese language. It takes a Portuguese infinitive verb form as input and delivers the corresponding conjugated forms. It is the only available tool for fully-fledged Portuguese verb conjugation, including the full range of pronominal conjugation forms. Its capacity includes the handling of pronominal conjugation, compound tenses, double forms of past participles, past participle forms inflected for number and gender, negative imperative forms, and courtesy forms for second person. Given that it is based in principled linguistic generalizations captured by regular expressions and the appropriate lexica of affixes, it is the only available conjugator to handle neologisms. Their eventual faults have been correct along the time as it has been put to use, and at present no defect is known.

For its online service and further details and pointers, go to http://lxconjugator.di.fc.ul.pt:
**LX-Definitions**

This is an application for definitions extraction, that scores in the range of 95-99% in terms of AUC, and in the range of 90–99% in terms of F-measure.

For its online service and further details and pointers, go to http://lxdefinitions.di.fc.ul.pt:

**LX-DepParser**

This is a parser of grammatical dependency relations for sentences of Portuguese that for each input sentence delivers a graph connecting its words and whose directed arcs represent grammatical dependencies and the labels at the said arcs represent the grammatical function of those dependencies. The evaluation of its performance obtained 91.2% in terms of labelled attachment score (LAS).

For its online service and further details and pointers, go to http://lxdepparser.di.fc.ul.pt:
LX-Inflector

This is a language processing tool for nominal lemmatization and inflection, taking a Portuguese word form that follows the nominal inflection paradigm and an inflection feature bundle, and delivers both the corresponding lemma and the indication of its feature bundle, and the resulting form that conveys the feature bundle entered. It is based on principled linguistic generalizations captured by regular expressions and the appropriate lexica of affixes, thus handling neologisms. The lemmatization function has 97.7% f-score.

For its online service and further details and pointers, go to http://lxinflector.di.fc.ul.pt:
LX-ListQuestions

This is an application for question answering to open domain list questions on the basis of the web of online Portuguese documents.

For its online service and further details and pointers, go to http://lxlistquestion.di.fc.ul.pt:
LX-Lemmatizer

This is a verbal lemmatizer that takes a Portuguese verb form as input and delivers a ranked list of the corresponding lemmata (infinitive forms) together with inflectional feature values. Its performance was evaluated as delivering 96.5% accuracy.

For its online service and further details and pointers, go to http://lxlemmatizer.di.fc.ul.pt:

![LX-CENTER](Image)

LX-NED

This is a named entity disambiguator that annotates the occurrence of an input expression with the Wikipedia entry it refers to in its context, with an f-score of 67.0%.
LX-NER

This is an identifier and classifier of named entity expressions for the Portuguese language. Its number-based part evaluates to an f-score of 85.6%, and the name-based to 85.7%.

For its online service and further details and pointers, go to http://lxner.di.fc.ul.pt:
LX-Parser

This is a stochastic parser that performs the syntactic analysis of Portuguese sentences in terms of their constituency structure. It achieves an f-score of 88% under the Parseval metric.

For its online service and further details and pointers, go to http://lxparser.di.fc.ul.pt:

LX-SenseAnnotator

This is an application to support the validation of the annotation of each word in corpora with the identifier of the concept from the Portuguese wordnet MWN.PT that it happens to convey in each occurrence.
LX-SRLabeler

This is a semantic role labeller for sentences of Portuguese that undertakes the constituency analysis of Portuguese sentences and the assignment of semantic role labels to their relevant constituents. It achieves an f-score of 82% in terms of Parseval metrics.

For its online service and further details and pointers, go to http://lxsrlabeler.di.fc.ul.pt:
LX-Syllabifier

This is a syllabifier of words for Portuguese, with 99.7% accuracy.

For its online service and further details and pointers, go to http://lxsyllabifier.di.fc.ul.pt:
LX-Tagger

This is a part-of-speech tagger with disambiguation and full coverage for the Portuguese language. For each word occurring in a text and from the possible different morpho-syntactic categories that word may have in the lexicon, it assigns a single tag to it that indicates the morpho-syntactic category that it bears in that occurrence in the text. It scores 96.8% accuracy.

For its online service and further details and pointers, go to http://lxsuite.di.fc.ul.pt:
LX-TimeAnalyzer

This is a temporal information extractor from Portuguese texts. Given an input text, it finds the following elements: Temporal expressions, which are expressions that occur in the input text and that refer to dates and times; Events terms, which are words that refer to events that happen or hold at some point in time; Temporal relations between these times and events, i.e. the temporal ordering among these entities (before, after, overlap), according to the input text. Its performance evaluates at the state of the art level along its many relevant dimensions.

For its online service and further details and pointers, go to http://lxtimeanalyzer.di.fc.ul.pt:
LX-Tokenizer

This is an identifier of the boundaries of relevant word-level tokens in Portuguese text. It seeks to cope with the ambiguity of strings that in some contexts are single-word tokens and in some other contexts are contractions, i.e. double-word tokens. It achieves an f-score of 99.7%.

For its online service and further details and pointers, go to http://lxsuite.di.fc.ul.pt (picture of its interface can be found above in the section of the LX-Tagger).

LX-Translator

This is an application for machine translation between Portuguese and English with written and spoken input. Its evaluation scored at the state of the art level, with 0.28 and 0.27 BLEU points, respectively, for the Portuguese->English and English->Portuguese directions.

For its online service and further details and pointers, go to http://lxtranslator.di.fc.ul.pt:
LX-WSD

This is a word sense disambiguator that annotates the occurrence of an input word with the MWN.PT wordnet concept it expresses in its context, with an f-score of 65.0%.

MWN.PT Searcher

This is a system to search in the lexical semantic network MWN.PT.

For its online service and further details and pointers, go to http://mwnpt.di.fc.ul.pt:
SIMBA

This is an application for multi-document summarization, with a performance scoring 0.4698 in terms of ROUGE-L metric.

For its online service and further details and pointers, go to http://lxsimba.di.fc.ul.pt:

XisQuê

This is an application for question answering to factoid, open domain questions on the basis of the web of online Portuguese documents. The system’s performance was assessed over a test-set of 60 questions — 15 for each type of question handled by the system — randomly selected from cards of the Trivial Pursuit game. The following results were obtained with the March 2008 version of the system. Recall: A short-answer was returned for 57% of the questions. Some answer (short- or long-) was returned for 98% of the questions; Precision: Considering the 5 tentative answers provided for each question, a correct answer (short or long) was found to 98% of the questions in the test-set. For 55% of the questions, a correct short-answer was found.
For its online service and further details and pointers, go to http://xisque.di.fc.ul.pt:
Annex D — Contributions: projects list

For the sake of readers’ convenience, this annex gathers in a single consolidated and lexicographically sorted list the projects referred to in the lecture and that supported the results presented.

For a complete list of my projects, the interested reader is referred to my curriculum vitae in the set of the documents supporting the present application.

ASSET—Intelligent Assistance for Everyone Everywhere

Funding: Portugal 2020 Program
Line: Co-promotion projects, announcement 08/SI/2015
Grant: 3279
Participants: University of Lisbon (DIFCUL) and Higher Functions Lda.
Goal: To lower the barrier to enter the market and improve the conditions of success for new business projects that are based on knowledge and technology intensive activities and hence that crucially need to offer support to their clients in the resolution, by the latter, of difficulties with the installation, operation and exploitation of remote devices and services. The goal is thus to lower this kind of access barrier by making available this kind of support at ever lower costs and in more agile ways, making it affordable even for small sized companies.
Hosting institution: DIFCUL
Period: April 2016 to March 2019

CEFR—Text Categorization for Language Learning Assessment

Funding: Camões-Instituto da Cooperação e da Língua I. P.
Grant: project commissioned
Participants: University of Lisbon (DIFCUL)
Goal: To develop an online service that categorizes input Portuguese excerpts according to the proficiency levels of the Common European Framework of Reference for Languages: Learning, Teaching, Assessment.
Hosting institution: DIFCUL
Period: January 2014 to June 2015

CLARIN–Common Language Resources and Technology Infrastructure

Funding: European Commission
Line: Research Infrastructures
Grant: EC-RI-080718 (FP7-2008)
Principal investigator: Steven Krauwer, Utrecht University (The Netherlands)
Participants: University of Lisbon and over 30 other partners from European Union countries and associated countries.
Goals: Design study and preparatory project for the construction of a pan-European research infrastructure for promoting and supporting the use of language resources and tools.
Hosting institution: DIFCUL
Period: January 2008 to June 2011

DP4LT–Deep Language processing for Language Technology

Funding: FCT-Fundaçao para a Ciência and Tecnologia
Line: Electrotechnical and Informatics Engineering, Intelligent Information Systems
Grant: PTDC/EEI-SII/1940/2012
Participants: University of Lisbon (DIFCUL)
Goals: To improve the robustness of deep language processing by tackling the issue of lexical creativity with new methods. Other important goals are the construction or enhancement of datasets, tools and deep grammar for the processing of Portuguese, to be released as resources to support research and innovation.
Hosting institution: DIFCUL
Period: April 2013 to November 2015
GramaXing–Computational Grammar for the Deep Processing of Portuguese

Funding: FCT-Fundação para a Ciência and Tecnologia, Programa Lusitânia
Line: Computational grammars
Grant: PLUS/PLP/50301/2003
Participants: University of Lisbon (DIFCUL), Saarland University (Computational Linguistics Department), Universidade do Minho (Centro de Estudos Humanísticos), Universidade Federal do Rio de Janeiro
Goals: Development of a computational grammar for the deep linguistic processing of Portuguese with HPSG, integrating the DELPHIN international consortium (University of Cambridge (United Kingdom), DFKI - German Research Center for Artificial Intelligence (Germany), Laboratory for Communication Science NTT (Japan), Norwegian University for Science and Technology (Norway), Saarland University (Germany), Stanford University / CSLI (USA), Tokyo University (Japan), University of Oslo (Norway), University of Sussex (United Kingdom), University of Washington (USA))
Hosting institution: DIFCUL
Period: April 2005 to July 2007

LTRC–Language Typology Resource Center

Funding: European Commission
Line: Thematic networks
Grant: HPRI-CT-2001-40028
Principal investigator: Martin Everaert, Utrecht University (The Netherlands)
Participants: Utrecht University (The Netherlands), Max Planck Institute (The Netherlands), University of Amsterdam (The Netherlands), University of Nijmegen (The Netherlands), Leiden University (The Netherlands), Tilburg University (The Netherlands), University of Lisbon, Free University of Berlin (Germany), University of Stockholm (Sweden), University of Surrey (United Kingdom), Konstanz University (Germany), Lancaster University (United Kingdom), ZAS/Berlin (Germany)
Goals: Creation of an online archive with the typological description of languages that includes tools for research such as typological Databases, expert systems for linguistic typology, extensive scientific grammars, corpora, etc.
Hosting institution: DI/FCUL
Period: October 2002 to October 2004
LT4eL-Language Technology for eLearning

Funding: European Commission
Line: Information Society Technologies: Specific Targeted Research Project
Grant: EC-STREP-004758 (FP6-2004)
Principal investigator: Paola Monachesi, Utrecht University (The Netherlands)
Participants: University of Lisbon, Utrecht University (The Netherlands), Charles University of Prague (Czech Republic), Polish Academy (Poland), Bulgarian Academy (Bulgaria), Open University (United Kingdom), University of Iași "Alexandru Ioan Cuza" (Romania), University of Cologne (Germany), Hamburg University (Germany), University of Malta (Malta), Tübingen University (Germany), University of Applied Sciences of Zurich (Switzerland).
Goals: This project uses language technology tools and web semantic techniques to improve the search for learning materials. The technology to develop in the project facilitates the customized access to knowledge in the scope of learning management systems and supports the cooperation and decentralization of content.
Hosting institution: DIFCUL
Period: December 2005 to May 2008

METANET4U–Enhancing the European Linguistic Infrastructure

Funding: European Commission
Line: Information and Communication Technology Policy Support Programme (ICT PSP), Competitiveness and Innovation Framework Programme (CIP), 4th Call for project proposals, 1 June 2010, Theme 6: "Multilingual Web", Objective 6.1
Grant: #270893
Principal investigator: António Branco
Participants: Faculty of Sciences of the University of Lisbon, Instituto Superior Técnico, The University of Manchester (United Kingdom), University of Iași "Alexandru Ioan Cuza" (Romania), Romanian Academy, Institute for Artificial Intelligence (Romania), University of Malta, Polythecnic University of Catalunya (Spain), University Pompeu Fabra (Spain).
Goals: Development of an international platform for the distribution and sharing of language processing tools and resources.
Hosting institution: DIFCUL
Period: February 2011 to January 2013

NEXING–Natural Negation Modeling and Processing

Funding: FCT-Fundaçao para a Ciência and Tecnologia, POSI
Line: Computational Processing of the Portuguese Language
Grant: PLP/34076
Participants: University of Lisbon (DIFCUL), Universidade de Coimbra (Faculdade de Psicologia) and Universidade do Algarve (Faculdade de Ciências Humanas and Sociais)
Goals: To improve the automatic mapping between linguistic form and meaning representation, on the one hand, and between meaning representation and knowledge representation, on the other hand, with respect to natural languages negation.
Hosting institution: DIFCUL
Period: March 2001 to June 2003

QTLeap–Quality Translation by Deep Language Processing Approaches

Funding: European Commission
Grant: #610516
Principal investigator: António Branco
Participants: Faculty of Sciences da University of Lisbon, DFKI – German Research Centre for Artificial Intelligence (Germany), Charles University in Prague (Czech Republic), Bulgarian Academy of Sciences (Bulgaria), Humboldt University Berlin (Germany), University of Basque Country (Spain), University of Groningen (The Netherlands), Higher Functions, Lda (Portugal).
Goals: To research on and deliver an articulated methodology for machine translation that explores deep language engineering approaches in view of breaking the way to translations of higher quality.

Hosting institution: DIFCUL
Period: November 2013 to October 2016

QueXting—Answering Questions in the Portuguese Web

Funding: FCT-Fundação para a Ciência and Tecnologia, Programa POSI
Line: Computational Processing of the Portuguese Language
Grant: PLUS/PLP/61490/2004
Participants: University of Lisbon (DIFCUL)
Goals: Make available and maintain a free online service for answering questions worded in Portuguese on the basis of the Portuguese web.
Hosting institution: DIFCUL
Period: August 2005 to December 2007

S4S—Smartphones for Seniors

Funding: ADI - Agência de Inovação
Grant: QREN
Principal investigator: Miguel Sales Dias, Microsoft
Participants: Microsoft, Devscope, Faculty of Sciences da University of Lisbon, Instituto Fraunhofer, Optimus, Universidade de Aveiro, Wit-software
Goals: Design and implementation of applications and interaction environment targeting senior users
Hosting institution: DIFCUL
Period: October 2011 to October 2013
SemanticShare–Ferramentas and Recursos para o Processamento Semântico

Funding: FCT-Fundação para a Ciência and Tecnologia
Line: Computational Processing of the Portuguese Language
Grant: PTDC/PLP/81157/2006
Participants: University of Lisbon (DIFCUL) and Pontifícia Universidade Católica do Rio Grande do Sul, Brazil
Goals: Development of corpora annotated with linguistic information that includes syntactic constituency and dependency representations, semantic roles, and logical form semantic representations, to support the development and evaluation of syntactic-semantic parsers for the Portuguese language, with ambiguity resolution in context
Hosting institution: DIFCUL
Period: February 2008 to January 2010

TagShare–Tools and Resources for Shallow Morpho-syntactic Tagging and Processing

Funding: FCT-Fundação para a Ciência and Tecnologia, POSI
Line: Computational Processing of the Portuguese Language
Grant: PLP/47058
Participants: University of Lisbon (DIFCUL) and Center of Linguistics of the University of Lisbon
Goals: To develop a set of tools and resources for the shallow processing of the Portuguese language: large-scale POS tagged corpus; POS taggers; lemmatizers; named-entity recognizers
Hosting institution: DIFCUL
Period: March 2004 to July 2006