Simões, Barreiro, Santos, Sousa-Silva & Tagnin (eds.) *Linguística, Informática e Tradução: Mundos que se Cruzam*, Oslo Studies in Language 7(1), 2015. 101–124. (ISSN 1890-9639 / ISBN 978-82-91398-12-9)
http://www.journals.uio.no/osla

AUTOMATIC ANONYMISATION OF A NEW PORTUGUESE-ENGLISH PARALLEL CORPUS IN THE LEGAL-FINANCIAL DOMAIN

ECKHARD BICK AND ANABELA BARREIRO

RESUMO

Este artigo apresenta o processo de anonimização automática de entidades mencionadas num novo corpo paralelo pesquisável do domínio jurídico-financeiro para o par de línguas português-inglês. O corpo resulta de memórias de tradução utilizadas em tradução profissional. Contém cerca de 40.000 pares de frases alinhadas, ou seja, frases que são traduções umas das outras. A anotação das entidades mencionadas foi feita com regras especiais da Gramática de Restrições otimizadas para o domínio jurídico-financeiro, que permitiram alcançar uma abrangência balanceada em termos de precisão de quase 90% para as entidades mencionadas candidatas (pessoa, organização, endereço e identificadores pessoais) e uma abrangência consideravelmente superior com modificações heurísticas e otimizadas para a produção. O corpo destina-se a estudos de tradução e à linguística computacional (tradução automática estatística) e será publicamente pesquisável, permitindo ao seu utilizador procurar uma palavra ou expressão e devolvendo os resultados da pesquisa em contexto na língua da busca e na sua tradução.

[1] INTRODUCTION

High quality parallel corpora are useful for many natural language processing (NLP) applications and represent an important resource for language and translation learning. However, parallel corpora available for research are scarse, and when available, they may not be of good quality. Many parallel corpora contain mistakes resulting from lexical variation or inappropriate use of the lexicon and terminology, which carries over into semantic errors and unsuitable translations Barreiro (2009). Despite quantity and quality limitations, researchers use parallel corpora for cross-language retrieval, mining terms for human and machine translation (MT), among other applications. For languages like Portuguese, the few parallel corpora available may be specific to a certain subject matter or domain, but normally do not exist for technical texts. Given the lack of parallel data available to train NLP systems, the corpus described in this paper represents an effort in making trustworthy technical data available for research pur-

poses, namely to train statistical machine translation (SMT) systems in the legal-financial domain.

One of the most important tasks in releasing a legal-financial domain corpus is to ensure that data privacy is maintained. According to the Universal Declaration of Human Rights, Article 12, everyone is entitled the right to privacy and protection of his/her personal data, information about his/her family, home, etc. Data and personally identifiable information protection applies to both individuals and corporations.

Named entities (NE) such as person or corporate name, address (physical and postal), ID number, date of birth, sex, registration data, affiliation, e-mail address, social security number, driver license, computer IP address, and so on, are examples of personally identifiable information. With respect to this legal right, there is a significant challenge for organisations to make data useful, but comply with information privacy regulations, so that personally identifiable information is not disclosed publicly. However, the effort involved in text anonymisation prevents many organizations otherwise willing to share data, from making their corpora publicly available for research purposes.

In this paper, we tackle the challenges of anonymisation of data in our corpus, keeping the data useful for research and while maintaining privacy. We will examine which types of NE are relevant to anonymisation and how they can be identified automatically, using pattern matches and contextual rules. We will also evaluate the results achieved with an adapted Named Entity Recognizer (NER) parser and discuss fall-back strategies for maximum recall.

[2] STATE OF THE ART

Corpora resources represent the driving force behind NLP systems and the source of data to train SMT systems. Several resources and corpora tools allow studying human translation and performing contrastive studies between Portuguese and English (cf. Santos (1996, chp. 8), Maia (2008), and Tagnin et al. (2009), among others). Tools for searchable corpora allow, for example, to search a word or expression in Portuguese and see how that word or expression was translated into English in different contexts. Searches can be simple text searches or advanced context searches exploiting categories like part of speech, syntactic function or semantics, and will often allow quantitative analysis, providing frequency lists, and so on.

There are several parallel corpora available for Portuguese as one of the languages involved in the corpus' translation pair, among them: the EuroParl[1], JR-

[1] http://www.statmt.org/europarl

CAcquis[2], NAT-QI[3], CorTrad[4], COMPARA[5], and Linguee[6]. Most of these parallel corpora are searchable, and therefore, constitute popular and useful tools for word and term searches.

Different corpus types present different challenges, and corpora from specialized domains are faced with problems such as data sparsity, lack of freely available sources and non-standard language that makes it difficult to use corpus tools developed for mainstream corpora (from the news and literature domain). If a corpus is to be used publicly outside a narrow circle of project researchers, further problems arise, not least copyright issues which have plagued mainstream corpora from the very outset. In the legal-financial domain, where the data is largely produced by public institutions, copyright is not the main issue, but rather the potentially high sensitivity/privacy of the data, which calls for either an impossible amount of signatures to allow their use or an effective anonymisation procedure.

An effective anonimization procedure is not a trivial task, especially if privacy is to be protected, while still retaining linguistic integrity and allowing researchers to look for interesting patterns. For text corpora, the recognition and classification of anonymisation-candidate named entities (ACNE) is the central challenge, while the actual anonymisation is relatively straightforward, as we will illustrate in (i)-(iv) in Section [5], taking the form of removal, category dummies or psydonymisation, depending on parameters such as statistical purpose (e.g. type occurencies) or desired fluency (e.g. syntactic analysis).

The first formal privacy protection model was the k-anonymity model proposed by Sweeney (2002) for structured datasets (e.g. patient records). The method consisted of removing attributes with sensitive information, such as name or address, driving license number or record, *inter alia*. Attributes represent quasi-identifiers that in combination can uniquely identify individuals. The k-measure is used to quantify the confusion risk between referents (e.g. patients) of the individual records. However, the measure and related methods are difficult or impossible to port from databases to text corpora, because corpora lack the clearly defined fields of databases. Thus, the challenge in database anonymisation is solely how to generalize information while retaining informativity, while the main problem in text corpora is the identification and classification of what is to be generalized/anonymised - something that is given from the start in database fields. This precludes a 100% safe anonymisation (de-identification) with automatic methods, whatever its (internal confusion) k-value is. On the other hand, the k-anonymity model assumes that the set of entries in the database (e.g. hospital patients) is

[2] http://langtech.jrc.it/JRC-Acquis.html
[3] http://linguateca.di.uminho.pt/nat
[4] http://nilc.icmc.usp.br/dispara/CorTrad/
[5] http://www.linguateca.pt/COMPARA/
[6] http://www.linguee.com/

PT-PT	EN-UK
h) Os que exerçam funções de administração ou de fiscalização em cinco sociedades, exceptuando as sociedades de advogados, as sociedades de revisores oficiais de contas e os revisores oficiais de contas, aplicando-se a estes o regime do artigo 76ª do &fA;, de 16 de Novembro;	h) Those, who have management or supervisory duties in five companies, excepting law firms, firms of official auditors and official auditors, subject in the latter case to the provisions of article 76 of Decree-Law no. 487/99, of the 16th of November;
i) Os revisores oficiais de contas em relação aos quais se verifiquem outras incompatibilidades previstas na respectiva legislação;	i) Official auditors, who are in any of the other circumstances of incompatibility provided in the corresponding legislation;
j) Os interditos, os inabilitados, os insolventes, os falidos e os condenados a pena que implique a inibição, ainda que temporária, do exercício de funções públicas.	j) Those, who are disqualified or debarred from the exercise of their rights, the insolvent, bankrupt and those on whom a sentence has been imposed, which involves disqualification from the exercise of public office, even if only temporarily.

TABLE 1: PT-EN legal-financial parallel corpus.

knowable, which is why they have to be anonymised against each other (hence the internal confusion measure). By contrast, a text corpus does not come with clear referents and needs NER just to identify the data "records" themselves. So in principle, the anonymisation background is the entire population, making the task less challenging in this regard. In addition, without a database structure, an internal confusion measure such as the k-value is not practically applicable. All in all, textual anonymisation is quite different from the anonymisation of data fields, with its own added problems, such as vagueness, importance of context, lack of consistency, among others. In the following sections ([3] to [6]), we will discuss how these issues can be addressed with NLP tools.

[3] CORPUS

Taking into consideration that the quality of SMT is ultimately dependent on the adequacy of the parallel corpora used for the task, and that good quality translations for a specialised domain are difficult or impossible to obtain when training MT systems on another, or more general domain, we have prepared such a specialized parallel corpus for the legal-financial domain. Apart from SMT researchers, we are also targeting human translators in need of contextualized and idiomatic translation examples. The corpus is based on translation memories used in the Metatrad[7] agency's professional translation activities, and comprises 40,000 sentences in Portuguese and English, corresponding to about 1 million tokens each.

[7] http://www.metatrad.com

[4] THE PALAVRAS NER FRAMEWORK

The PALAVRAS parser Bick (2000) is a rule-based parser using the Constraint Grammar paradigm, specifically the open source CG3 compiler[8]. PALAVRAS uses contextual disambiguation and mapping rules on morphologically multi-tagged input, where each token receives one or more readings lines (a so-called cohort). The core version of the system covers part-of-speech (POS), inflection, syntactic function and dependency links or constituent structure. However, various special grammar modules have been added over time for specific research projects or applications, such as semantic roles, semantic prototypes, valency, anaphora and NER Bick (2014). The parser has been applied to a host of Portuguese language corpora (among others, all Linguateca[9] corpora), and research versions have addressed transcribed speech, historical text and various non-standard written domains.

PALAVRAS NER participated twice in Linguateca's joint NER tasks, and performed at the top of the field. The first version (avalia SREC, Bick (2003)), taking a more static approach, tried to fix multi-word names (MWEs) before running the system's grammars - either by simple lexicon-lookup or by pattern-recognition in the preprocessor - and the only allowed post-grammar token alteration was fusion of adjacent name chains. This technique was replaced by a more dynamic, grammar based NE chunking approach in the second version Bick (2006), used for the HAREM shared task Santos et al. (2006). In this system, which we are using here, preprocessor-generated name candidate MWEs are fed to the morphological analyzer not as a whole, but in individual token parts. Thus, parts of unknown name candidates will be individually tagged for word class, inflection and, most importantly, semantic prototype class, which is used as a prime trigger for NE classification and used by the NE type mapping rules (cf. [5.3]). In addition, each part is tagged as either `@prop1` (leftmost part) or `@prop2` (middle and rightmost parts), and both tag types can be added or removed by contextual rules. At the same time, the NE category set was expanded from 6 super-categories to 41 fine-grained categories with a functional rather than lexematic definition. For our anonymisation task, we internally maintained the fine-grained set, but selected the "individual human" category `@hum` as the anonymisation category `<NAME_PERSON>` and lumped the membership group category with administrative/institutional organisations and companies into `@org` (anonymisation category `<NAME_ORGANIZATION>`).

[8] http://visl.sdu.dk/constraint_grammar.html
[9] http://www.linguateca.pt/ (2000-2014)

[5] ANONYMISATION

Anonymisation consists of the identification, categorization and neutralisation of sensitive identifying information from data. Specifically, we are addressing the task of turning a set of documents into a corpus that may be publicly used for research purposes. Identifying information can be names of people, names of organizations, social security numbers, postal and physical addresses, among others. In its broadest sense, anonymising data can be performed by the four basic methods illustrated in (i)-(iv):

(i) replacement of identifying entities with category dummies or place holders (e.g. `<NAME_PERSON>`), pseudonymization (e.g. `John Doe`) or substitution of numbers or letters (e.g. `99-99-9999` for dates)

(ii) suppression or omission of identifying entities from the released data (replacement of a proper name with `(...)` or `[-]`)

(iii) generalization or replacement of specific data (a birth date `27-02-1978`) with less data (the year of birth `1978`)

(iv) perturbation or random changes to the data (e.g. the sequence of characters `&fA;` standing for the name of a Decree-Law in Portuguese, `Decreto-Lei nº 487/99`, as represented in Table 1)

For unstructured data sets like text corpora, with a desire of maintaining textual cohesion, (i) and (ii) are most relevant. For corpus-size data sets, anonymisation is difficult or impossible to perform without automatic tools, and independently of which method is used for the actual anonymisation, the task presupposes the existence of a well-working module for NER and classification, optimally supported by a robust morphosyntactic tagger.

Because in most cases anonymisation is necessary only for certain NER types, and because false negatives are more problematic in the treatment of sensitive data than false positives, the NER process should be optimized for high recall, rather than precision, for types such as person/organisation names and corresponding identifying number expressions. This optimization need and the type of data to be anonymised, optimally calls for a tailor-made solution, as Medlock (2006) points out:

> *The inherent subjectivity of anonymisation means that different instances of the task may exhibit different characteristics even within the same domain. In light of this, it is probably impractical to deploy a solution requiring a large amount of annotated training data, bearing in mind that such training data may not generalise within the same domain, let alone across domains. In reality, application of an NLP-based anonymisation procedure would probably*

> *be carried out on an instance-by-instance basis, with rapid adaptation to the characteristics of the required solution through the use of weakly-supervised machine learning techniques.*

While we agree with Medlock on the high domain and text dependence of anonymisation, we will here follow another methodological approach for exactly this reason (domain and text dependence), and try to show that linguist-written rules supplementing a rule based parser are an effective and (in our view methodologically better) way to address both NER in general, and domain-dependent anonymisation in particular. The most relevant problem with Medlock's HMM approach is that it is statistical and needs labeled training data, which does not exist for our corpus. Even if training data were produced (manually), this would not allow the system to work well on a new domain. Also, the statistical setup does not allow users to prioritize and fix individual annotation error types, because a statistical system works as a whole, as a black box. Linguistic rules on the other hand, once written, are individually accessible and allow effective tracing, identification and fixing of errors when run on a new corpus. Thus, we will explore and evaluate how the existing PALAVRAS NER resource Bick (2003, 2006) can be used and adapted for the translation memory anonymisation of the Portuguese side of the corpus. For the actual search interface, NER marking and anonymisation will be carried over to the English side automatically with existing translation word alignment tools.

[5.1] *Preprocessing and Postprocessing*

In order to run a text parser on a corpus with a data structure, it is necessary to separate text from corpus meta-information such as paragraph id's, time stamps, author or speaker information, etc.. In the case of PALAVRAS, this means enclosing meta-information in angular brackets `<...>`, as illustrated in the preprocessed corpus header (`<20080805~134716 u 0 PT-PT>`) in Table 2. In addition, the parallel English text has to be "protected" against Portuguese analysis in the same fashion.

Other tasks for the preprocessor are the normalisation of meta-characters (the corpus uses hexadecimal `&...;` codes, such as `&'92;`, `&'93;`, and `&'94;` in Table 2), as well as OCR errors, where possible (e.g. extra spaces in numbers, confusion of `1/l`, `º/°/o` or `,/.`). In the annotated corpus, text is line-tokenized, including punctuation, and each token followed by a number of tag fields, among them NER category. In order to recreate the corpus, tokens have to be extracted and stripped of non-relevant tags. Because PALAVRAS is a syntactic parser, it splits elements like `do`, `à`, etc. into syntactic primitives, here prepositions and articles, and fuses MWEs (among them, name MWEs) into single tokens. The postprocessor has to reconstruct running text from this, attach punctuation and un-bracket metatext. Finally, and most importantly, the NER tags selected for

	PT-PT	EN-UK
Raw corpus	**20080805 134716 u 0 PT-PT** Relativamente à opinião dos formandos quanto à possibilidade de os CET**&'92**; serem vistos pela população em geral como cursos de **&'93;**segunda categoria**&'94;**,	**ENG-UK** So far as trainees**&'92** opinion regarding the possibility that CETs will be seen by the general public as **&'93;**second rate**&'94;** courses
Pre-processed corpus	**<20080805 134716 u 0 PT-PT>** Relativamente à opinião dos formandos quanto à possibilidade de os CET's serem vistos pela população em geral como cursos de "segunda categoria"	**ENG-UK** So far as trainees' opinion regarding the possibility that CETs will be seen by the general public as "second rate" courses

TABLE 2: Preprocessing.

	PT-PT
Parser output	O [o] <art> <dem> DET M S seu [seu] <poss 3S/P> <si> DET M S nome [nome] <f> <ac-cat> N M S é [ser] <vK> V PR 3S IND VFIN Ana=Borges [Ana=Borges] <hum> PROP F S $, com [com] PRP domicílio [domicílio] <build> N M S profissional [profissional] <h> ADJ M/F em [em] <sam-> PRP a [o] <-sam> <art> DET F S Av.=República=nº=50,=3º=piso [Av.=República=nº=50,=3º=piso] <address> PROP F S $, Lisboa [Lisboa] <civ> PROP F S $? <ENG-UK ... >
Post-processed corpus	O seu nome é <NAME_PERSON> com domicílio profissional na <NAME_ADDRESS>, Lisboa?

TABLE 3: Annotation example.

anonymisation have to be inserted as <NAME_. . . . > place-holders and the respective token removed. For certain unclassified name tokens, the postprocessor performs its own heuristic anonymisation (cf. [5.3.2]), treating all-uppercase names as organisations and compound names as person names.

Note that the extract illustrated in Table 3, apart from two ACNE, contains a third NE, Lisboa, which has also been classified as civitas <civ>. Geographical locations were considered public domain in our current scheme, but could easily be anonymised, given the full NER mark-up, or, in this case, fused into the address ACNE.

[5.2] *General Grammar Adaptations*

Since the NER grammar itself relies on grammatical context and needs to target words with the right POS, look up lexical properties of recognized words, etc., the quality of the underlying POS and morphological tagging is important. Ordinarily, PALAVRAS can achieve F-Scores of 98-99% for POS, but for our bilingual legal-financial domain corpus, the parser had precision problems with the proper noun class (with a certain ensuing recall loss distributed across the confusion classes). The reason for this are graphical properties of the corpus, in particular the high incidence of uppercasing.

Text type (language)	**% Uppercase words**
VEJA (pt-br)	14.45%
Leipzig internet corpus (en)	16.61%
TM3 law corpus (pt)	29.08%
TM3 nouns & proper nouns	29.51%
Leipzig internet corpus (de)	37.61%

TABLE 4: Uppercase incidence.

As can be seen from the comparison Table 4, our corpus uses twice as much uppercasing as ordinary Portuguese or English text, and almost as much as German, which uppercases all nouns as an orthographical rule. In particular, 1/3 of all nouns in the corpus were uppercase, turning uppercase from a safe into an unsafe predictor of name-hood. A further problem was that 21.40% of the uppercase words had not only the initial, but all letters uppercase, making it difficult to distinguish usually safe abbreviation names like `ONU` or `OTAN` from ordinary words in all-uppercase. Because of the ensuing highly increased ambiguity between proper nouns (`PROP`) and other parts of speech, it was necessary to change, amend and add rules in PALAVRAS' core grammar.

False Negative Names

As a default, PALAVRAS' morphological analyser will try to recognize words with uppercase initial as names, while analyzing everything else as a chain of morphemes in order to assign it POS and inflection categories. It will do the same even for upper-cased material in three cases: (a) at sentence start, (b) for noun/adjective material, and, of course, in the face of (c) multi-word all-uppercasing. For ordinary cases of increased uppercasing, such as newspaper headings or book titles, this is a good strategy, but in a corpus like ours, with many uppercased sequences, it leads to overgeneration of non-names. The grammar therefore needs to append (locally ambiguous) proper noun readings to uppercased material risking POS disambiguation errors, using so-called morphological `APPEND` rules. The simplified `CG` rule in (v) tackles cases where a name has been interpreted as a verb

or masculine noun, but where a feminine article a in combination with quotes (<*1> + <*2>) makes a name reading as brand/product/vehicle likely (e.g. a Bramir, or a Imperial):

(v)
```
("%u$1"v <HEUR> PROP F S) TARGET ("<([a-z]+)>"r + <*>) (0
<*1> + <*2>) (-1C ("<a>" <art>)) (0C V OR NMS) ;
# a "Bramir", a "Imperial"
```

The most difficult are cases where the initial-uppercase clue (for namehood) is lost because the whole word is in uppercase, e.g. EVITA COSTA (V: evitar, N: costa). Still, many extreme cases (e.g. (vi)) can be ruled out heuristically, even in otherwise uppercased context. For instance, rules can rule out multiple derivation or forbid certain affixes specifically.

(vi)
```
Cardim (N: cardo+im), Salvor (N: salva+or), Portimão (N:
porto/porta/porte+im+ão), Lombador (N: lombada+or), Godinho
(ADJ/N: godo+inho), Etar (V: eta+ar)
```

False Positive Names

The legal-financial corpus often marks key terms (as defined entities) by writing them in all-uppercase (e.g. débito da CONTA, descritos no ANEXO). This may trigger a (wrong, but in other contexts meaningful) interpretation as a name abbreviation. Thus, the rule in (vii) targets all-uppercase strings with three to six letters, if they are not flanked (NEGATE -1, NEGATE 1) by other all-uppercase tokens or line boundaries on both sides (>>>, <<<).

(vii)
```
APPEND ("%U$1"v <HEUR> PROP M/F S) TARGET ("<([a-z]{3,6})>"r
<allupper>) (NEGATE -1 <allupper>) (NEGATE 1 <allupper>)
(NEGATE -1 >>> LINK 2 <<<) ;
```

Strings of this and similar type, such as AS (article or A.S.?), CET and PAI need to be contextually disambiguated. Thus, rules exploit the fact that an all-uppercase word in parenthesis is more likely a name abbreviation than, say, a function word. On the other hand, a plural article or a plural ending in -s help discard a company name in favour of a noun abbreviation (e.g. os CET, os SPVs).

[5.3] *The NER Grammar*

Person Names

Person names are arguably the most prototypical ACNE type, and represent clearly sensitive information, asking for high recall[10]. Our system harvests category-relevant information from both the lexicon and the sentence context. Person name strings are built from left to right, with either a +HUM noun (e.g. titles, professionals, nationals, nouns ending in '-ista') or a lexicon sanctioned first name at the head. Though the former need not be anonymised, they provide a useful clue even in the absence of a recognized first name. The rule in (viii) allows attributes (e.g., o atleta profissional N.N.) in between the noun head atleta and the name N.N., but more complex rules exist to cover cases with interfering prepositional phrases (PP's) (e.g. atleta profissional de futebol N.N.), title chains or inverted, predicative cases (e.g. N.N. atleta de profissional).

(viii)
```
MAP (@hum @prop1) TARGET (<HEUR> PROP) OR (<hum> PROP) OR
(<H> PROP) (*-1 (<Hprof>) OR (".*ista"r N) OR (<Hnat>) OR
(<Hetn>) BARRIER NON-ATTR LINK NOT 0 @hum) ;
# O Atleta profissional Pedro Alvarez
```

Following HAREM conventions, titles are regarded as part of a person name. The (simplified) rule in (ix) targets only the first title in a row, and makes an exception for addresses (where person names can be part of a street name).

(ix)
```
MAP (@hum @prop1) TARGET N-TITLE (**1 PROP OR <*> + N-HUM
BARRIER (*) - N-TITLE LINK NOT 0 N-TITLE) (NOT -1 N-TITLE OR
N-STREET) ; # Sr. Alvarez, o Sr. Dr. Teófilo Alvarez, # Tio
Zeca, padre Melanaos, Exmo. Sr. Dr.= Fonseca da Paz
```

The actual name-part chaining is achieved by CG mapping @prop1 "beginning-of-name") and @prop2 ("in-name") tags in addition to the category tag @hum, allowing a later filter program to fuse the name parts into syntactic units for further processing, frame-based category mapping and disambiguation. The filter program inserts |-markers between title nouns and the name proper, and only the latter will be anonymised. Rules like (x) and (xi) allow person names to grow to the right.

(x)
```
MAP (@prop2) TARGET PRP-DE OR ("\(di|v[ao]n\)"r) (*1 (<hum>
PROP) OR (<HEUR> PROP) OR (<H> PROP) BARRIER (*) - <art>) ;
# name growing right
```

[10] This is definitely true for the proper noun part of person names, while categories like HAREM's OFFICIAL, or titles without proper nouns (e.g. Sr. Dr. Juíz) have no great need for anonymisation. The only exception for the proper noun person names are cases where a name is used to denote works of art (e.g. listen to Mozart) and possibly names in publications - where we follow HAREM conventions in using a different category, PUBLICATION

(xi) `MAP (@prop2) TARGET ("[A-Z][a-z]+"r PROP) (*-1 (@hum) CBARRIER (*) - @prop2 ) ; # PROP chain element looking left (*-1) for a @hum header with nothing (*) but other second elements (@prop2) in between (BARRIER)`

Special person name contexts in the legal-financial domain are settlements or patents named after people (e.g. `acórdaõ Lindemann` or `patente Kobashi`).

In a person name context, upper-case words wrongly tagged as other word-classes, can be marked as proper name material by the grammar (e.g. rule (xii) for the numeral `Cem` in `Sr Cem Sürük`).

(xii) `SUBSTITUTE (<*>) (<*> <prop>) TARGET (<*>) (-1 <Htit> + @hum) ;`

NE type mapping rules are ordered into sections in the grammar, with one section for each type. However, if no rule from the ordinary person name section was applicable, and if no later rules assign a different category either, then a second round of more heuristic person name mapping is performed. For example, a heuristic proper noun appearing first in a chain of proper nouns will be tagged `@hum`, if it is initial-uppercase rather than all-uppercase, and not preceded by an article or brand-noun[11], as illustrated in (xiii).

(xiii) `MAP (@hum @prop1) TARGET PROP (0 ("[A-Z][a-z]+"r <HEUR>)) (NOT -1 N-HUM-person OR PROP OR <art> OR N-BRAND) (1 PROP) ;`

Organization Names

Internally, our grammar distinguishes between different types of organisation according to the PALAVRAS and HAREM schemes 1-7.

(i) organisation (`@org`) - the umbrella category, e.g. international, NGO;

(ii) company (`@company`): e.g. `Embraer`, `A.S.`, `Ltda.`;

(iii) administrative units[12] (`@admin`): government, parliament, assembly;

(iv) institution (`@inst`): institute, laboratory, museum, university;

[11] The rule has been simplified, real rules often have multiple exceptions to cover special cases. Here, the brand case is constrained to `<foreign>`-marked proper nouns, there is a town name context exception for `São`, and the `PROP` chaining also allows the preposition `de`.

[12] This is a HAREM category and was also used for countries and towns, if they functioned as agents or cognizers. The distinction is not upheld by PALAVRAS, but only mapped later using semantic role inference, where desired. Furthermore, PALAVRAS tags place-bound administrative units as institutions, alongside shops, hotels etc.

(v) functional bodies of organisations (@suborg): boards, councils, committees;

(vi) groups with members (@grouporg): clubs;

(vii) special plural cases: @grouphum (e.g. families) and @groupofficial.

The subcategory distinctions in schemes 1-7 are not strictly necessary for anonymisation and can be lumped for this purpose, but they are useful for other corpus work, and are maintained in the grammar that works with subcategory-specific rules and sets. Scheme 1 is mapped last, using heuristic rules and the parser's lexicon (which does not recognize the subdistinctions). Scheme 4 is not anonymised on its own, only where the parent organisation appears adjacently (e.g. Conselho de Administração da|Embraer). Like person names, organisation names can be triggered by specific noun heads, that are defined as sets in the Constraint Grammar, N-COMPANY (e.g. alugadora, banco, caixa, companhia, editora, empresa, sociedade), N-ADMINISTRATION (e.g. assembleia, câmara, parlamento), N-GROUP (e.g. delegação, equipe, pessoal), among others. These trigger nouns are treated as part of the name if in uppercase and followed by a preposition (SPB, Sociedade Portuguesa de Bioquímica), but not if followed by a proper noun or all-uppercase (e.g. a alugadora Aires Baeta). In many cases, the @org category can also be triggered by a tail token at its end[13]. Thus, it is typical of corporations and formal clubs that they affix a legal-financial typer marker, such as AS, &, Co., GmbH, Lda., S.A.R.L., or generic name parts such as Holding, Consulting, Telecom, Associados. This is also exploited by the NER rules through special sets, that can then look both left (N-CLUB) in (xiv) and right (N-CLUB-POST) in (xv).

(xiv)
```
MAP (@company @grouporg @prop1) TARGET PROP (-1 N-CLUB)
    (NOT -1 <*>) ; # S.C. Braga
```

(xv)
```
MAP (@company @grouporg @prop1) TARGET PROP (NOT 0 <prop2>)
    (1 N-CLUB-POST) ; # Boavista FC
```

Note that the rules above add a @company tag alongside the normal @grouporg for sport clubs. This allows later disambiguation rules to treat the club as a company if the name string continues with an acronym, such as S.A.D., which stands for Sociedade Anónima Desportiva.

A more heuristic distinctor for organisations names is a definite article immediately left of a proper noun, or - safest - an all-uppercase abbreviation. Articles do occur with other name types, but less frequently with person names than

[13] Tail tokens also occur with person names, but they are rare (e.g. Neto, Neta, Filho), unless one also counts prepositional phrases like da Silva, dos Santos, etc.

organisation names. Provided that other name types have been targeted with their own safe rules already, it is therefore a good bet for otherwise unclassifiable names to categorize them as @company after the letter a[14] and @grouporg after the letter o. Because full names are much easier to classify than abbreviations, an internal tagging memory was used to resolve uppercase abbreviations that had already occurred earlier in the text in parentheses after a corresponding long form.

Addresses

Though the existing PALAVRAS NER module already treated addresses as a separate NER category, it did not perform well on the bilingual legal-financial domain corpus at hand, in part simply because international address formats (e.g. English, Dutch, etc.) appeared next to the "known" Portuguese ones (e.g. 10a Belmont Street, NW1 8HH, Londres), but also because of the large orthographical variation in the corpus, possibly caused by OCR or keyboard (typewriter?) limitations. Thus, there were around 20 different variants of nº, to name just one example, including n.°, n°, n.e, no.s, no., n9., n*, n", n,º, etc. plus uppercase variants, with similar variation in ordinals before words like piso and andar, or as affixes (e.g. 89-3º), as well as use of ordinal abbreviations in other words (e.g. 2o dt°, Esqº). In order to identify address NE, we again defined head nouns and tail words, as illustrated in rules (xvi) and (xvii).

(xvi) `LIST N-ADDRESS = <Lpath> "Av" "Av." "Av.a" "Av. [A-Z].*"r ... "rua" "R." "Rúa" "Via" ...`

(xvii) `LIST N-ADDRESS-POST = "Avenue" "Bd" "Boulevard" "Rd" "Road" "St" "Street" "Sq" "Square" ....`

The latter was necessary, because English addresses place the closed-class part of street names last (e.g. Hampton Road), while Portuguese (and other Romance languages) have closed-class material first (e.g. Via Appia). A third possibility is seen in German and Dutch addresses where the closed-class items are not separate words, making the use of regular expressions necessary (Bergstrasse, Meulengracht). In addition, Portuguese/Continental and English addresses place street number differently, so they mark either right or left boundaries of street addresses. @prop2 rules were used to let addresses span right over further uppercase material, added numerical material and "subaddress" words (e.g. casa, lote, piso, esq., r/c), allowing also interfering commas, letters, hyphens, slashes, the preposition de, articles and the nº token in all its variants. Though identified as such, person names inside addresses were not allowed to prevent address string

[14] Provided, of course, the parser has correctly disambiguated a as not being a preposition.

from growing right, i.e. from the head `Avenida` to the last part `Esq.` or `Piso` across the person names in bold face in the examples in (xviii) and (xix). This means that it is the larger address NE that gets marked rather than the smaller person NE inside it (`Júlio Dinis` and `Fernão de Magalhães`, in the examples).

(xviii) `Avenida *Júlio Dinis*, n.º 2 3o Esq.`

(xix) `Avenida *Fernão de Magalhães*, nº 1862.º-14º Piso`

A special topic concerns town names with postal area codes, which were treated as addresses when appearing on their own, but otherwise fused into adjacent address strings. Internationally, postal codes vary a lot, and number-only codes in particular need a recognized place name or address as context. Conversely, once identified, postal codes can help identify lexically unknown place names. In some cases, address heads or tail words are identified in connection with proper nouns, but without a number extension, subaddress or postal code. These are first tagged ambiguously as `@address @site`, and later treated by the disambiguation grammar with full context, lumping these cases together with other site words such as `estação`, `estádio`, `mina`, and `shopping`, among others. Corpus-wise, we decided that street names, etc. used on their own are not precise enough to need anonymisation.

Identifying Numerical Expressions

Numerical expressions can help identify a person or company either directly or indirectly. Indirect numerical identifiers such as age, weight, income, etc. can help a detective choose between several people, but are not identifiers on their own, and obviously numerals in addresses present no problem because they will be anonymised together with the address as a whole. Problematic, on the other hand, are direct numerical identifiers that are long enough to be unique, or appear in a feature-attribute pair that makes them unique. Examples for such numerical ACNEs are:

(i) telephone or telefax numbers

(ii) email or personal/company websites, passport numbers, and tax numbers (IRC, CIRC, RFI, NIF - person, NIPC - company)

(iii) bank accounts, NIB (IBAN)

(iv) invoices, file numbers, `NUIPC (identificador de processo crime)`

The safest way to identify these cases is with a triggering head noun in the left-hand context, defined as a set including the above abbreviations as well as

variants of telefone/telef/tel/tel/fax/telefax, passaporte, etc., and more general expressions such as imatricul*, identific*, etc. In these contexts, virtually any numerical expression of a certain length, with a mixture of digits, letters and -/. will be an individual identifier. Heuristic numerical pattern matches (bold in the rule exemplified in (xx)) can also be used in the absence of privacy trigger nouns, based on the trigger word nº (número) alone, but only in the absence of competing specific triggers for public identifiers.

(xx)
```
MAP (@nameid) TARGET (<cif> NUM) (*-1 ("nº") OR ("número")
BARRIER (*) - (@nameid) - IT LINK NEGATE *-1 N-PUB OR N-COPY
OR (<media>) OR (<ABBR>) CBARRIER N-PRIVATE OR NON-KOMMA -
<prop2> - PRP-COM - PRP-DE - ("nº") - <artd> - <*>) \bf{(0
(".*[0-9][0-9]= [0-9][0-9][0-9A-Z=]+"r) OR ("[0-9=]{7,}"r) OR
("[0-9].*[.-=][0-9]{4,}.*"r) OR ("[0-9]+/[0-9].*"r))} ;
```

Together with the ACNE type @nameid, we introduced @pub, for public identifiers. This name type uses similar rules and number patterns, but looks for a special N-PUB set (e.g. aviso, circular, decreto, despacho, lei, parecer, resolução) or the N-COPY set used for publication names and works of art (e.g. <sem-r> readable: books, etc., <sem-w> watchables: films, etc., <sem-l> listenables: concerts, etc.). The same sets are used as exceptions (e.g. NEGATE) in the heuristic @nameid rule quoted in (xx). Leaving aside other purposes and application of tagging private and public identifiers, it should be born in mind that anonymisation for these categories is easily and undestructively achieved by substituting, for example, the digit 9 for all other digits in the text. So far, we only considered a similar option (letter substitution) for the last identifier category, emails and web addresses (e.g. aaaaa.aaa@aaa.aa, aaaa://aaa.aaaa.aa/), but the substitution method should be considered as a last resort to achieve the highest possible anonymisation recall at publication time in the face of legal constraints.

[6] EVALUATION

[6.1] *Evaluating the Adapted PALAVRAS Parser*

An evaluation subset of 6,800 aligned paragraphs, with 80,800 Portuguese words, was extracted randomly from the total data set, based on the last digits of the paragraph id's. The Portuguese part of the evaluation data was automatically annotated for NER strings and their categories, using the adapted PALAVRAS parser. Since purely numerical NERs can be easily treated with a coverall digit replacement operation, our focus was on non-numerical candidates for anonymisation (human names, organisation names and addresses), including only one numerical category, individual identifiers. These categories were then inspected and

	Cases	Recall	Precision	F_1-score
hum	263	87.83	87.50	87.66
org	871	93.69	86.53	89.97
address	38	81.58	91.18	86.11
nameid	54	60.71	87.18	71.58
all	1229	88.32	86.88	86.68

TABLE 5: Performance by category.

ACNEs	Recall	Precision	F_1-score
untyped	90.71	89.22	89.96
typed	88.32	86.88	87.59
untyped, chunked (0.5)	88.36	86.91	87.63
typed, chunked (0.5)	86.26	84.84	85.54
untyped, chunked	86.02	84.61	85.31
typed, chunked	84.19	82.81	83.49

TABLE 6: Performance according to different evaluation metrics.

evaluated with regard to NER span and category. In addition, the text was manually annotated for false negatives of the same name types. The whole evaluation process was performed solely by one author, based on his linguistic expertise, albeit without parallel multi-annotator controls. All in all, the parser found 1,142 out of 1,259 possible ACNEs, making for a recall of 90.71%, and suggested 128 false positives, equalling a precision of 89.22%. About 52% of the false positives were non-ACNE name types, 48% were uppercase nouns. Confusion across anonymisation categories was fairly rare (2-3%), the most common error being to read uppercase person name abbreviations (e.g. `AH`, `JB`) as company names. For individual ACNE type recognition, overall recall was 87.78% and precision 86.98%, with organisations performing best, and name id's performing worst.

In 48 cases, the recognized ACNEs were too short, in 6 cases too long, amounting to a 5% chunking error rate. Typical cases were missing address details or organisation type extensions (e.g. `95_Wilton_Road & 201, Londres, SW1V 1`, or `VESTAS_Mediterranean A / S`).

If chunking errors are included in the accuracy calculation, both recall and precision drop with a couple of percentage points, if any mismatch is counted as an error (CoNLL evaluation scheme Sang & De Meulder (2003)), or 1 percentage point if partial hits are counted half (MUC evaluation scheme[15]).

[15] http://www.itl.nist.gov/iaui/894.02/related_projects/muc/muc_sw/muc_sw_manual.html

On the background of a real-world anonymisation task, it should be born in mind that almost all span (chunking) errors are irrelevant to anonymisation, since, for example, CEP codes without street addresses or an A/S suffix without a company name, are already quite anonymous. There are few published evaluation scores for text anonymisation in the literature, let alone for the same language and domain, but one comparable approach is Medlock's work on an English email corpus Medlock (2006), where F-scores between 54.42 and 63.99 (depending on the metrics) were achieved for selective anonymisation (roughly our set of categories) and 65.76-73.87 for blanket anonymisation (all potential NER types), achieved with the LingPipe HMM tagger[16]. However, a direct performance comparison is problematic, because Medlock uses a pre-tagged gold corpus rather than results inspection, and intentionally limited training corpus size. Another comparison, for typed NER, albeit on the easier newspaper domain, is the CoNLL 2003 shared task[17], where the best systems achieved F-Scores of 88.76 for English and 72.41 for German, possibly reflecting the before-mentioned difficulty of identifying names in the face of noun-uppercasing. PALAVRAS own typed F-Score on the mixed HAREM domain was 63.0/68.3 for absolute/relative category classification.

[6.2] *Increasing Recall*

As discussed earlier, recall is more important for anonymisation than precision, and as intended, evaluation shows that for the overall tagging task this goal was achieved. However, the R-P difference is not big, and mainly valid for the <org> class, while <address> and <nameid> have a low recall. Obviously, recall for the latter categories can be increased by better pattern matching rules, addition of further address fields and id trigger words, and this was done after the above round of evaluation. But on the other hand, a more radical recall-increasing approach is desirable for a real world application, where anonymisation close to 100% is necessary if human post-editing is to be avoided. A case-for-case inspection of false negatives shows that add-on strategies can be used exploiting the following patterns:

(i) treating all-uppercase strings as ACNEs, or - trusting the parser's POS disambiguation - all-uppercase proper nouns (in the evaluation run, all such false negatives had been recognized as PROP, just not the right type of PROP). Typing as <org> would also be possible (only 25% were not <org>).

(ii) treating all compound proper names as ACNEs, i.e. strings where the parser fused 2 or more upper case tokens into one, and tagged it as PROP. These cases were about equally distributed between person and organisation names.

[16] http://www.alias-i.com/lingpipe/
[17] http://www.cnts.ua.ac.be/conll2003/ner/

(iii) treating single-token PROP as ACNEs, if the parser marked them as <foreign>. Again, these cases covered a mixture of person/organisation types.

(iv) treating camel case as ACNE (of <org> type).

(v) treating all numerical expressions as ACNEs. These were mostly of <nameid> type, but <address> in cases where uppercase letters were followed by digits.

The above strategies capture 88.8% of all false negatives. Of the remaining 13 cases, one was partially recognized already (person name within organisation name), and would thus get anonymised anyway; the rest consisted of ordinary words used as names (e.g. Tranquilidade) or ambiguous with names at sentences start (e.g. Marques), names with case errors (e.g. o opbbr é uma sociedade, o Oi) or mistyped/untyped PROP, the latter sometimes as part of what the parser regarded as a longer PROP chain. Given this distribution of cases, almost total anonymisation recall could be achieved by treating all PROP-tagged strings as ACNEs. Table 7 below shows how the individual strategies affect recall, and - for the non-numerical types - precision.

The price in precision loss for applying the above strategies is, of course, fairly high. The safest strategy is all-uppercase PROP, where recall gain out-weighs precision loss 5:1 and where recall for the main affected category, <org>, climbed to over 96%. Treating all complex PROP as ACNEs is much less safe, and would sink precision into the 50% bracket. However, only applying this strategy to complex PROP not otherwise categorized, still matches most false positives of this type[18], while leading to a more tolerable precision loss, only a little above the corresponding recall gain. It is beneficial especially for person names (8% recall gain), bringing them on par with <org> coverage. Camel case and the <foreign> tag are much more "expensive" in precision terms, and risk including typos and, for the latter, a good portion of ordinary English words (> 40%). General numerical anonymisation, finally, captures virtually all id and address information and is unproblematic to use - irrespective of precision loss - because textual cohesion suffers much less from digit replacement than it does when upper case noun chains and proper nouns are replaced with dummies.

We conclude from the above that apart from numerical anonymisation, two fallback strategies are cost-efficient enough to be used - treating remaining all-uppercase as <org> and unclassified compound proper nouns as <hum>. All in all, this achieves a recall for ACNEs of 98.24%, arguably good enough for purely

[18] The target group of compound names, person names, are mostly cases where all elements of the MWEs are individually proper nouns, while compound names with uppercase noun elements often belong to other classes. It is exactly this trait that makes it likely that the parser already has found a classification for them, based on its knowledge about semantic noun classes.

	False negative	R gain	P loss	Cumulative recall, untyped	Typed recall effect for main category	Typed recall gain
no recall heuristics				90.29%		
all-upper PROP	30	2.35%	0.47%	92.64%	<org> 96.16%	2.97%
compound PROP	41	3.43%	5.51%	96.23%	<hum> 95.82%	7.99%
numerical expressions	20	1.76%		97.91%	<nameid> 100%	39.29%
uppercase + numerical	4	0.25%		**98.24%**	<address> 92.16%	10.58%
<org>				99.13%		
<hum>				98.10%		
<address>				100.00%		
<nameid>				100.00%		
<foreign> PROP	6	0.50%	3.77%	98.83%		
<camelcase> PROP	2	0.17%	0.50%	98.91%		
other PROP	10	0.84%		99.75%		
other	3					

TABLE 7: Effect of recall heuristics.

automatic corpus treatment. As a futher safety measure, we provide the option of including publication names and public identifiers in the anonymisation, because the former may contain person names, and the latter may be confused with private (person) identifiers.

[6.3] *Parallel Corpus Anonymisation*

Anonymisation of the English part of our parallel corpus could of course be performed by independent anonymisation using the same techniques as for the Portuguese part, specifically by using the English sister parser of PALAVRAS, EngGram. However, we opted for a different, alignment-based solution, where ACNEs marked in the Portuguese text were aligned with matching strings in the parallel English sections, transferring the already established NE category tags. This method ensures that the same category definitions and span conventions are used in the two languages, and also automatically establishes referent links between Portuguese and English ACNEs, which is useful because many paragraphs contain many ACNEs, and in anonymised form, without the actual name string, it is not always easy for the reader to establish which goes where in the translation. Alignment is achieved in 3 steps:

(i) All Portuguese ACNEs are numbered, and where the individual strings match corresponding English strings, the latter are tagged/anonymised with the same category and number. This method captures most person names, addresses and numerical name identifiers, because these name types do not differ much across languages. To guard against typing/OCR errors and small orthographical differences, search strings were case-insensitive and adapted as regular expressions with optional dummy characters replacing "unsafe" characters (dots, strings, spaces, accents, etc.).

(ii) A pure pattern-based ACNE identification was performed for numerical expressions with variability across languages (dates) and to identify name identifiers that were not present in the Portuguese part because they were missing, omitted or anonymised in that language (e.g. `&fA;`). If possible (e.g. dates), these new English ACNEs were then back-aligned with not-yet aligned Portuguese ACNEs and numbered correspondingly.

(iii) The remaining unaligned Portuguese ACNEs were typically multi-part organisation names, whose English equivalent was a part-by-part translation, or all-uppercase abbreviations where the Portuguese and English letter order differed (e.g. `NATO - OTAN`). In these cases, we tried to match English uppercase strings of similar makeup, for example matching a name with a lowercase word in the middle with an English corresponding sequence of uppercase-lowercase-uppercase words. Even this kind of alignment was

quite successful, not least, because we used already-aligned material to constrain the left and right borders of the search space.

PT-PT	EN-UK
(A) As Partes são duas sociedades consti tuídas sob o domínio integral da <_ORGANISATION>, sociedade adjudicatária da Fase A do denominado "Concurso das Eólicas", conforme Contrato celebrado com a (agora designada) "<_ORGANISATION_ ADMIN>" ("<NAME3_ORGANISATION_ ADMIN>") em <NAME4_DATE>, nos termos do qual, e dos respectivos anexos, a <NAME5_ORGANISATION> e a <NAME6_ORGANISATION> assumiram os direitos e obrigações relacionados co m as actividades de promoção dos Parques Eólicos e do Projecto Industrial previs tos no mesmo Contrato com a "<NAME7_ORGANISATION_ ADMIN>", respectivamente;	(A) The Parties are two companies incorporated under the exclusive control of <NAME1_ORGANISATION>, a company, which has been awarded the contract for Phase A of the "Wind power Tender", in accordance with a Contract with the <NAME2_ORGANISATION_ ADMIN> ("<NAME3_ORGANISATION_ ADMIN>"), as it is now designated, signed on the <NAME4_DATE>. According to the terms of the said Contract with the <NAME7_ORGANISATION_ ADMIN> and the annexes thereof, <NAME5_ORGANISATION> and <NAME6_ORGANISATION> respectively assumed the rights and obligations in relation to the promotion of the Wind Parks and Industrial Project envisaged in the said Contract;

TABLE 8: Annotation example.

[7] CONCLUSIONS AND FUTURE WORK

We have presented a new 1 million token Portuguese-English parallel corpus, covering the legal-financial domain, and shown how an existing general-purpose NER-parser can be adapted for robust text anonymisation, achieving F-scores of 80-90% on the NER task as such, and over 98% ACNE recall for the anonymisation-optimized system as a whole. We were also able to show that alignment can be used to propagate anonymisation between languages. Website publication of the corpus with a suitable search-interface is planned for the immediate future, but we also need to investigate how well our anonymisation method carries over into other domain or language pairs, so that a more general database and search tool for translation memories can be created.

ACKNOWLEDGMENTS

We would like to thank Metatrad for making it possible to create the corpus described here, and for allowing us to make it publicly available for searching. We also would like to thank Hugo Gonçalo Oliveira and Miriam Leite for relevant comments that helped improve this paper. Anabela's work was funded by FCT through grant SFRH/BPD/91446/2012).

REFERENCES

Barreiro, Anabela. 2009. *Make it Simple with Paraphrases: Automated Paraphrasing for Authoring Aids and Machine Translation*: Universidade do Porto PhD dissertation.

Bick, Eckhard. 2000. *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*: Aarhus University PhD dissertation.

Bick, Eckhard. 2003. Multi-Level NER for Portuguese in a CG Framework. In Jorge Baptista, Isabel Trancoso, Maria das Graças Volpe Nunes & Nuno J. Mamede (eds.), *Computational Processing of the Portuguese Language: 6th International Workshop, PROPOR 2003. Faro, Portugal, June 2003 (PROPOR 2003)*, 118–125. Springer.

Bick, Eckhard. 2006. Functional Aspects in Portuguese NER. In Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira & Maria Carmelita Dias (eds.), *Computational processing of the portuguese language, proceedings of propor 2006*, 80–89. Springer.

Bick, Eckhard. 2014. Palavras, a constraint grammar-based parsing system for portuguese. In Tony Berber Sardinha & Thelma de Lurdes São Bento Ferreira (eds.), *Working with portuguese corpora*, 279–302. Bloomsbury Academic.

Maia, Belinda. 2008. Corpógrafo V4 - Tools for Educating Translators. In Elia Yuste Rodrigo (ed.), *Topics in Language Resources for Translation and Localisation*, 57–70. John Benjamins Pub. Co.

Medlock, Ben. 2006. An Introduction to NLP-based Textual Anonymisation. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik & Daniel Tapias (eds.), *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, 1051–1056.

Sang, Erik F. Tjong Kim & Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL 2003*, .

Santos, Diana, Nuno Seco, Nuno Cardoso & Rui Vilela. 2006. HAREM: An Advanced NER Evaluation Contest for Portuguese. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odjik & Daniel Tapias (eds.), *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, 1986–1991.

Santos, Diana Maria de Sousa Marques Pinto dos. 1996. *Tense and aspect in English and Portuguese: a contrastive semantical study*: Instituto Superior Técnico, Universidade Técnica de Lisboa PhD dissertation.

Sweeney, Latanya. 2002. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10(5). 557–570.

Tagnin, Stella O. E., Elisa Duarte Teixeira & Diana Santos. 2009. CorTrad: a multiversion translation corpus for the Portuguese-English pair. *Arena Romanistica* 4. 314–323.

CONTACTS

Eckhard Bick
University of Southern Denmark
eckhard.bick@mail.dk

Anabela Barreiro
INESC-ID
anabela.barreiro@inesc-id.pt