UNIVERSIDADE DE ÉVORA
Mestrado em Engenharia Informática

Resolução de Anáforas e o seu Impacto em Sistemas de Recuperação de Informação

Luís André da Rocha Rosário

Orientador: Prof. Doutor Paulo Quaresma

Outubro de 2007

UNIVERSIDADE DE ÉVORA

Mestrado em Engenharia Informática

Resolução de Anáforas e o seu Impacto em Sistemas de Recuperação de Informação

Luís André da Rocha Rosário

Orientador: Prof. Doutor Paulo Quaresma

Outubro de 2007

Prefácio

Este documento contém uma dissertação intitulada *Resolução de Anáforas e o seu Impacto em Sistemas de Recuperação de Informação*, um trabalho realizado pelo aluno Luís André da Rocha Rosário[1], do Mestrado em Engenharia Informática, na Universidade de Évora.

O autor da dissertação é licenciado em Engenharia Informática e de Computadores, pelo Instituto Superior Técnico, da Universidade Técnica de Lisboa.

O orientador da dissertação é o Professor Doutor Paulo Quaresma[2], do Departamento de Informática, da Universidade de Évora.

A presente dissertação foi entregue em Outubro de 2007 e já inclui as críticas e sugestões feitas pelo júri.

[1] lrosario@eseb.ipbeja.pt
[2] pq@di.uevora.pt

Resumo

Este trabalho apresenta dois estudos de corpora, uma avaliação da metodologia de *Centering*, aplicada à resolução de anáforas pronominais de terceira pessoa, em língua portuguesa e a verificação do impacto dessa resolução nos sistemas de recuperação de informação SENTA (*Software for the Extraction of N-ary Textual Associations*) e *Google Desktop*.

O estudo dos dois corpora, jurídico e jornalístico, tem por objectivo identificar características específicas das relações anafóricas pronominais, nos respectivos documentos. Estas características servem de ponto de partida para o processo de resolução de anáforas. O objectivo final é o de verificar se existem vantagens em integrar a resolução de anáforas em sistemas de recuperação de informação.

Abstract

Anaphora Resolution and its Impact on the Information Retrieval Systems

This work presents two corpora studies, an evaluation of *Centering* for third person pronouns resolution, in Portuguese texts and the verification of the impact of that resolution on the information retrieval systems SENTA (*Software for the Extraction of N-ary Textual Associations*) and *Google Desktop*.

The legal and journalistic corpora's study is to identify specific features related to the pronominal anaphoric relations in the texts. These features serve as background knowledge for the pronominal anaphora resolution process. The final goal of this study is to verify the integration of the pronoun resolution into information retrieval systems.

Agradecimentos

Aproveito esta oportunidade para agradecer às pessoas que permitiram que este trabalho se concretizasse.

Agradeço à minha esposa toda a ajuda e incentivo mostrado ao longo deste trabalho.

Agradeço ao meu orientador Prof. Doutor Paulo Quaresma toda a ajuda e compreensão.

Agradeço à equipa do projecto SENTA pela disponibilidade de processamento dos meus documentos.

Índice

Listas de Figuras

Lista de Tabelas

Lista de Abreviaturas

C – Cobertura

Cb – *Backward-looking Center*

Cf – *Foward-looking Centers*

Cp – *Preferred Center*

P – Precisão

SENTA – *Software for the Extraction of N-ary Textual Associations*

SN – Sintagma Nominal

SRI – Sistema de Recuperação de Informação

SV – Sintagma Verbal

RA – Resolução de Anáforas

RI – Recuperação de Informação

TREC – *Text Retrieval Conference*

TSC – Taxa de Sucesso Crítico

U – *Utterance*

V – Verbo

1. Introdução

Esta dissertação apresenta um estudo que remete não só para a área do processamento de linguagem natural, como também para a área interdisciplinar da recuperação de informação. Partindo da descrição de um processo de resolução de anáforas pronominais, em língua portuguesa, baseado na metodologia de *Centering*, explica-se a utilização dessa resolução em dois sistemas de recuperação de informação: SENTA (*Software for the Extraction of N-ary Textual Associations*) e *Google Desktop*.

1.1. Motivações

Uma das motivações para este trabalho surge do interesse pela criação de sistemas computacionais que permitam simular a capacidade humana do uso da linguagem e da manipulação de informação. Ou seja, dois objectivos que se encontram no âmbito das ciências computacionais e que se traduzem, por exemplo, no reconhecimento e sumarização de discursos, nos sistemas de tradução, na interpretação de diálogos, na resposta automática a perguntas e na extracção de informação, entre outros.

Têm sido desenvolvidos muitos estudos nestes domínios e a sua utilidade é visível. Contudo, existem obstáculos que limitam essas aplicações e que, ao mesmo tempo, constituem um desafio para o constante desenvolvimento das mesmas. Um dos obstáculos diz respeito à necessidade de conhecimento do mundo real para a interpretação de situações que ocorrem na língua natural. Veja-se o exemplo (1.1.).

(1.1.) **O Ricardo e o Jorge** assistiram aos **jogos**, **eles** foram espectaculares.

No exemplo, o pronome pessoal **eles** refere uma entidade já introduzida no discurso: **jogos**, no entanto, essa associação, embora possa ser correctamente feita por um ser humano, já não é tão óbvia e simples para um sistema computacional. Pois, na realidade existem, nesta frase, duas entidades possíveis para se associarem ao pronome: **O Ricardo e o Jorge** e **jogos**. Esta é uma tarefa referente aos sistemas de resolução de anáforas, a qual se procura desenvolver, de forma a tornar-se o mais eficiente possível.

Já é certo, portanto, que a resolução de anáforas é fundamental para as aplicações em processamento de língua natural, acima referidas. Actualmente, procura-se verificar se essa resolução se pode tornar também importante na área da recuperação de informação.

Nos últimos anos, as inovações científicas, tecnológicas, culturais e sociais têm acontecido de forma rápida e as pessoas precisam constantemente de actualizar os seus conhecimentos, para se adaptarem às mudanças. É neste sentido que a *web* constitui uma das principais fontes de informação actuais, em que, através de motores de pesquisa, destacando-se o *Google*, a informação é distribuída em grande volume e passa por um processo constante de criação, actualização, armazenamento e procura. Estas são funções dos sistemas da recuperação de informação que tem vindo a ser uma área de estudo cada vez mais importante, onde se tentam ultrapassar obstáculos como o da necessidade de interpretar os diversos tipos de documentos disponíveis em vários formatos e o da ordenação de resultados, segundo uma relevância e como resposta a uma pesquisa. O principal objectivo de um sistema de recuperação de informação é, assim, recuperar informação (contida em documentos) que possa ser útil ou relevante para o utilizador. Estes sistemas ordenam os documentos de acordo com o grau de relevância em relação ao que foi pedido pelo utilizador.

Actualmente são poucos os estudos sobre o desenvolvimento e a eficácia de processos de resolução de anáforas na língua portuguesa, uma das mais faladas no mundo. Este facto constitui, por si só, um motivo para o trabalho aqui desenvolvido. Para além disto, se se considerar que se pode obter benefícios com a integração da resolução automática de anáforas nos sistemas de recuperação de informação, essa motivação aumenta. É pois importante que os vários milhões de pessoas que falam o português se integrem na "Sociedade da Informação", actualizando os seus conhecimentos e adaptando-se às constantes mudanças do mundo actual.

Neste seguimento, o presente estudo utiliza como um dos corpora de trabalho, um conjunto de documentos jurídicos (Pareceres da Procuradoria Geral da República de Portugal), no sentido de aqui prestar também um contributo para os sistemas de recuperação de informação do Projecto PGR (Procuradoria Geral da República de Portugal). Este projecto teve como objectivo o agrupamento e visualização automática de documentos legais num sistema cooperativo de pesquisa de informação na *web*.

1.2. Objectivos

Este trabalho tem como principal objectivo verificar o impacto da resolução de anáforas pronominais em sistemas de recuperação de informação textual. Se esse impacto trouxer benefícios para o melhoramento de alguns sistemas de recuperação de informação, então podemos afirmar que, perante a necessidade de uma pesquisa de informação, em língua portuguesa, de um utilizador, poderá obter-se uma resposta mais útil ou mais relevante para esse mesmo utilizador. Este objectivo faz com que, na prática, se pretenda melhorar o desempenho dos sistemas de recuperação de informação, nomeadamente, o SENTA e o *Google Desktop*, de modo a torná-los mais eficazes ao nível da relevância da informação recuperada.

Para chegar a este objectivo final, propôs-se, neste trabalho, atingir um outro objectivo, que é o de avaliar o método de *Centering* na resolução de anáforas pronominais pessoais de terceira pessoa, em textos de língua portuguesa. Ou seja, pretende-se criar, através do referido método, um sistema computacional de resolução de anáforas que determine o antecedente de um termo anafórico, num discurso em língua portuguesa.

1.3. Principais Contribuições

De uma forma geral, este trabalho representa uma contribuição para a área de processamento de linguagem natural e de recuperação de informação. Mais especificamente contribui com um estudo sobre o processo de resolução de anáforas pronominais pessoais de 3ª pessoa, em língua portuguesa, e com um sistema computacional que as resolve. Por outro lado, foi feito um estudo sobre o impacto da integração dessa resolução de anáforas pronominais em sistemas de recuperação de informação, nomeadamente os sistemas SENTA e *Google Desktop*.

Em resumo, as contribuições deste trabalho são:

- A apresentação de um estudo comparativo de diferentes metodologias de resolução de anáforas pronominais.

- A implementação de um sistema de resolução de anáforas pronominais pessoais, baseado na teoria original do *Centering*.
- A aplicação do método proposto de resolução de anáforas a dois corpora, um jurídico e outro jornalístico.
- A avaliação dos resultados obtidos no processo de resolução de anáforas pronominais.
- A implementação de um sistema de substituição de pronomes pessoais pelo seu referente anafórico.
- A análise comparativa das respostas dadas pelos sistemas de recuperação de informação (SENTA e *Google Desktop*), com e sem resolução de anáforas pronominais.
- A avaliação do impacto da resolução de anáforas nos sistemas de recuperação de informação utilizados.

1.4. Estrutura

A dissertação é constituída por seis capítulos, em que no primeiro se faz uma introdução, onde constam as motivações e os objectivos deste estudo, bem como a forma como se encontra estruturado.

Os restantes cinco capítulos serão a seguir resumidos:

Capítulo 2 – Este capítulo pode dividir-se em duas partes: uma referente ao estado da arte da resolução de anáforas e outra referente ao estado da arte da temática da recuperação de informação.

No que diz respeito ao tema da resolução de anáforas, começa-se por definir anáfora. Descreve-se a seguir o processo de resolução de anáforas; os conhecimentos que este implica; e algumas abordagens para esse processo. É ainda feita uma reflexão sobre a evolução e tendências nesta área.

Relativamente à recuperação de informação, é apresentada a sua definição e de seguida, descreve-se o seu processo; o pré-processamento (filtros); as técnicas de

indexação; e os modelos de recuperação de informação. Também aqui se expõe um subcapítulo sobre a evolução e as tendências nesta área.

Capítulo 3 – Neste capítulo descreve-se a abordagem proposta para a resolução de anáforas pronominais, o *Centering*. São também descritos os sistemas de recuperação de informação utilizados: SENTA e *Google Desktop*.

Capítulo 4 – Esta parte do trabalho é reservada à descrição do tipo de anáforas resolvidas; do corpus; e da implementação da resolução de anáforas, onde se explicam as etapas de identificação e filtragem e de classificação e selecção. É ainda descrita a forma como se substituiu o referente anafórico pelo seu antecedente; e comparam-se as respostas dadas pelos sistemas de recuperação de informação utilizados, onde são abordadas as comparações das respostas dos sistemas SENTA e *Google Desktop*.

Capítulo 5 – Aqui começa-se por descrever os métodos usados na avaliação e de seguida apresentam-se os resultados da resolução de anáforas pronominais e os resultados do processo de recuperação de informação (SENTA e *Google Desktop*). Por fim, são analisados os resultados.

Capítulo 6 – Neste capítulo chega-se a uma conclusão, referindo o que foi feito em relação ao objectivo inicial deste trabalho, as limitações que se encontraram e, por fim as propostas de trabalho futuro.

No final, são apresentadas as referências bibliográficas, anexos referentes aos corpora e outras tabelas complementares.

2. Estado da Arte

Neste capítulo é descrito o "estado da arte" da resolução de anáforas (em 2.1.), bem como, necessariamente, o da recuperação de informação (em 2.2.). Com este estudo, pretende-se compreender os princípios fundamentais subjacentes a estas duas áreas de investigação.

2.1. Resolução de Anáforas (RA)

Antes de iniciar este tópico, há a necessidade de definir o conceito de anáfora, reconhecer as suas variantes, para só depois se passar a compreender o processo de resolução de anáforas.

2.1.1. Anáfora / Referência

A anáfora é, na sua essência, um processo de referenciação, e por conseguinte, a sua definição associa-se ao conceito de **referência.**

Estabelecer referências é uma das funções da linguagem, que se entende pela relação que une uma expressão linguística a um objecto do mundo real ou imaginário. Isto significa que, por meio da linguagem, (por exemplo, de sintagmas nominais que descrevem entidades do mundo), um ouvinte, ou leitor consegue identificar seres, factos, eventos ou acções no universo do discurso. Sendo o discurso constituído por um conjunto de frases que se referem a um determinado assunto, ele só poderá apresentar sentido e coesão, através de um processo de referenciação.

Há, no entanto, que salientar dois tipos de referência, segundo [Halliday & Hasan (1976)]:

- **situacional** ou **exofórica**: O referente[3] não se encontra no texto. A interpretação da referência depende de um contexto situacional específico. Veja-se o exemplo a seguir:

[3] Entenda-se como referente a entidade a que se remete por meio de um elemento de referência.

(2.1.) **Aquela** é a mais interessante. (diante de uma série de obras de arte);

- **textual** ou **endofórica**: quando o referente se encontra expresso no próprio texto. Dentro desta classificação incluem-se a referência textual anafórica (2.2.), em que o referente é introduzido anteriormente, e a referência textual catafórica (2.3.), em que o mesmo é introduzido *a posteriori*.

(2.2.) Ela procurou os familiares, mas não **os** encontrou.
(2.3.) Realizara todos os seus sonhos, menos **este**: o de entrar para a Universidade.

O tipo de referência aqui em causa é a referência textual anafórica.

Assim, a partir da introdução de uma entidade (ser, facto, evento ou acção), no discurso, pode ocorrer a retoma dessa entidade com o emprego de termos representados, por exemplo, por um pronome. Estes são, não só um recurso estilístico, evitando a repetição de uma expressão já mencionada, mas também um recurso de coesão textual, cuja função é estabelecer relações textuais. Veja-se o exemplo (2.4.):

(2.4.) *"**O património**, ou, talvez mais acertadamente, os direitos que **o** constituem podem ser objecto de actos e negócios jurídicos (13).*
*Há, nomeadamente que assegurar a **sua** gestão, mediante actos e negócios..."*

Neste caso, os termos representados pelos pronomes **o** e **sua** são termos de referência que retomam a mesma entidade, anteriormente introduzida no discurso: **O património**.

A partir desta noção de referência e mais especificamente, de referência textual anafórica, podemos definir a anáfora. Uma **anáfora** é uma relação discursiva que ocorre quando um referente é introduzido e, mais adiante, retomado por meio de algum elemento de referência, tal como um pronome. A anáfora é assim composta por um objecto de referência, a que chamamos de **antecedente** e um **termo anafórico**, que representam a mesma entidade do mundo real ou imaginário. Antecedente e termo anafórico são, por essa razão, **co-referentes**. Em (2.4.) temos os pronomes **o** e **sua** que desempenham o papel de termos anafóricos e **O património** como seu antecedente. Os

dois pronomes representam a mesma entidade do mundo real, sendo, por isso, também co-referentes.

O processo de determinação do antecedente de um termo anafórico é denominado **resolução** ou **cálculo**.

2.1.1.1. Tipos de Anáfora

De acordo com as designações gramaticais, podemos referir vários tipos de anáfora.

Anáfora Pronominal, como o próprio nome indica, diz respeito a todas as anáforas na forma de pronomes pessoais (2.5.), demonstrativos (2.6.), possessivos (2.7.) e relativos (2.8.). Os pronomes pessoais de 3ª pessoa são invariavelmente considerados termos anafóricos, pois designam qualquer ser referido no discurso, enquanto que os pronomes de 1ª e 2ª pessoa designam sempre a pessoa que fala ou com quem se fala, podendo não ser previamente introduzidos no discurso, ou não ser identificados sem recorrer a informação contextual, não linguística.

Tomemos como exemplo o seguinte:

(2.5.) Rega bem **as plantas**, para que **elas** cresçam com vigor.

(2.6.) Eles ensinaram-me **uma canção**, mas não é **aquela** que tu gostas.

(2.7.) **O cantor** animou o público com a **sua** actuação.

(2.8.) Aquele é **o pintor que** mais sucesso teve na exposição.

Anáfora Nominal é um tipo de anáfora em que o termo anafórico é constituído por um sintagma nominal, SN, que pode tomar a forma de: a) repetições literais, b) adjectivos que qualificam núcleos omitidos em sintagmas nominais, c) hiperónimos[4], d)

[4]Hiperónimos são vocábulos de sentido mais genérico em relação a outro. Por exemplo, "árvore" é um hiperónimo de macieira, pereira, etc.

hipónimos[5], e) holónimos[6], f) merónimos e g) um SN do mesmo campo lexical do termo antecedente.

Veja-se um exemplo de cada situação.

(2.9.) Tenho **um telemóvel** e gosto muito de ter **um telemóvel** (situação a)).

(2.10.) Comprei **um pássaro vermelho** e outro amarelo. **O vermelho** é o meu favorito (situação b)).

(2.11.) Naquelas terras havia **macieiras, pereiras e laranjeiras** e o Gonçalo tinha sempre **as árvores** bem tratadas (situação c)).

(2.12.) Dei tudo por **aquele animal.** Era **um gato especial** (situação d)).

(2.13.) **A cozinha** pegou fogo. **A casa** ficou irreconhecível (situação e)).

(2.14.) Não gosto muito de **legumes** só como **espinafres.** (situação f)).

A **Anáfora Verbal** é assim chamada porque é constituída por um sintagma verbal, SV, ou pelos seus constituintes. O seu antecedente tem também a forma de um SV. Veja-se:

(2.15.) A senhora **deu uma esmola ao pedinte** e assim **fez** a sua irmã.

Anáfora Adverbial

(2.16.) A família viaja **para o Algarve** todos os Verões e **lá** passam as suas férias.

[5]Hipónimos são palavras de sentido mais específico em relação ao de outro mais geral, em cuja classe está contido. Por exemplo, "gato" é um hipónimo de "animal".
[6]Holónimos são termos que representam um todo, de que os merónimos são as partes. É o caso em que "árvore" é holónimo de "tronco" e "tronco" é merónimo de "árvore".

Este exemplo é um caso em que a anáfora tem a forma de um advérbio de lugar, e o seu antecedente é a descrição de um ou mais acontecimentos. A anáfora pode ainda encontrar-se sob a forma de advérbio de tempo ou de modo.

A **Catáfora** é um caso particular de anáfora, em que o termo anafórico precede o antecedente (2.17.)

(2.17.) **Ela** chega sempre tarde porque trabalha longe de casa. **A Joana** é incansável.

Segundo [Mitkov (2002)], para além das anáforas pronominais, verbais, adverbiais e da catáfora, acima referidas, podem ainda ocorrer outros tipos de anáfora, os quais se passam a descrever.

A **Anáfora Definida,** é uma forma de anáfora nominal que ocorre sempre que uma descrição definida[7] é antecedida de: a) uma expressão com o mesmo núcleo e refere-se à mesma entidade no discurso, b) um núcleo diferente, mas que se refere à mesma entidade, c) um elemento não co-referente.

Veja-se um exemplo de cada situação.

(2.18.) Há **um filme** muito bom no cinema. **O filme** é sobre a Idade Média.

(2.19.) Planeavam assaltar **o banco,** mas **o edifício** estava vigiado.

(2.20.) Nós visitámos **um museu. As esculturas** eram fantásticas.

No caso do **Substantivo Anafórico,** o referente diz respeito ao núcleo do sintagma nominal e não ao sintagma completo (2.21.).

[7] A descrição definida constitui um grupo de palavras começado por um artigo definido e que tem um nome como núcleo. [Strawson (1950)]

(2.21.) Em vez de uma **caixa** de madeira, ele preferiu **uma** de cartão.

Anáfora Indirecta ou Associativa ocorre sempre que a referência ao antecedente é feita pelo leitor ou ouvinte indirectamente, sem que esteja explícito no discurso. Exige por vezes conhecimento adicional por parte do leitor / ouvinte. Veja-se o exemplo (2.22.).

(2.22.) **A Terra** é um planeta fantástico. **Os continentes** formaram-se ao longo de muitos milhões de anos.

Este tipo de anáfora indirecta ou associativa remete para o conceito de Anáfora Profunda, de [Hankamer & Sag (1976)], que é detectada apenas através de conhecimento semântico ou pragmático. Esta espécie de anáfora surge, segundo este autor, em oposição à Anáfora de Superfície, em que se diz que há uma relação de superfície entre o termo anafórico e o antecedente quando ela se detecta puramente a partir da estrutura sintáctica do texto.

A **Anáfora Vazia 0,** também chamada de **Elipse**, é invisível, pois não ocorre no texto na forma de palavras ou frases. Esta é uma representação sofisticada das anáforas, que visa reduzir a quantidade de informação sob forma abreviada. As formas mais comuns são: redução de pronomes, nomes e verbos (2.23.).

(2.23.) Das **roupas** que comprei, muitas ( ) estavam em saldos.

O conceito de Anáfora Nula ou Vazia foi originalmente criado por [Chomsky (1981)[8]]. Este tipo de anáfora é considerado, por este autor, como uma "categoria vazia" que não é uma simples ausência, porque uma ausência não pode possuir propriedades diferenciadas. Pelo contrário, é uma categoria linguística real com uma matriz

[8] Segundo a teoria de Chomsky, as várias línguas usadas pelos seres humanos através do mundo podem ser caracterizadas em termos da variação de conjuntos de parâmetros (como o parâmetro *pro-drop*, que estabelece se um sujeito explícito é obrigatório, como no caso da língua inglesa, ou se pode ser opcionalmente *deixado de lado* (suprimido ou elidido), como no caso do português, italiano e outras.

gramatical, embora sem matriz fonológica: Ou seja, o termo anafórico é foneticamente nulo, mas está presente na sua forma sintáctica e semântica.

De acordo com [Eckert & Strube (2000)], pode ainda considerar-se um outro tipo de anáfora, que a seguir se descreve.

A **Anáfora Deíctica** pode referenciar um acontecimento, um facto, uma ideia, um conceito ou uma acção, na forma de um predicado de uma oração, de uma oração, de um conjunto de orações, ou mesmo do texto como um todo. Veja-se o exemplo (2.24.).

(2.24.) **Era preciso tomar medidas** e o Manuel sabia-**o**.

Ainda segundo [Hirst (1981)], as anáforas podem ser classificadas de acordo com o número de frases envolvidas na sua resolução, constituindo as duas situações que a seguir se referem.

A **Anáfora Limitada**, também chamada de referência intrafrásica, ocorre quando o termo anafórico e o antecedente se encontram na mesma frase, como no caso dos pronomes reflexivos[9] (2.25.).

(2.25.) Naquele tempo, **os alunos** esforçavam-**se** mais.

A **Anáfora Discursiva** ou referência interfrásica ocorre sempre que o termo anafórico e o antecedente se encontram em frases diferentes (2.26.).

(2.26.) **Os amigos** estavam juntos. Porém, desta vez, **eles** não se entenderam e *todos* começaram a discutir.

[9] O pronome é reflexivo ou recíproco quando o objecto directo ou indirecto representa a mesma pessoa ou a mesma coisa que o sujeito do verbo. [Cunha & Cintra (1991)]

Expostos os diferentes tipos de anáforas, importa referir que para o estudo em questão, o interesse recai sobre as Anáforas Pronominais pessoais de terceira pessoa, como uma forma de referência textual anafórica.

2.1.2. Processo de Resolução de Anáforas

As anáforas são fenómenos referenciais de natureza bastante complexa e a sua interpretação é indispensável para a construção do sentido dos textos. A resolução de anáforas é uma questão crucial no Processamento da Linguagem Natural e constitui um dos principais temas de pesquisa para a linguística computacional. Do ponto de vista computacional, o problema da resolução anafórica põem-se quando existe mais do que um candidato a antecedente e é preciso aplicar um processo de decisão, para escolher um, dentre os candidatos possíveis.

Para a determinação de antecedentes de uma anáfora podem ser utilizadas muitas estratégias, individualmente, ou em conjunto, fazendo uso da informação linguística e cognitiva. A informação linguística é obtida principalmente através da análise morfológica e sintáctica, enquanto que a informação cognitiva se encontra ao nível de uma análise semântica e discursiva. Este conhecimento permite desenvolver algoritmos com estruturas de dados que representam as diferentes entidades mencionadas nos respectivos discursos. Na resolução de anáforas, essas listas de entidades são utilizadas como listas de potenciais candidatos a antecedentes de expressões anafóricas. Para uma melhor compreensão do processo de resolução de anáforas, passa-se a descrever os diferentes tipos de conhecimentos aí empregues.

2.1.2.1. Conhecimentos empregues na Resolução de Anáforas

O **Conhecimento Lexical e Morfológico** abrange o léxico / palavras e a estrutura da sua forma, fora do discurso em que estão inseridas.

A concordância de género e de número entre o termo anafórico e o antecedente pode, em muitos casos, ser suficiente para resolver a anáfora. Esta concordância pode

também ser útil para descartar hipóteses que morfologicamente não fazem sentido. Veja-se o seguinte exemplo:

(2.27.) **As casas** já estão pintadas. Mas se não fosse a Sandra e o José **elas** ainda estariam por pintar.

O exemplo demonstra ambas as situações acima referidas. Encontra-se concordância de género e número entre o pronome anafórico **elas** e o antecedente **As casas** e, ao mesmo tempo, descartam-se os candidatos a antecedente **Sandra** e **José**, pois não concordam em número com o termo anafórico.

A concordância em género e número é extremamente útil na resolução de anáforas pronominais, contudo existem excepções, veja-se:

(2.28.) **A mãe** foi jantar fora com **o pai. Eles** divertiram-se.

Neste exemplo, e em termos de análise morfológica, o pronome **Eles** não tem antecedente possível. Num algoritmo que exija concordância entre género e número, o antecedente deste pronome não será encontrado.

O **Conhecimento Sintáctico** é bastante importante para a resolução de anáforas, pois não só fornece informação sobre o que é, por exemplo, um sintagma nominal, um pronome ou um verbo, como também faz a divisão da frase em orações. Estas informações permitem estabelecer regras que determinam se um elemento é sintacticamente compatível com o termo anafórico ou se é eliminado como um possível antecedente. Veja-se o exemplo:

(2.29.) **Os pássaros** levantaram voo. Os meninos viram-**nos.**

Os sintagmas nominais são, por natureza, sintacticamente compatíveis com os pronomes. Neste sentido, o pronome –**nos** pode ser resolvido com um dos sintagmas nominais: **Os meninos** ou **Os pássaros**. No entanto, se tivermos em consideração a regra que estabelece que um pronome não-reflexivo e um sintagma nominal, na mesma

oração, não podem ser co-referentes, o pronome em questão só terá uma única solução: **Os pássaros.**

A partir do nível de **Conhecimento Semântico** torna-se cada vez mais complexo o tratamento computacional, dada a natureza das informações tratadas e a ambiguidade inerente à língua natural. O conhecimento semântico contribui para o processo de resolução de anáforas através de informações disponíveis em dicionários ou ontologias[10]. Os termos são analisados de acordo com os seus traços semânticos.

Quando a análise morfológica, lexical e sintáctica não resolve directamente a anáfora, a componente semântica pode ser muito útil neste sentido. Veja-se o exemplo em (2.30.).

(2.30.) **O cão** estava no pátio, quando viu o osso. **Ele** comeu o petisco com satisfação.

Se ao processo de resolução for adicionada informação semântica, que explicite que a acção de **comer** está associada a uma entidade animada, o pronome **Ele** só pode ser resolvido com o sintagma nominal **O cão** e não com o sintagma nominal **o osso.**

O **Conhecimento de Discurso** pode ser útil, no caso em que existam vários candidatos a antecedente de um termo anafórico e os conhecimentos anteriores não dão a melhor resposta. Este pode acrescentar conhecimento, que prefere uma das hipóteses em relação às outras. Essa preferência pode ser baseada na coesão[11] do discurso. Isto é, se num dado momento do discurso, o tema / assunto for um determinado e corresponder a um dos candidatos possíveis a antecedente, esse será o escolhido.

Analisemos o exemplo:

[10] Ontologia é uma especificação formal e explícita de um fenómeno do mundo real de conhecimento consensual. É formal porque é utilizada computacionalmente. É explícita, porque diz respeito a conceitos, propriedades, relações, funções, restrições e axiomas que são explicitamente definidos. [Borst (1997)].

[11] Coesão é a manifestação linguística da coerência, advém da maneira como os conceitos e relações subjacentes são expressos na superfície textual.

(2.31.) **A Paula** assistiu a um jogo de futebol... **Paula** levou a amiga a casa. **Ela** ficou cansada..

O pronome **Ela** pode ser resolvido com qualquer dos candidatos a antecedente: **casa, amiga** e **Paula**. A escolha recai no antecedente **Paula**, pois este constitui o tema actual do discurso.

Em forma de conclusão e relativamente aos vários tipos de conhecimento empregues no processo de resolução de anáforas, considera-se que o conhecimento lexical e morfológico e o conhecimento sintáctico são do tipo restritivo, excluem elementos incompatíveis. Os restantes conhecimentos, semântico e de discurso, podem ser classificados como preferenciais, isto é, permitem enunciar regras que preterem uns elementos em benefício de outros.

2.1.2.2. Diferentes Abordagens

Têm sido propostas diferentes estratégias para a resolução de anáforas, as quais, segundo [Deoskar (2004)], se podem classificar em duas grandes categorias: *Knowledge-Rich* e *Knowledge-Poor*, conforme a complexidade ou simplicidade do tipo de conhecimento (sintáctico, semântico e de discurso), empregue no algoritmo de resolução.

As **Abordagens baseadas na Sintaxe (*Knowledge-Rich*)** pressupõem a existência de uma árvore sintáctica que é pesquisada, com o intuito de descobrir antecedentes, com base em restrições sintácticas e morfológicas. Um exemplo clássico desta abordagem é a de [Hobbs (1978)].

O algoritmo de Hobbs é bastante simples. Em primeiro lugar, executa uma procura, *em-largura-primeiro* (*breadth-first*), da esquerda para a direita, (cada nó de profundidade N é visitado primeiro que qualquer nó de profundidade N+1), na árvore sintáctica da frase onde ocorre o pronome; dando preferência ao candidato a antecedente mais próximo. Caso não seja encontrado nenhum candidato, o processo é repetido em frases anteriores, preferindo sempre as mais próximas.

O algoritmo selecciona possíveis antecedentes e verifica se, para cada um deles, há concordância de género e de número com o pronome. Este algoritmo também tem em atenção restrições sintácticas, dentre as quais as mais importantes são as seguintes:

- um pronome não-reflexivo e o seu antecedente não podem ocorrer na mesma oração;
- o antecedente do pronome deve preceder o pronome.

Inicialmente, este algoritmo foi aplicado em três textos diferentes, de língua inglesa, onde os pronomes resolvidos eram: *he*, *she*, *it* e *they*. Dos trezentos pronomes encontrados, ele resolveu com sucesso 88,3% dos casos. Quando a este a algoritmo se adicionaram algumas restrições de selecção, conseguiu-se atingir uma taxa de sucesso de 91,7%. Contudo, estes resultados foram obtidos nos casos em que não havia conflitos, isto é, em que não existia mais do que um candidato a antecedente. No caso contrário, em que foram encontrados mais candidatos a antecedente, a taxa de sucesso foi de 81,8%.

Por se tratar de um algoritmo bastante simples, que apenas contempla conhecimento de tipo morfo-sintáctico, este é tido em menor consideração. No entanto, consegue competir com outros mais recentes.

As **Abordagens baseadas no Discurso (*Knowledge-Rich*)** utilizam conhecimento do discurso para poderem escolher o melhor candidato a antecedente. Um exemplo clássico desta abordagem é o *Centering*. Foi introduzida por [Grosz *et al* (1986)] e depois estendida por [Brennan *et al* (1987)]. Estes últimos consideram a teoria de *Centering* como um sistema de regras e restrições que governam as relações entre o tema e o discurso e algumas escolhas linguísticas efectuadas pelos participantes do discurso, como o emprego de pronomes. O *Centering* permite determinar o elemento dominante do discurso, a cada momento e prevê que este se vá modificando ou não (preferindo sempre a manutenção).

Esta é uma proposta especialmente útil para a resolução de referências interfrásicas, nas quais se torna necessário um critério limitador para a quantidade de candidatos a considerar.

O resultado obtido com o algoritmo de Brennan, Friedman e Pollard, em comparação com o algoritmo de Hobbs, mostra que ambos respondem de forma semelhante, ao caso de cem frases pertencentes a um texto ficcional, de língua inglesa. Contudo, o algoritmo de Hobbs obteve uma melhor performance (89% contra 79%), num conjunto de textos jornalísticos [Tetreault (2001)].

Esta abordagem será descrita com mais detalhe no capítulo 3 (Abordagem Proposta).

Abordagens Híbridas (*Knowledge-Rich*) são abordagens que fazem uso de vários tipos de conhecimento para ordenar os possíveis candidatos a antecedente. Um dos mais conhecidos algoritmos que tem como base esta abordagem é o de [Lappin & Leass (1994)], designado por RAP (*Resolution of Anaphora Procedure*). Visa resolver anáforas inter e intrafrásicas pronominais de terceira pessoa, na língua inglesa. Este modelo pressupõe o cálculo da saliência de elementos no discurso, com base na estrutura sintáctica de cada frase e numa representação simples do discurso. O algoritmo é composto por várias etapas, que se passam a descrever:

- um filtro sintáctico, que actua dentro de cada frase e permite descartar dependências sintácticas entre pronomes e sintagmas nominais. [Lappinn & McCord (1990a)];
- um filtro morfológico que ignora sintagmas nominais que não concordem em pessoa, género e número com o pronome;
- um procedimento para encontrar pronomes pleonásticos[12];
- um algoritmo de ligação que identifique os possíveis candidatos a antecedente de um pronome reflexivo. [Lappinn & McCord (1990b)];
- um procedimento para atribuição de valores a vários factores de saliência de um sintagma nominal (função sintáctica, paralelismo sintáctico, frequência e proximidade);

[12] Pronomes pleonásticos são pronomes que se usam unicamente para enfatizar a entidade anteriormente referida.

- um procedimento para identificar sintagmas nominais, com ligações entre si, que formam uma classe de equivalência. Esta classe tem como valor de saliência a soma dos valores de cada elemento;
- um procedimento de decisão, que escolhe o candidato a antecedente do pronome melhor classificado.

Em síntese, o algoritmo começa por extrair todos os possíveis candidatos a antecedente e elimina aqueles que sejam morfológica e sintacticamente incompatíveis. Aos restantes candidatos é atribuída uma pontuação de saliência, de acordo com a sua função gramatical, que permite a escolha do melhor candidato a antecedente.

No estudo original, este algoritmo obteve uma taxa de sucesso de 86%, num corpus composto por manuais de computadores, em língua inglesa. Esta taxa de sucesso encontra-se ao nível dos algoritmos atrás referidos.

As **Abordagens baseadas no Corpus (*Knowledge-Rich*)** são o mais possível desprovidas de pressupostos teóricos e associam outros tipos de conhecimento aos elementos do corpus. Um exemplo é a proposta de [Ge *et al* (1998)], que ao algoritmo de Hobbs acrescenta um modelo probabilístico para a resolução de anáforas pronominais. Este modelo foi obtido através da utilização de um corpus de treino, constituído por um pequeno conjunto de textos jornalísticos na língua inglesa, que já continha a identificação e resolução de anáforas. Para a criação do referido modelo foram contemplados os seguintes tipos de informação:

- distância entre o pronome e o antecedente proposto;
- restrições sintácticas do algoritmo de Hobbs;
- restrições de número, género e de estado animado / inanimado do antecedente proposto;
- interacção entre o nó acima do pronome (da árvore sintáctica) e o antecedente proposto;
- número de vezes que o antecedente proposto é referido.

Neste estudo, calcularam-se as probabilidades na resolução de anáforas, no corpus de treino, tendo sido depois usadas para resolver as anáforas no corpus de teste. Foram identificados 2477 pronomes pessoais (*he*, *she* e *it*), e obteve-se uma taxa de sucesso de 82,9%.

Estes autores investigaram ainda a importância relativa de cada um dos tipos de informação enunciados anteriormente. No caso em que consideraram apenas os dois primeiros tipos de informação, acima referidos, obtiveram uma taxa de sucesso de 65,3%. Depois de acrescentarem informação relativa ao género e ao estado animado / inanimado, a taxa subiu para 75,7%. Acrescentando ainda a estes tipos de informação aquele que diz respeito à interacção entre o pronome e o antecedente, a taxa aumentou apenas 2,2%. Finalmente, com a utilização da informação sobre o número de ocorrências do antecedente proposto, atingiu-se o valor já referido de 82,9%.

Esta taxa de sucesso encontra-se também ao nível dos algoritmos atrás referidos.

As **Abordagens *Knowledge-Poor*** surgiram da necessidade de evitar sistemas de conhecimento complexos empregues no algoritmo de resolução, que exigem um trabalho intensivo e consomem muito tempo.

Um dos casos de abordagem *Knowledge-Poor* é o algoritmo robusto de [Mitkov (2002)]. Este algoritmo de resolução de anáforas pronominais aplica-se após uma fase de pré-processamento, na qual se faz a identificação das categorias gramaticais das palavras, dos sintagmas nominais e das frases. O algoritmo é constituído por três etapas, que se passam a descrever.

- Encontram-se os sintagmas nominais, seguindo a ordem da direita para a esquerda, a partir da posição do pronome, até ao limite de duas frases imediatamente anteriores.
- Cria-se um grupo de candidatos a antecedente, a partir dos sintagmas nominais anteriormente encontrados, compatíveis em género e em número com o pronome.
- Atribui-se a cada um dos candidatos várias pontuações, de acordo com indicadores de preferência (*antecedent indicators*) e escolhe-se o candidato melhor classificado.

Os indicadores de preferência utilizados neste tipo de resolução são os seguintes:

- **Primeiro sintagma nominal** – o primeiro sintagma nominal (SN) da frase tem uma pontuação +1;
- **Verbos de indicação** – SNs que se seguem imediatamente a um grupo predefinido de verbos (verbos de indicação) têm uma pontuação +1;
- **Reiteração lexical** – SNs que ocorrem repetidos num parágrafo têm uma pontuação de +1, se se repetem uma vez, ou uma pontuação de +2, se se repetem mais vezes.
- **Preferência por secção de tópico** – SNs que ocorrem na mesma secção (de um tópico) do pronome, tem uma pontuação de +1.
- **Padrão de colocação** – SNs com o mesmo padrão de colocação do pronome, tem uma pontuação de +2.
- **Referência imediata** – SNs que ocorrem na seguinte construção: "... *(You) V1 SN ... con (you) V2 it (con (you) V3 it)*", onde *con* é um dos seguintes {*and / or / before / after*}, têm a pontuação de +2.
- **Instruções sequenciais** – SNs que ocorrem na posição SN1 numa construção do tipo: "...*To V1 SN1, V2 SN2. (Sentence). To V3 it, V4 SN4*", são os antecedentes mais prováveis do pronome *it*.
- **Preferência por termos específicos** – SNs identificados como termos específicos da área temática abordada têm uma pontuação +1.
- **Indefinição** – SNs que são indefinidos têm uma pontuação de -1.
- **Sintagmas nominais não preposicionais** – SNs que pertencem a sintagmas preposicionais têm uma pontuação de -1.
- **Distância referencial** – SNs que ocorram na oração anterior da mesma frase têm pontuação +2; na frase anterior têm a pontuação +1; na segunda frase anterior não têm pontuação. Nas restantes frases têm a pontuação -1.

Em caso de empate, utiliza-se a seguinte ordem de indicadores com a melhor pontuação: **Referência imediata, Padrão de colocação, Verbos de indicação** e **Distância referencial.**

O algoritmo de Mitkov obteve uma taxa de sucesso de 89,7%, encontrando-se ao nível dos outros atrás mencionados, baseados em abordagens *Knowledge-Rich.*

Como conclusão, importa referir que as abordagens *Knowledge-Rich* fazem, habitualmente, um pré-processamento manual dos dados, o que influencia positivamente as suas taxas de sucesso. Em oposição, as abordagens *Knowledge-Poor* integram, na sua maioria, um pré-processamento automático. No entanto, no caso do algoritmo robusto de [Mitkov (2002)], o resultado do pré-processamento é corrigido manualmente, de modo a que se possa comparar, de uma forma justa, o sucesso das diferentes abordagens.

2.1.2.3. Evolução e Tendências

Numa perspectiva evolutiva dos processos de resolução de anáforas, as primeiras abordagens [Sidner (1979)], [Brennan *et al* (1987)] e [Lappin & Leass (1994)], entre outras, baseavam-se no conhecimento da linguagem, o que era difícil de representar e de processar. Perante este problema surgiram estratégias *Knowledge-poor*, como é o caso da abordagem de [Baldwin (1997)], ou mesmo do algoritmo de [Mitkov (1998)], em que o algoritmo de resolução de pronomes dispensa conhecimentos complexos de sintaxe, semântica e análise de discurso. Para estes sistemas de resolução de anáforas pronominais é essencial a eficácia do pré-processamento, o qual constitui actualmente um problema importante por resolver, e para o qual se procuram alternativas, por exemplo, recorrendo à *web* [Market *et al* (2003)], ou tentando outras abordagens, que utilizam técnicas de aprendizagem automática.

Deve ainda salientar-se que a maior parte dos estudos realizados nesta área foram feitos em língua inglesa, francesa, japonesa e espanhola, pelo que poucos se têm dedicado à língua portuguesa. Destes poucos refira-se os casos de [Aires *et al* (2004)] que fez a avaliação da teoria de *Centering* segundo [Brennan *et al* (1987)], na resolução

de anáforas pronominais pessoais, em textos jurídicos, com uma taxa de sucesso de 51%; e [Coelho (2005)] que fez a avaliação do algoritmo de Lappin e Leass, na resolução de anáforas pronominais pessoais, incluindo pronomes reflexivos / recíprocos, nos mesmos textos, com uma taxa de sucesso de 43,56%. O presente trabalho constitui, por este motivo, mais um contributo para o estudo da resolução de anáforas pronominais em língua portuguesa.

Hoje em dia os sistemas de resolução de anáforas, sejam elas pronominais ou não, são usados em muitos domínios, como por exemplo em interfaces de língua natural, geração automática de resumos, tradução automática, recuperação de informação, entre outros.

2.2. Recuperação de Informação (RI)

«*O objectivo preliminar de um sistema de recuperação de informação é recuperar todos os documentos que são relevantes, como resposta a uma solicitação do utilizador, com uma quantidade mínima de documentos não-relevantes.*» [Baeza-Yates & Ribeiro-Neto (1999)].

O termo recuperação de informação (*information retrieval*) foi referido pela primeira vez por Calvin Mooers[13], na sua dissertação de mestrado, em 1948, passando a ser utilizado na literatura científica, a partir de 1950. Desde essa altura, a RI tornou-se uma área de pesquisa autónoma interdisciplinar, com acelerado desenvolvimento e com grande ligação às ciências computacionais.

A partir da segunda metade do século XX, a grande quantidade de informação produzida no desempenho das inúmeras actividades humanas é, cada vez mais, editada em formato digital, constituindo vastos repositórios de informação (acessíveis através de redes e sistemas de computadores). Por outro lado, a necessidade de encontrar informação relevante faz, desde logo, ressaltar a importância de sistemas e ferramentas eficientes para a recuperação de informação, criada continuamente a ritmos vertiginosos. Estas duas constatações explicam a razão pela qual a RI tem sido, cada vez mais, tema de estudo e investigação.

Os sistemas de recuperação de informação (SRI) são classificados de acordo com o tipo de informação (documento) que manipulam, o qual pode ser textual, visual, áudio e multimédia. O principal objectivo de um SRI é recuperar informação (contida nos documentos) que possa ser útil ou relevante para o utilizador. Tal informação (de interesse do utilizador) é designada necessidade de informação. Para obter documentos do seu interesse, o utilizador tem de traduzir uma necessidade de informação numa consulta. Na sua forma mais comum, esta consulta é um conjunto de palavras-chave que são usadas para recuperar documentos, a partir de uma vasta colecção. A presença de documentos não relevantes entre os documentos recuperados por uma consulta é praticamente certa, no entanto os SRIs devem a todo o custo minimizar a sua presença.

---

[13] Inicialmente, Calvin Mooers apresentou a seguinte definição: «*A RI trata dos aspectos intelectuais de descrição da informação e sua especificação para a pesquisa, e também de qualquer sistema, técnicas ou máquinas que são empregues na realização desta operação.*»

Uma forma simples de obter rapidamente um conjunto de respostas, para uma consulta do utilizador, seria determinar quais os documentos de uma colecção que contêm as palavras da consulta. No entanto, isto certamente não seria suficiente para satisfazer a necessidade de informação do utilizador. Para que esta tarefa seja feita de forma mais eficiente, os SRIs ordenam os documentos de acordo com o grau de relevância em relação ao expresso na consulta do utilizador.

A noção de relevância é um conceito fundamental em recuperação de informação e é um elemento chave para classificar / ordenar os documentos recuperados.

O presente estudo aborda apenas a RI de tipo textual, em que a informação é descrita em linguagem natural, e a palavra é a menor unidade de análise.

2.2.1. Processo de Recuperação de Informação

O processo típico de recuperação de informação passa pelas seguintes etapas:

- criação de uma representação dos documentos;
- interpretação da consulta;
- pesquisa de documentos;
- apresentação dos resultados.

Uma representação simplificada do processo de recuperação de informação é apresentada na seguinte figura.

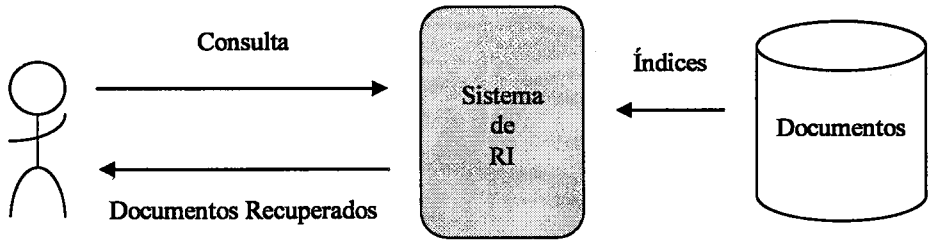

Figura 1 – Processo de recuperação de informação.

A etapa de **criação da representação dos documentos** consiste essencialmente num processo de indexação (manual ou automático), precedido ou não de um pré-processamento. Nesta fase é analisado o conteúdo de cada documento e extraídos conceitos que o identificam e que definem pontos de referência para a consulta. Em 2.2.1.2. serão descritas as principais técnicas de indexação.

A análise do documento pode envolver a interpretação do seu conteúdo e permite a abstracção de assuntos que não estão directamente explícitos na superfície textual. No entanto, a automatização do processo de indexação só é possível através de uma simplificação, em que se considera que os assuntos só podem ser extraídos a partir da sua estrutura textual. Neste processo, nem todas as palavras do documento têm a mesma importância, por isso, se se tiver em consideração a totalidade das palavras do documento (sistemas de texto integral) gera-se muito "ruído" e dificulta-se a tarefa de recuperação. Para reduzir esse "ruído" faz-se um pré-processamento dos documentos, com o objectivo de diminuir o conjunto de palavras usadas na representação (o que leva também a um melhor desempenho destes sistemas). Em 2.2.1.1. será descrito com mais pormenor o pré-processamento.

A saída final do pré-processamento (dos documentos) será um conjunto de classes de palavras, uma para cada radical (*stem*) detectado. A representação dos documentos será então uma lista de termos, também denominada de índice. Por outras palavras, um documento será indexado por uma determinada classe, se nesse documento ocorrerem palavras consideradas relevantes que pertencem a essa classe.

Existem ainda outras formas de melhorar o desempenho dos SRIs, como é o caso da utilização de *Thesaurus*[14] que define relacionamentos conceptuais entre palavras (por exemplo: sinónimos) e permite aumentar as hipóteses de recuperar documentos relevantes. Outra técnica, o *Clustering*, torna o processo de recuperação mais rápido, pois permite *a priori* o agrupamento de documentos que, de alguma forma, se relacionam entre si.

A etapa de **interpretação da consulta** consiste em analisar a pergunta do utilizador e fazer a sua interpretação. Para se obter uma consulta que possa ser processada, o sistema pode também utilizar as técnicas de pré-processamento da etapa

[14] O termo *Thesaurus* teve origem no ano de 1852 com a publicação do dicionário de Peter Mark Roget, intitulado "*Thesaurus of English words and phrases*". Nesse dicionário, as palavras não foram dispostas por ordem alfabética, mas antes de acordo com as ideias que exprimem.

anterior. Pode-se também utilizar um *Thesaurus* para substituir as palavras-chave de uma consulta quando a consulta original não retornar resultados ou retornou poucos; ou ainda, gerar várias consultas semelhantes quando se quer mais resultados ou mais precisos.

Na etapa de **pesquisa de documentos** obtêm-se os documentos relevantes a que a consulta faz referência. Faz-se a comparação do índice da consulta, resultantes da interpretação, com o índice dos documentos ou, em alternativa, com os perfis dos *clusters* de documentos.

Na última etapa, de **apresentação de resultados,** é disponibilizado ao utilizador, um conjunto de citações de documentos relevantes para a consulta efectuada. Os resultados apresentados podem ou não estar ordenados pela sua relevância, já que alguns SRIs não integram o seu cálculo.

2.2.1.1. Pré-processamento (Filtros)

No pré-processamento podem utilizar-se uma série de filtros que permitem a redução do tamanho dos documentos e a sua normalização, simplificando desta forma a consulta. Segundo [Frakes & Baeza-Yates (1992)], as operações de filtragem mais comuns são as que a seguir se enunciam:

- remoção de palavras utilizando uma lista de *stopwords* – é uma lista formada por palavras consideradas irrelevantes para a tarefa de recuperação, como, por exemplo, as preposições;
- transformação de maiúsculas em minúsculas;
- remoção de caracteres especiais e redução de sequências de espaços em branco para um só;
- normalização do formato de datas e números;
- extracção automática de palavras-chave;

- *ranking* de palavras – selecção de palavras baseada no contexto e no formato do documento;
- *stemming* de palavras – redução das palavras ao seu radical (*stem*), por eliminação de prefixos e sufixos.

Após a aplicação dos filtros, o texto original será transformado num conjunto de classes (termos), uma para cada radical (*stem*) detectado.

2.2.1.2. Técnicas de Indexação

De acordo com [Baeza-Yates & Ribeiro-Neto (1999)], as principais técnicas de indexação são: o Ficheiro Invertido, o Ficheiro de Assinaturas e a Árvore de Sufixos.

O **Ficheiro Invertido** é o exemplo mais comum de um índice lexicográfico (índice ordenado). É construído a partir de todos os termos representativos dos documentos, que opcionalmente podem também ter associado uma frequência ou um peso de relevância. Este índice é, portanto, uma lista ordenada de termos, em que cada um deles tem ligações aos documentos a que dizem respeito. A lista é invertida, pois em vez de ser ordenada pela localização no documento, passa a ser ordenada (alfabeticamente) pelo termo.

Este tipo de indexação é encontrado na maior parte dos sistemas comerciais e uma das suas características consiste na rápida pesquisa dos termos. Contudo, apresenta desvantagens, em relação ao espaço ocupado, que pode chegar aos 300% dos documentos originais; e também um alto custo para a actualização do índice.

O **Ficheiro de Assinaturas** é um exemplo de um índice baseado em *hashing*. Neste método, cada documento é representado por uma sequência de blocos de bits de tamanho fixo, que é referida como a assinatura do documento. Para se conseguir esta assinatura, aplica-se uma técnica de *hashing* a cada termo representativo do documento, obtendo-se assim, a respectiva assinatura; depois divide-se as assinaturas dos termos em blocos e é aplicada uma técnica de superimposição de código (geralmente uma função do tipo OR) para obter a assinatura de cada bloco. As assinaturas de todos os

documentos são armazenadas sequencialmente num ficheiro – o ficheiro de assinaturas – que é muito menor que o conjunto original de documentos. As comparações entre consultas e documentos são feitas através deste ficheiro de assinaturas.

Este tipo de indexação tem as vantagens de ser capaz de tolerar erros de tipos e grafia e de ocupar só 10% a 20% do espaço ocupado pelos documentos originais. No entanto, por ser sequencial é mais lento na pesquisa de termos que o método do Ficheiro Invertido.

A **Árvore de Sufixos** (*PAT Tree* ou árvore de posições) é uma árvore n-ária de posições que vê o documento como uma longa *string*. Cada posição no texto corresponde a uma *substring* (sufixo) que começa na posição indicada e termina arbitrariamente em direcção ao fim do texto. Este índice é indicado para consultas complexas ao nível da frase.

É um tipo de indexação em que a pesquisa de excertos de textos é bastante mais rápida, mas por outro lado, também é mais difícil de construir e de manter que os anteriores.

Atendendo a todo o seu processo, um SRI pode adoptar um modelo ou várias características de um ou mais modelos. O modelo de RI determina as especificações da representação dos documentos e da consulta e define como estes são comparados. Os principais modelos utilizados em SRIs serão a seguir descritos.

2.2.1.3. Modelos de RI

Os modelos de RI são modelos conceptuais ou abordagens genéricas para a recuperação de informação. Apesar de, inicialmente, terem sido desenvolvidos apenas para documentos textuais, estes modelos podem ser utilizados em qualquer tipo de documento.

Segundo [Baeza-Yates & Ribeiro-Neto (1999)], existem 12 modelos de recuperação de informação. Os modelos de RI podem dividir-se em modelos clássicos e modelos estruturados. Nos modelos clássicos, cada documento é descrito por um conjunto de termos representativos - também chamados de termos de indexação - que tenta representar o assunto do documento e sumariar o conteúdo de forma significativa.

Nos modelos estruturados, podem especificar-se, para além dos termos representativos, algumas informações acerca da estrutura do documento (como secções a serem pesquisadas, fontes de letras, proximidade das palavras, etc.).

Os modelos clássicos são três: o modelo Booleano [Paice (1984)], o modelo Vectorial [Salton & McGill (1983)] e o modelo Probabilístico [Maron & Kuhns (1960)]. Para cada um deles, há modelos alternativos que os estendem. De seguida passam-se a explicar os referidos modelos.

O modelo **Booleano** é baseado na teoria dos conjuntos, é um modelo simples, mas considerado o mais ineficiente. Para cada consulta, são recuperados os documentos que possuem os termos nas condições especificadas pelo utilizador. A consulta pode integrar operadores booleanos[15] *or*, *and* e *not* para estabelecer relações específicas de ocorrência entre os termos. A maior desvantagem deste modelo reside na sua rigidez binária, ou seja, os documentos são analisados sob o critério dualista: relevante / não relevante (não há valores intermédios), o que não permite a ordenação dos resultados. Existem alguns modelos alternativos ao Booleano, que a seguir se apresentam.

- O modelo de **Lógica Difusa (Fuzzy)** estende o conceito da representação dos documentos por termos, assumindo que cada consulta determina um conjunto difuso e que cada documento possui um certo grau de pertença a esse conjunto, usualmente menor do que 1.
- O modelo **Booleano Estendido** procura ultrapassar o problema das decisões binárias do modelo clássico através da atribuição de pesos aos termos índice, aproximando o modelo original do modelo Vectorial.

O modelo **Vectorial** trata os documentos como "sacos de palavras" (*bags of words*), os quais são representados como vectores num espaço de dimensão *n*, onde *n* é o total de termos índices de todos os documentos no sistema. Neste modelo, que não é binário, pode-se calcular um grau de semelhança que os documentos devem ter para poderem ser considerados relevantes e que simultaneamente os permite ordenar

[15] Estes operadores receberam o nome de George Boole, matemático inglês, que foi o primeiro a definilos como parte de um sistema de lógica, em meados do século XIX.

(*ranking*). O modelo Vectorial é a base da grande maioria dos SRIs, dos quais se destacam aqueles que funcionam na Internet, embora estes também utilizem outras técnicas para determinar o *ranking* de documentos. De seguida serão apresentados alguns modelos que propõem estender a funcionalidade do modelo Vectorial.

- O modelo **Vectorial Generalizado** questiona a independência dos termos índices, assumida nos modelos clássicos, e abre a possibilidade de se considerar que certos conceitos, representados por estes termos, estejam relacionados.
- O modelo de **Indexação Semântica Latente** questiona a relevância das palavras-chave como candidatos a termos de indexação e tenta estabelecer relações conceptuais entre documentos e consultas.
- O modelo de **Redes Neuronais** utiliza as redes neuronais para estabelecer padrões de relacionamento entre as consultas e o conjunto de documentos. Cada consulta "dispara" um sinal que activa os termos índice, que por sua vez propagam os sinais aos documentos relacionados. Estes, por sua vez, devolvem os sinais a novos termos índices, em interacções sucessivas. O conjunto resposta é definido através desse processo e pode conter documentos que não compartilhem nenhum termo índice com a consulta, mas que tenham sido activados durante o processo.

O modelo **Probabilístico** pressupõe a existência de um conjunto ideal de documentos, que satisfaz cada uma das consultas. Este conjunto pode ser recuperado através de uma tentativa inicial, com um conjunto de documentos e refinado com o *feedback* do utilizador (documentos considerados relevantes), em sucessivas interacções. Os modelos que procuram estender o modelo Probabilístico são referidos de seguida.

- O modelo de **Redes de Inferência** associa variáveis aleatórias ao evento de resposta de um certo documento a uma determinada consulta. Essas variáveis podem ser alteradas à medida que novos eventos são observados.

- O modelo de **Redes de Crença**, semelhante ao anterior, faz uma tradução dos documentos e das consultas em subconjuntos de um espaço de conceitos. A cada documento, associa-se a probabilidade de que o mesmo cubra os conceitos presentes no espaço de conceitos. Cada consulta é mapeada no espaço de conceitos que, por sua vez, está ligado ao espaço de documentos.

2.2.1.4. Evolução e Tendências

Nos anos 50, uma massa crítica de cientistas, engenheiros e empreendedores começaram entusiasticamente a trabalhar o problema e a solução da recuperação de informação. Nos anos seguintes, esse trabalho tornou-se uma actividade relativamente ampla e organizada, que deu origem a debates estimulantes e a acalorada argumentação acerca das melhores e mais adequadas soluções. Nos anos 60, introduziram-se os modelos clássicos de RI, foram efectuadas muitas experiências e começaram a utilizar-se métricas para a avaliação de SRIs. Na década de 70, surgiram os primeiros processadores de texto e muitos documentos começaram a ficar disponíveis em formato digital. Foi na década de 90 que se deu um grande desenvolvimento nesta área, pois as tecnologias passaram da fase experimental para a fase de utilização. Em 1992, o Departamento de Defesa dos Estados Unidos, em conjunto com o Instituto Nacional de Padrões e Tecnologia, do mesmo país, patrocinou a primeira edição da *Text Retrieval Conference* (TREC), com o objectivo de reunir a comunidade mundial de recuperação de informação em torno da avaliação das várias metodologias desta área. Também no final desta década, surgiram os primeiros motores de pesquisa *web*. Estes SRIs tiveram necessidade de rever e melhorar o processo de recuperação de informação convencional, devido às características da informação disponível na *web* (distribuída, em grande volume, não estruturada e heterogénea) e aos diferentes perfis dos seus utilizadores. Neste sentido, destacou-se o motor de pesquisa *Google,* que apostou na cobertura alargada da informação presente na *web* e na sua acessibilidade universal, na rapidez de execução (através de um processamento distribuído), na simplicidade das suas interfaces e sobretudo no retorno de resultados relevantes. Este último aspecto, a determinação da relevância dos documentos, foi elemento essencial para o sucesso alcançado. O cálculo da relevância é feito em função de cinco critérios: i) *pagerank* - pontuação dada a uma página em função do número de hiperligações de outras páginas

que lhe fazem referência; ii) coincidência textual entre a pergunta e partes do documento; iii) medidas de proximidade - cálculo de semelhança entre as palavras da pergunta e do documento; iv) ordem dos termos da pergunta (pressupõe-se que os termos são introduzidos por ordem decrescente de importância); e v) características visuais de apresentação (pressupõe-se também que palavras com determinadas formatações são mais importantes que outras) [Brin & Paige (1998)].

Actualmente os SRIs mais usados são os motores de pesquisa para a *web*, dos quais resultam diversas frentes de pesquisa, em que se tentam melhorar as técnicas de recolha de documentos, para manter uma colecção actualizada. Verifica-se também a necessidade de interpretar os diversos tipos de documentos disponíveis em vários formatos. Outro problema reside na necessidade de lidar, de forma eficiente, com o grande volume de informação, não só em termos de criação de índices, mas também em termos de ordenação de resultados. A qualidade questionável do conteúdo é também uma questão que se tenta resolver. Por último, refira-se a necessidade de lidar com um conjunto diversificado de utilizadores, nomeadamente na disponibilização de *interfaces* simples para utilizadores não-especialistas; na recuperação de erros e identificação de consultas ambíguas.

Perante este quadro, a RI apresenta continuamente novos desafios e configura-se como uma área de estudo cada vez mais importante.

3. Abordagem Proposta

O objectivo deste estudo é verificar o impacto da resolução de anáforas pronominais nos sistemas de recuperação de informação. Neste sentido, foram escolhidos um método de resolução de anáforas pronominais e dois sistemas de recuperação de informação.

Numa primeira fase, é implementada a metodologia de *Centering* original [Grosz *et al* (1986)], para a resolução de anáforas pronominais pessoais de terceira pessoa, em textos de língua portuguesa (descrita em 3.1.) e, posteriormente, é avaliada a influência dessa resolução nas respostas fornecidas pelos sistemas SENTA[16] (*Software for the Extraction of N-ary Textual Associations*) e *Google Desktop* (descritos em 3.2).

A seguir descrever-se-ão cada um dos três sistemas atrás referidos.

3.1. Aplicação do *Centering* à Resolução de Anáforas Pronominais

A teoria de *Centering* foi inicialmente proposta por [Grosz *et al* (1986)] e posteriormente revista em [Grosz *et al* (1995)], com o objectivo de modelar a coerência local de um discurso. Entenda-se como coerência num discurso a forma como as entidades são introduzidas e se relacionam entre si de modo coerente. A coerência é local quando é limitada a um segmento do discurso, ou seja, limitada a um grupo de enunciados[17] sequenciais (que constituem esse segmento).

A teoria de *Centering* baseia-se também na noção de saliência que permite determinar o elemento dominante de cada enunciado do discurso, através da preferência de uma entidade em detrimento de outras, de acordo com as suas características gramaticais.

Segundo [Brennan *et al* (1987)], o *Centering* estabelece um conjunto de regras e restrições que regem as relações entre o tema e o discurso e algumas escolhas linguísticas efectuadas pelos participantes do discurso, como o emprego de pronomes. A aplicação do *Centering* na resolução de pronomes anafóricos é assim justificada, uma vez que a pronominalização visa centrar a atenção sobre aquilo que se diz. Assim,

[16] Sistema On-Line disponível em http://senta.di.ubi.pt

[17] Considera-se neste trabalho que um enunciado seja um fragmento textual delimitado por ponto (.), ponto de exclamação (!), ponto de interrogação (?), ponto-e-vírgula (;), reticências (...) ou dois pontos (:).

perante a ocorrência de um pronome, o tema actual ou os anteriores são candidatos altamente prováveis a termo antecedente.

De acordo com esta metodologia e do ponto de vista estrutural, um discurso é constituído por uma sequência de segmentos de discurso. Por sua vez um segmento do discurso é composto por uma sequência de enunciados ou *Utterances* (U_1, U_2, ..., U_m). Os enunciados possuem centros que são entidades que ligam enunciados dentro do mesmo segmento. Os centros de um enunciado U_n podem ser classificados da seguinte forma:

- $Cf(U_n)$ (*foward-looking Centers*) é uma lista das entidades referidas em U_n (pronomes e grupos nominais), ordenada segundo o critério de saliência baseado na função gramatical: sujeito > objecto > outras. Estas entidades fornecem possíveis ligações para o próximo enunciado. $Cp(U_n)$ (*preferred Center*) é a entidade melhor classificada (dominante) da lista $Cf(U_n)$;
- $Cb(U_n)$ (*backward-looking Center*) é a entidade dominante do enunciado anterior, ou seja, $Cp(U_{n-1})$. Se não existir um enunciado prévio, o $Cb(U_n)$ é nulo.

Outro conceito importante é a relação de transição entre enunciados. Considerando os centros anteriormente definidos, há três tipos diferentes de transições: *center Continuing*, *center Retaining* e *center Shifting*. O tipo de transição é determinado por dois factores: pela igualdade $Cb(U_n)$ e $Cb(U_{n-1})$ e pela igualdade $Cp(U_n)$ e $Cp(U_{n-1})$. Veja-se a tabela:

	$Cb(U_n) = Cb(U_{n-1})$	$Cb(U_n) \neq Cb(U_{n-1})$
$Cp(U_{n-1}) = Cp(U_n)$	*Continuing*	*Shifting*
$Cp(U_{n-1}) \neq Cp(U_n)$	*Retaining*	

Tabela 1 – Factores que determinam as transições do *Centering*.

Cada transição possui um significado que abaixo se apresenta.

- *Continuing*: $Cb(U_{n-1}) = Cb(U_n) = Cp(U_n)$ - neste tipo de transição dá-se continuidade ao assunto do discurso, mostra-se a intenção de o continuar.
- *Retaining*: $Cb(U_{n-1}) = Cb(U_n) \neq Cp(U_n)$ - o assunto é mantido, no entanto, demonstra-se a intenção de introduzir no discurso um novo assunto.
- *Shifting*: $Cb(U_{n-1}) \neq Cb(U_n)$ - é caracterizado por uma mudança de assunto.

Além das transições, existem algumas restrições e regras que permitem ligar correctamente os pronomes às entidades a que estes se referem. As restrições são três:

1. Para cada enunciado U_n só existe um $Cb(U_n)$.
2. Todas as entidades de $Cf(U_n)$ são efectivamente referidas no enunciado U_n.
3. $Cb(U_n)$ é a entidade melhor classificado de $Cf(U_{n-1})$ referida em U_n.

As regras são duas:

1. Se alguma entidade de $Cf(U_n)$ é referida por um pronome em U_{n+1}, então $Cb(U_{n+1})$ também o será.
2. A ordem de preferência na resolução de referências anafóricas pronominais é: *Continuing* > *Retaining* > *Shifting*.

Como exemplo do que foi descrito considere-se o seguinte segmento de discurso:

(3.1.)
U_1: O juiz leu a sentença.
U_2: Ele aplicou a pena máxima.
U_3: O advogado de defesa não concordou com ele.
U_4: Ele não interpôs recurso.
U_5: Ele achou que não tinha hipóteses.

Veja-se agora a análise do segmento de discurso.

Enunciado	$Cb(U_n)$	$Cf(U_n)$	$Cp(U_n)$	Transição
U_1	Ø	{ juiz, sentença }	juiz	Ø
U_2	juiz	{ Ele = juiz, pena máxima }	juiz	*Continuing*
U_3	juiz	{ advogado de defesa, ele = juiz}	advogado de defesa	*Retaining*
U_4	advogado de defesa	{ Ele = advogado de defesa, recurso }	advogado de defesa	*Shifting*
U_5	advogado de defesa	{ Ele = advogado de defesa, hipóteses }	advogado de defesa	*Continuing*

Tabela 2 – Análise do enunciado segundo a teoria de *Centering*.

A tabela ilustra todas as transições possíveis: *Continuing*, *Retaining* e *Shifting*. Os enunciados U_1 e U_2 referem-se ao "juiz" que é a entidade melhor classificada de Cf. Em U_3 o Cb continua a ser o "juiz", mas o "juiz" já não é a entidade melhor classificada de Cf, o que indica uma intenção de mudança de assunto. Em U_4 é concretizada essa mudança. E finalmente em U_5 volta-se a manter o assunto (centro) "advogado de defesa".

Depois de aqui apresentada a abordagem proposta para a resolução de anáforas pronominais refira-se ainda que outros autores propuseram extensões a esta teoria. É o caso de [Brennan *et al* (1987)], os quais apresentaram um refinamento da transição de *Shifting*. Também [Walker (1998)] defende que a procura dos antecedentes não deve ser limitada aos enunciados do segmento do discurso. Por sua vez [Strube (1998)] apresenta uma alternativa em que o Cb e os estados de transição são substituídos por uma *S-list* (*single linked list*) constituída pelas entidades do discurso, ordenadas por saliência.

Outros estudos podem ser consultados em [Poesio *et al* (2000)] que propõe a verificação da validade da primeira restrição e da primeira regra do *Centering*; ou [Kibble (2001)] que apresenta uma reformulação da segunda regra.

3.2. Sistemas de Recuperação de Informação utilizados

3.2.1. SENTA (*Software for the Extraction of N-ary Textual Associations*)

O sistema SENTA resultou de um estudo de [Gil (2002)], o qual tomou como referência o trabalho de [Dias *et al* (1999)], relacionado com a procura de métodos, puramente estatísticos, que permitem a identificação directa de expressões relevantes, que ocorrem frequentemente em corpora de dimensão elevada, constituídos por textos escritos numa qualquer língua natural.

Este sistema de recuperação de informação faz a extracção automática de expressões relevantes, formadas por sequências de unidades lexicográficas (por exemplo: caracteres, palavras, sinais de pontuação, etiquetas), contíguas ou não contíguas, que sejam assumidas como unidades sintáctico-semânticas, com significado próprio.

O SENTA é composto por um algoritmo, o *GenLocalMaxs*, que permite a identificação de expressões relevantes; e por uma medida, a *Expectativa Mútua*, para a quantificação da força de associação entre os componentes de cada sequência de unidades lexicográficas ou *n-grama*.

Os *n-gramas* são padrões textuais representados sob a forma de uma máscara, contendo elementos e opcionalmente algumas omissões ("__"), permitindo desta forma altos níveis de abstracção, como se pode ver pelos seguintes exemplos:

(3.2.) Força Área (2-gramas)
(3.3.) Milhões de contos (3-gramas)
(3.4.) Campeonato __ de Futebol (3-gramas)
(3.5.) Comissão __ __ das Privatizações (3-gramas)

Estes padrões textuais são, por exemplo, substantivos compostos, expressões idiomáticas, verbos compostos, locuções preposicionais, ou locuções adverbiais. Genericamente, são expressões que ocorrem mais vezes do que o simples acaso faria prever [Dias *et al* (1999)].

Para além da identificação de expressões relevantes e do cálculo da medida de quantificação da força de associação, este sistema também regista a frequência absoluta (número de ocorrências) de cada expressão.

3.2.2. *Google Desktop*

A aplicação *Google Desktop* surge da ideia de transportar a tecnologia de pesquisa *web* do *Google* para um ambiente confinado ao próprio computador.

O *Google Desktop* é uma ferramenta de pesquisa local, disponibilizada gratuitamente pela empresa *Google*. Possui uma interface *web* simples, semelhante à do motor de pesquisa *Google.com,* que torna possível a utilização do *browser* para procurar informação no próprio computador.

Após a sua instalação, o *Google Desktop* começa a indexar integralmente todos os documentos armazenados no computador (todas as palavras de cada documento). Essa indexação acontecerá somente quando o computador estiver inactivo por mais de 30 segundos, não prejudicando o seu desempenho normal. Todavia, dependendo do número de documentos, esse processo poderá levar várias horas. O índice criado vai sendo actualizado continuamente, à medida que são alterados e adicionados novos documentos.

O *Google Desktop* pode indexar e gerir uma grande variedade de recursos, incluindo: documentos do *Office*, documentos multimédia, documentos compactados, mensagens de correio electrónico, documentos presentes no histórico do *browser* e sessões de *chat*.

Quando é efectuada uma pesquisa no *Google Desktop,* é exibida uma página que mostra os resultados de pesquisa mais relevantes, cada um dos quais inclui o nome do ficheiro e um fragmento com os termos da sua pesquisa em destaque. Esta situação é a seguir exemplificada.

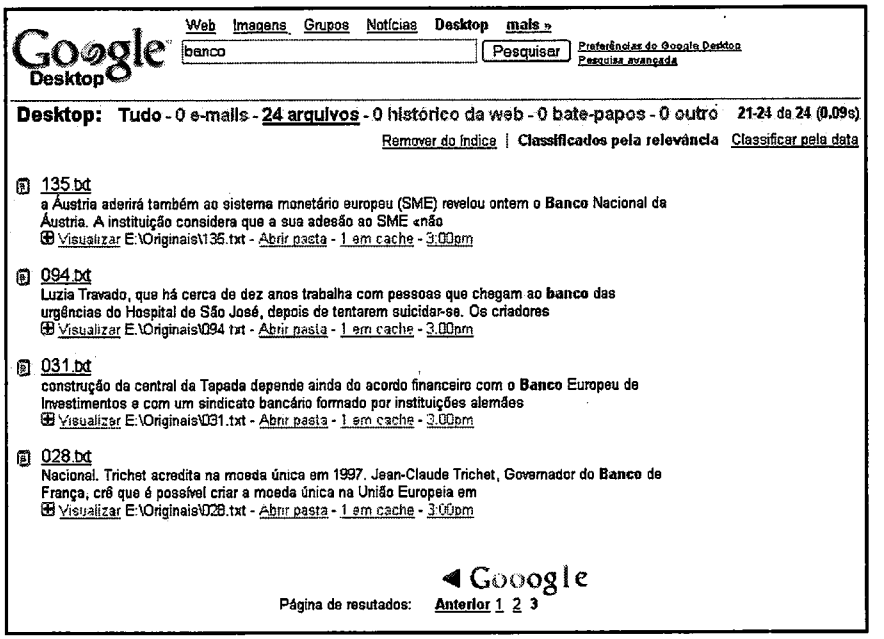

Figura 2 – Resultados de uma pesquisa no *Google Desktop*

A ferramenta de pesquisa *Google Desktop* que já vai na versão 5 suporta 29 línguas entre as quais o português. A localização em múltiplas línguas é uma evolução essencial para expandir a utilização da ferramenta, já que a sua principal funcionalidade é procurar documentos no PC, que à partida foram concebidos na língua de origem do seu utilizador.

4. Estudo de Caso

Este capítulo começa por descrever as escolhas realizadas, no que diz respeito ao tipo de anáforas que se procuram resolver (em 4.1.) e ao tipo de textos que constituem o corpus linguístico (em 4.2.). Posteriormente, é explicada a forma como foi implementado o método de resolução de anáforas pronominais em textos de língua portuguesa (em 4.3.); a substituição do referente anafórico pelo seu antecedente (em 4.4.); e a comparação das respostas dadas pelos sistemas de recuperação de informação SENTA e *Google Desktop*, em relação aos textos, antes e depois da substituição (em 4.5.).

4.1. Tipo de Anáforas Resolvidas

Este estudo incide sobre o tipo de anáfora pronominal e envolve a maioria dos pronomes pessoais de 3ª pessoa. O motivo desta escolha baseia-se no facto de os pronomes, e sobretudo os de 3ª pessoa, serem intuitivamente referências anafóricas num discurso. Por outro lado, os pronomes pessoais têm sido frequentemente a primeira escolha em métodos de resolução de anáforas, nomeadamente no método de *Centering*.

Mais concretamente, os pronomes aqui resolvidos são os pronomes pessoais rectos e oblíquos não reflexivos de 3ª pessoa[18]: ele(s), ela(s), o(s), a(s), lhe(s), incluindo as formas especiais dos pronomes quando ligados a certas terminações verbais: lo(s), la(s), no(s), na(s), quando contraídos com preposições: dele(s), dela(s), nele(s), nela(s), ou quando combinados: lho(s), lha(s).

4.2. Descrição do Corpus

Este estudo utiliza dois tipos de corpora diferentes. Um é constituído por 39 Pareceres da Procuradoria-Geral da República de Portugal (exemplo no Anexo A); o outro é formado por 160 textos jornalísticos do "Público" (exemplo no Anexo B).

[18] Doravante estes pronomes serão referidos apenas por "pronomes" ou "pronomes pessoais", para simplificar o discurso.

Inicialmente, estava previsto utilizar apenas um corpus de trabalho jurídico neste estudo, no entanto, quando se obtiveram os primeiros resultados no processo de RI, verificou-se que esse corpus era de pequena dimensão e poderia comprometer os resultados finais. Por este motivo e por ser interessante tratar um tipo de linguagem diferente, recorreu-se à utilização de um outro corpus de trabalho de maior dimensão, o jornalístico.

Para além da definição dos dois corpora de trabalho, o processo de resolução de anáforas, assim como um dos sistemas do processo de recuperação de informação (SENTA) exigem a definição de um corpus de teste (amostra), pois de outra forma seria impossível estudar todos os fenómenos tratados.

Neste sentido e por motivos de operacionalidade, no processo de resolução de anáforas pronominais, foi constituído apenas um corpus de teste a partir do corpus de trabalho jurídico. Por sua vez, no processo de recuperação de informação, também se trabalhou apenas com um corpus de teste, mas este já construído a partir do corpus de trabalho jornalístico.

Os **39 documentos jurídicos** (numerados de 1 a 39), do primeiro corpus de trabalho, foram seleccionados de entre um total de 3.017. A selecção fez-se da seguinte forma: os 3.017 documentos foram agrupados por ano (de 1940 a 2001); para cada ano foi calculada a média do número de pronomes pessoais por documento; escolheram-se os documentos pertencentes aos 3 anos com a média mais alta: 1992, 1994 e 1995, conforme a tabela seguinte; por fim, dos 45 documentos obtidos retiraram-se 6 que não continham enunciados.

Anos	Pronomes	Documentos	Média
...	...	...	...
1991	700	11	63,6
1992	**964**	**12**	**80,3**
1993	543	11	49,4
1994	**1356**	**20**	**67,8**
1995	**1137**	**13**	**87,5**
1996	1173	26	45,1
1997	339	12	28,3
1998	2702	57	47,4
1999	3480	87	40
2000	2948	619	4,8
2001	0	14	0

Tabela 3 – Média do número de pronomes por documento jurídico em cada ano.

Curiosamente, no ano de 2001, verificou-se a inexistência deste tipo de pronomes pessoais em todos os documentos; o que pode ser considerado como uma estratégia para evitar erros de interpretação.

Depois de constituído o corpus de trabalho, formou-se o corpus de teste, que se obteve da forma que se passa a explicar. Dos 39 documentos iniciais excluíram-se 8 (1, 5, 17, 18, 19, 27, 32 e 33) que não continham pronomes pessoais. Os restantes 31 foram divididos em 5 grupos, conforme a seguinte tabela:

Número de Pronomes	1-20	21-40	41-60	61-80	81-...
Documento	7, 10, 14, 15, 25, 30	2, 3, 4, 12, 20, 23, 28, 29, 35, 36	8, 9, 11, 12, 24, 34, 38	16, 22, 37, 39	6, 21, 26, 31

Tabela 4 – Agrupamento de documentos jurídicos em função do número de pronomes.

Seleccionou-se o primeiro documento de cada grupo, excepto do grupo com mais documentos, em que se seleccionaram os dois primeiros. Como resultado desta selecção, formou-se o corpus de teste constituído pelos documentos identificados com os números: 7, 2, 3, 8, 16 e 6. Considera-se portanto, que este conjunto de 6 documentos seja uma amostra representativa do corpus de trabalho jurídico.

Após uma análise morfo-sintáctica do corpus de trabalho jurídico, foi possível identificar correctamente 1.385 pronomes pessoais, num total de 105.523 sintagmas nominais e 368.414 palavras; o que representa respectivamente 1,31% e 0,38% do total.

Quanto à distância entre o pronome pessoal e o seu antecedente, verificou-se que este se encontrava no mesmo enunciado do pronome, em 84,8% dos casos; o que pode ser justificado em parte, pelo facto dos enunciados, neste domínio linguístico, serem longos e estruturalmente complexos, isto é, compostos geralmente por mais de três orações.

Por sua vez, **os 160 documentos jornalísticos** (numerados de 1 a 160), do segundo corpus de trabalho, foram seleccionados de entre um total de 292.917 documentos, pertencentes aos anos de 1994 e 1995. A selecção fez-se da seguinte forma: os 292.917 documentos foram agrupados por dia; para cada dia foi calculada a média do número de pronomes pessoais por documento; por fim, escolheram-se os documentos pertencentes ao dia com a média mais alta (06/12/1994), conforme se pode observar na tabela abaixo.

Data	Pronomes	Documentos	Média
...	...	...	...
07/10/1994	607	170	3,57
15/03/1994	613	163	3,76
14/03/1995	607	159	3,82
28/11/1995	617	159	3,88
07/03/1995	753	191	3,94
22/02/1994	719	177	4,06
11/01/1994	721	175	4,12
08/01/1994	624	151	4,13
24/09/1995	605	142	4,26
26/07/1994	645	151	4,27
19/04/1994	669	152	4,40
06/12/1994	**769**	**160**	**4,81**

Tabela 5 – Média do número de pronomes por documento jornalístico em cada dia.

Como resultado desta selecção, formou-se o corpus de trabalho constituído pelos 160 documentos, identificados com a data de 6 de Dezembro de 1994.

Para formar o corpus de teste, dos 160 documentos iniciais retiraram-se 60, que não continham pronomes pessoais. Os restantes 100 documentos foram divididos em 19 grupos, conforme a tabela que se segue.

Número de Pronomes	Documentos
1	1, 14, 16, 31, 35, 44, 50, 51, 52, 56, 63, 65, 67, 68, 79, 98, 99, 102, 107, 115, 118, 119, 120, 122, 125, 130, 134, 141, 142, 158
2	33, 48, 55, 64, 66, 77, 78, 83, 85, 91, 109, 131, 148, 149, 152
3	29, 40, 45, 53, 71, 73, 74, 82, 111, 114, 121, 123, 145, 146
4	12, 37, 43, 81, 86, 101, 106, 108, 117, 128, 147, 159
5	7, 93, 96, 155
6	13, 150, 156
7	90, 144, 151
8	42, 154
9	92, 124
10	39, 95, 127, 153
11	94
12	58
15	110, 143
16	4, 157
19	103
22	113
31	160
101	97
174	100

Tabela 6 – Agrupamento de documentos jornalísticos em função do número de pronomes.

Seleccionou-se o primeiro documento de cada grupo, excepto no caso do primeiro grupo, em que se seleccionaram os 4 primeiros documentos e nos casos dos segundo, terceiro e quarto grupos, em que se seleccionaram os 2 primeiros documentos. Como resultado desta selecção, formou-se um corpus de teste constituído pelos documentos identificados pelos números: 1, 14, 16, 31, 33, 48, 29, 40, 12, 37, 7, 13, 90, 42, 92, 39, 94, 58, 110, 4, 103, 113, 160, 97 e 100. Considera-se, assim, que este conjunto de 25 documentos seja uma amostra representativa do corpus de trabalho jornalístico.

Após uma análise morfo-sintáctica do corpus de trabalho jornalístico, foi possível identificar correctamente 715 pronomes pessoais num total de 25.452 sintagmas nominais e 92.994 palavras, o que representa respectivamente 2,81% e 0,77% do total.

Quanto à distância entre o pronome pessoal e o seu antecedente, verificou-se que em 56,1% dos casos este se encontrava no mesmo enunciado do pronome; quando se considerava o mesmo enunciado e o enunciado imediatamente anterior as ocorrências subiam para 87,6%; o que pode ser justificado em parte, pelo facto dos enunciados,

neste domínio linguístico, serem estruturalmente mais simples que os enunciados do corpus jurídico.

4.3. Implementação da Resolução de Anáforas

Cada um dos corpora de trabalho (jurídico e jornalístico) foi objecto de uma análise morfo-sintáctica parcial, utilizando o *parser* Palavras [Bick (2000)]. Este analisador morfo-sintáctico, desenvolvido no âmbito do projecto VISL (*Visual Interactive Syntax Learning*)[19], analisa todos os enunciados, mesmo os incompletos e incorrectos. O seu *output* pode ser visto no exemplo (4.2.), dado pelo enunciado (4.1.).

(4.1.) O juiz leu a sentença.

(4.2.)

```
UTT:cl(fcl)
S:g(np)
=D:pron(det "o" <artd> DET M S)      o
=H:n("juiz" <left> M S)     juiz
P:v(fin "ler" <fmc> PS 3S IND VFIN) leu
Od:g(np)
=D:pron(det "o" <artd> DET F S)      a
=H:n("sentença" <right> F S)         sentença
.
```

A partir da análise deste exemplo, pode referir-se que o enunciado é representado por uma única oração (UTT: cl), composto por um sujeito (S:), seguido de um predicado (P:) e por um objecto directo (Od:). Por sua vez o sujeito e o objecto directo são sintagmas nominais (g(np)) compostos por um determinante (D:) e um núcleo (H:).

A análise dos documentos, do corpus jurídico, revelou um número apreciável de erros que resultou da utilização de abreviaturas (por exemplo: "art." – artigo) e siglas (por exemplo: "T.C.E." – Tratado da Comunidade Europeia). A fim de não se

[19] VISL é um projecto de investigação e desenvolvimento do *Institute of Language and Communication*, da *University of Southern Denmark* (http://visl.sdu.dk/).

comprometer os resultados finais, optou-se por identificar e substituir, nestes textos, todas as abreviaturas e siglas pela sua forma não abreviada (pré-processamento). Por este motivo, foi necessário efectuar uma segunda análise morfo-sintáctica.

Sobre o resultado desta análise, em forma de árvores sintácticas, foi aplicada uma segunda ferramenta: "Palavras Extractor"[20] [Gasperin *et al* (2003)] que gerou três ficheiros no formato XML (*eXtensible Markup Language*): i) *words* – lista de palavras do documento e respectivos identificadores, ii) *pos* – lista com informação morfológica das palavras do documento e iii) *chunks* – lista com informação da análise morfo-sintáctica dos enunciados do documento. Veja-se agora o exemplo de cada um dos ficheiros, dado pelo enunciado (4.1.).

Exemplo do ficheiro *words*, (4.3.):

```
<words>
<word id="word_1">O</word>
<word id="word_2">juiz</word>
<word id="word_3">leu</word>
<word id="word_4">a</word>
<word id="word_5">sentença</word>
<word id="word_6">.</word>
</words>
```

[20] "Palavras Extractor" é uma ferramenta desenvolvida no Departamento de Informática da Universidade de Évora.

Exemplo do ficheiro *pos*, (4.4.):

```
<words>
<word id="word_1">
<art canon="o" gender="M" number="S">
  <secondary_art tag="artd"/>
</art>
</word>
<word id="word_2">
<n canon="juiz" gender="M" number="S"/>
</word>
<word id="word_3">
<v canon="ler">
<fin tense="PS" person="3S" mode="IND"/>
</v>
</word>
<word id="word_4">
<art canon="o" gender="F" number="S">
  <secondary_art tag="artd"/>
</art>
</word>
<word id="word_5">
<n canon="sentença" gender="F" number="S"/>
</word>
<words>
```

Exemplo do ficheiro *chunks*, (4.5.):

```
<text>
<paragraph id="paragraph_1">
<sentence id="sentence_1" span="word_1..word_6">
<chunk id="chunk_1" ext="sta" form="fcl" span="word_1..word_6">
<chunk id="chunk_2" ext="subj" form="np" span="word_1..word_2">
<chunk id="chunk_3" ext="n" form="art" span="word_1">
</chunk>
<chunk id="chunk_4" ext="h" form="n" span="word_2">
</chunk>
</chunk>
<chunk id="chunk_5" ext="p" form="v_fin" span="word_3">
</chunk>
<chunk id="chunk_6" ext="acc" form="np" span="word_4..word_5">
<chunk id="chunk_7" ext="n" form="art" span="word_4">
</chunk>
<chunk id="chunk_8" ext="h" form="n" span="word_5">
</chunk>
</chunk>
</chunk>
</sentence>
</paragraph>
</text>
```

A estes ficheiros XML foram aplicadas várias folhas de estilo XSL (*eXtensible Stylesheet Language*), transformando-os em novos documentos XML, com a finalidade de identificar os sintagmas nominais a que os pronomes pessoais fazem referência. Este processo divide-se em duas etapas fundamentais: "Identificação e Filtragem" e "Classificação e Selecção". A primeira etapa diz respeito à identificação correcta das referências anafóricas pronominais e dos seus possíveis candidatos; a segunda etapa classifica e selecciona as hipóteses de resolução. Ambas as etapas são a seguir descritas.

4.3.1. Etapa de Identificação e Filtragem

Nesta etapa procedeu-se à identificação de todas as referências anafóricas pronominais válidas e de todas as suas possíveis soluções. Esta informação foi organizada e filtrada, de forma a poder passar-se para a fase de classificação e selecção.

A etapa de identificação e filtragem é constituída por um pré-processamento e pelas seguintes 7 folhas de estilo: *01anaforas.xsl*, *02candidatos.xsl*, *03cria_cf.xsl*, *04filtro.xsl*, *05cria_cb.xsl*, *06cria_cb_cf.xsl* e *07cria_item.xsl*.

No pré-processamento identificaram-se todas as referências anafóricas pronominais de cada documento. Corrigiram-se manualmente, nos ficheiros XML, falsas referências anafóricas pronominais, nomeadamente, palavras em língua estrangeira (por exemplo: *la* - do castelhano e *elle* - do francês), artigos definidos (por exemplo: *o parecer*) e preposições (por exemplo: *venham a frequentar*). Foram ainda retiradas, de forma automática, outras falsas referências, tais como pronomes relativos (por exemplo: *o que*, *a qual*) e um caso concreto de pronomes pessoais, que referenciam acções e adjectivos (por exemplo: *a convenção não foi ratificada, mas não o foi porque...*). Este caso caracteriza-se por a referência anafórica pronominal surgir na forma de pronome pessoal *o*, (ou nas suas formas especiais: *-no* e *-lo*), acompanhado de um sintagma verbal, em que o verbo *ser, fazer* ou *dizer* é o verbo principal.

Ainda nesta etapa, aplicaram-se as folhas de estilo XSL acima referidas para cada um dos documentos, as quais se passam a descrever.

01anaforas.xsl

Nesta primeira folha implementou-se a parte automática do pré-processamento, referida anteriormente.

Para além disso, geraram-se elementos *sentence*, representativos de cada enunciado, e incluiu-se uma lista de elementos do tipo: *<anaphor span="word_x" word="palavra" type=""/>* ou do tipo: *<anaphor span="word_x" pointer="word_y..word_z" type="np"/>*, em que a primeira tem como objectivo identificar referências anafóricas pronominais e a segunda possíveis soluções (sintagmas nominais). Veja-se o exemplo (4.6.):

(4.6.)

```
<sentence id="sentence_8">
<anaphor span="word_362" pointer="word_361..word_362" type="np"/>
<anaphor span="word_366" pointer="word_365..word_367" type="np"/>
<anaphor span="word_370" pointer="word_369..word_370" type="np"/>
<anaphor span="word_387" pointer="word_387..word_389" type="np"/>
<anaphor span="word_389" word="eles" pointer=""/>
<anaphor span="word_399" pointer="word_397..word_399" type="np"/>
</sentence>
```

No caso de referências anafóricas pronominais, os atributos *span* e *word* correspondem à identificação do pronome e o atributo *pointer* fica por atribuir, à espera de uma solução. Por outro lado, no caso dos sintagmas nominais, o atributo *span* corresponde à identificação do núcleo do sintagma nominal, o atributo *pointer* à identificação do próprio sintagma nominal e o atributo *type* é sempre preenchido com o valor *np*, para melhor operacionalidade.

02candidatos.xsl

Esta folha de estilo é responsável por criar uma lista de todos os candidatos a solução, isto é, de todos os sintagmas nominais do documento.

A partir dos resultados do passo anterior, criou-se uma lista de elementos do tipo: *<candidate span="word_x" sid="y" pos="z"/>*. O atributo *span* corresponde à identificação do núcleo do sintagma nominal, o atributo *sid* corresponde à identificação de cada enunciado e, por sua vez, o atributo *pos* diz respeito à posição de cada candidato na lista. Veja-se o exemplo (4.7.):

(4.7.)

```
<candidate-set>
<candidate span="word_2" sid="1" pos="1"/>
<candidate span="word_4" sid="1" pos="2"/>
<candidate span="word_8" sid="1" pos="3"/>
...
</candidate-set>
```

03cria_cf.xsl

Esta folha de estilo permite criar uma lista limitada de candidatos compatíveis para cada referência anafórica pronominal.

A partir dos resultados dos dois passos anteriores, criou-se, para cada referência anafórica pronominal, uma lista de candidatos gramaticalmente compatíveis em género e número. Esta lista é ainda uma lista limitada, pois só possui candidatos pertencentes até alguns enunciados anteriores, por razões de performance computacional, mas sobretudo por imposição da teoria do *Centering*, que actua sobre o discurso em termos locais e não em absolutos. Este número de enunciados varia de acordo com as

características dos corpora. No caso do corpus jurídico ficou definido uma pesquisa até três enunciados anteriores. Para o corpus jornalístico, uma vez que os seus enunciados são estruturalmente mais simples, estipulou-se até oito enunciados anteriores.

Neste passo foram criados elementos do tipo *cf*. Veja-se o exemplo (4.8.):

(4.8.)

```
<sentence id="sentence_8" anaphors="1">
<anaphors><anaphor span="word_389"/></anaphors>
<cf id="8">
<anaphor span="word_362" pointer="word_361..word_362" type="np"/>
<anaphor span="word_366" pointer="word_365..word_367" type="np"/>
<anaphor span="word_370" pointer="word_369..word_370" type="np"/>
<anaphor span="word_387" pointer="word_387..word_389" type="np"/>
<anaphor span="word_389" word="eles" pointer="word_377" sid="8"/>
<anaphor span="word_389" word="eles" pointer="word_353" sid="7"/>
<anaphor span="word_399" pointer="word_397..word_399" type="np"/>
</cf>
</sentence>
```

Cada elemento *cf* tem um atributo *id* que indica o número do enunciado. Contém elementos *anaphor* com os sintagmas nominais do enunciado, identificados com os atributos *span*, *pointer* e *type="np"*; e com as possíveis soluções de cada pronome pessoal do enunciado em questão, identificados respectivamente pelos atributos *span*, *word*, *pointer* e *sid*.

04filtro.xsl

Aqui eliminaram-se alguns casos particulares de candidatos obtidos no passo anterior, que se verificou, na prática, não poderem ser solução. Nomeadamente, sintagmas nominais que antecedem imediatamente o pronome ou a partícula *que* seguida de pronome, pertencentes ao mesmo enunciado. Neste sentido, foram definidas as seguintes regras de restrição (filtros):

(1) *N1 +N2 + Pron*

(2) *N1 +N2 + 'que' + Pron*

(3) *N1 +N2 + 'que' + não-Prep + Pron*

(4) *N1 + N2 + 'que' + X + Y + [W + Z] + Pron*

Para que estas 4 regras se cumpram e os sintagmas nominais sejam filtrados é necessário que o núcleo dos sintagmas nominais esteja localizado na posição N1 ou na posição N2 e que na posição N2 não ocorra uma vírgula. Ainda na regra 3, entenda-se *não-Prep* como uma partícula que não é preposição. Finalmente, na regra 4, entenda-se *X, Y, W* e *Z* como representações de qualquer partícula, em que *X* e *Y* são obrigatórias e *W* e *Z* são opcionais.

05cria_cb.xsl

Na sequência dos resultados obtidos no passo anterior, a presente folha de estilo criou o elemento *cb* para cada enunciado. Este elemento corresponde ao primeiro elemento de *cf* do enunciado anterior. Veja-se o exemplo (4.9.):

(4.9.)

```
<sentence id="sentence_8" anaphors="1">
<anaphors><anaphor span="word_389"/></anaphors>
<cb id="8"><anaphor span="word_310" pointer="word_309..word_313"/></cb>
<cf id="8">
<anaphor span="word_389" word="eles" pointer="word_377" sid="8"/>
<anaphor span="word_389" word="eles" pointer="word_353" sid="7"/>
</cf>
</sentence>
```

Para além disso, foram eliminados de *cf*, todos os elementos *anaphor* com o atributo *type="np"*, uma vez que já cumpriram o seu propósito e já não são necessários.

06cria_cb_cf.xsl

Efectua-se, nesta folha de estilo, uma reorganização dos elementos *cf* que se receberam do passo anterior. Até aqui, os elementos *cf* continham elementos descendentes *anaphor*, mas a partir deste passo foram criados elementos *cb_cf*, que passaram a ser descendentes de *cf* e incluem apenas um elemento *anaphor*. Existem tantos *cb_cf* quanto os diferentes elementos *anaphor*. Veja-se o exemplo (4.10.):

(4.10.)

```
<sentence id="sentence_103" anaphors="2">
<cb id="103"><anaphor span="word_3113" pointer="word_3112..word_3113"/>
</cb>
<cf id="103"/>
<cb_cf>
<anaphor span="word_3149" word="lhe" pointer="word_3145" sid="103"/>
<anaphor span="word_3149" word="lhe" pointer="word_3143" sid="103"/>
<anaphor span="word_3149" word="lhe" pointer="word_3140" sid="103"/>
...
<anaphor span="word_3149" word="lhe" pointer="word_3024" sid="99"/>
</cb_cf>
<cb_cf>
<anaphor span="word_3213" word="lhes" pointer="word_3180" sid="103"/>
<anaphor span="word_3213" word="lhes" pointer="word_3131" sid="103"/>
<anaphor span="word_3213" word="lhes" pointer="word_3098" sid="101"/>
...
<anaphor span="word_3213" word="lhes" pointer="word_3009" sid="99"/>
</cb_cf>
</sentence>
```

As combinações aqui apresentadas já respeitam, integralmente, as restrições impostas pela teoria de *Centering*, evitando assim, a necessidade de recorrer aos filtros, também enunciados nessa teoria.

07cria_item.xsl

Esta é a última folha de estilo da etapa Identificação e Filtragem, a qual também efectua uma reorganização da informação obtida no passo anterior. Criou um novo elemento *item* descendente do elemento *cb_cf* que passa a incluir um par *cb*, *cf* . Este par é composto pelo *cb* do enunciado em questão e por um *cf* que contêm um *anaphor* do passo anterior. São gerados tantos *item* quantos os *anaphor* existentes de *cf*. Veja-se o exemplo (4.11.):

(4.11.)

```
<sentence id="sentence_8" anaphors="1">
<cb id="8"><anaphor span="word_310" pointer="word_309..word_313"/></cb>
<cf id="8"/>
<cb_cf>
<item>
<cb><anaphor span="word_310" pointer="word_309..word_313"/></cb>
<cf><anaphor span="word_389" word="eles" pointer="word_377" sid="8"/>
</cf>
</item>
<item>
<cb><anaphor span="word_310" pointer="word_309..word_313"/></cb>
<cf><anaphor span="word_389" word="eles" pointer="word_353" sid="7"/>
</cf>
</item>
</cb_cf>
</sentence>
```

Refira-se ainda que, após a aplicação desta folha de estilo, cada *item* possui uma possível solução. E são estas soluções que irão ser processadas na próxima etapa deste método de resolução de anáforas.

4.3.2. Classificação e Selecção

Na segunda e última etapa, classificação e selecção, classificaram-se todas as soluções encontradas, segundo o tipo de transição proposto pela teoria adoptada; seleccionou-se a melhor solução, em função da regra de prioridade definida; e finalmente obteve-se o texto da solução, que irá substituir o respectivo pronome.

A presente etapa é constituída pelas folhas de estilo: *08classificacao.xsl*, *09seleccao.xsl* e *10substituicao.xsl*, que a seguir se descrevem.

08classificacao.xsl

Partindo dos resultados do passo anterior, atribui-se a cada elemento *item*, um dos seguintes tipos de transição: *continuing*, *retaining* e *shifting*, de acordo com a tabela 1 (Factores que determinam as transições do *Centering*). Esta atribuição é feita em função de duas comparações. A primeira, $Cb(U_n) = Cb(U_{n-1})$, que distingue a transição

continuing e a transição *retaining* da transição *shifting*; e a segunda, $Cp(U_{n-1}) = Cp(U_n)$, que distingue a transição *continuing* da *retaining*.

No fim deste passo, cada elemento *item* tem um atributo *classification*, com um valor que corresponde à respectiva transição. Veja-se o exemplo (4.12.).

(4.12.)

```
<sentence id="sentence_8" anaphors="1">
<cb_cf>
<item classification="shifting">
<cb><anaphor span="word_310" pointer="word_309..word_313"/></cb>
<cf><anaphor span="word_389" word="eles" pointer="word_377" sid="8"/>
</cf>
</item>
<item classification="shifting">
<cb><anaphor span="word_310" pointer="word_309..word_313"/></cb>
<cf><anaphor span="word_389" word="eles" pointer="word_353" sid="7"/>
</cf>
</item>
</cb_cf>
</sentence>
```

09seleccao.xsl

Após ter sido encontrada a classificação para cada elemento *item*, é escolhido, para cada enunciado (*sentence*), um *item*, de acordo com a ordem decrescente de preferência das transições. A ordem seguida é: 1º - *continuing*, 2º - *retaining* e 3º - *shifting*. Em caso de empate, é escolhida a transição que implique uma menor distância entre o pronome e a solução. No entanto, existe uma excepção à regra de empate, que em caso de várias hipóteses de *shifting*, para os pronomes –lo(s) e –la(s), escolhe-se preferencialmente a mais próxima do início do enunciado. Veja-se o exemplo (4.13.) que resulta da aplicação desta regra ao exemplo anterior (4.12.):

(4.13.)

```
<sentence id="sentence_8" anaphors="1">
<cb_cf>
<item classification="shifting">
<cf><anaphor span="word_389" word="eles" pointer="word_377" sid="8"/>
</cf>
</item>
</cb_cf>
</sentence>
```

10substituicao.xsl

Após ter sido encontrada a solução para cada pronome do documento, neste passo, cria-se um elemento *pron* com o contexto textual da ocorrência do pronome e um elemento *n* com as palavras que constituem a solução. Veja-se o exemplo (4.14.):

(4.14.)

```
<anaphor-set anaphors="14" resolved="14">
<pron span="word_389" pron="eles">desde_que um de eles não</pron>
<n pointer="word_377" n="cargos">os cargos</n>

<pron span="word_644" pron="ela">1991 . De ela resulta</pron>
<n pointer="word_634" n="entrada">entrada</n>

...
</anaphor-set>
```

Com a aplicação deste conjunto de dez folhas de estilo encontrou-se uma solução para cada referência anafórica pronominal, a qual está de acordo com a teoria de *Centering*, abordada neste estudo. O seguinte esquema ilustra todo este processo.

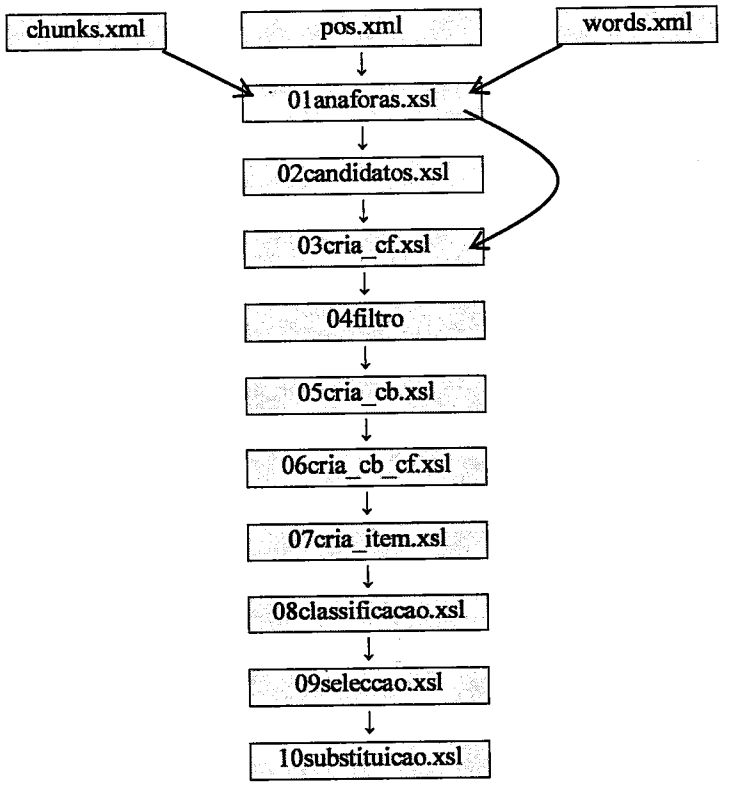

Figura 3 – Etapas do algoritmo de resolução de anáforas pronominais.

4.4. Substituição do Referente Anafórico pelo seu Antecedente

A partir do processo anterior obteve-se um ficheiro XML com a solução para cada pronome pessoal encontrado, como foi exemplificado em (4.14.).

Nesta fase intermédia foi desenvolvida uma ferramenta em linguagem C, que recebe como *input* o contexto textual do pronome, a respectiva solução e o ficheiro de texto onde esse pronome ocorre; e é devolvido, como resultado, um novo ficheiro de texto em que o pronome foi substituído pela respectiva solução.

O algoritmo desenvolvido tenta fundamentalmente encontrar no texto, o pronome em questão, através do seu contexto textual e proceder à sua substituição. O contexto textual é composto por duas partículas (palavras ou pontuação) que antecedem o pronome, o próprio pronome e ainda a partícula que o segue.

O contexto textual é, portanto, formado pela junção de quatro partículas que resultam da análise morfo-sintáctica do texto original. Esta análise separou algumas palavras contraídas e ignorou alguns símbolos, o que constitui um entrave à identificação correcta do pronome. Para tentar ultrapassar este obstáculo procedeu-se à

contracção automática dos pronomes com as preposições que os antecedem e fez-se uma adição semi-automática de símbolos não existentes na análise morfo-sintáctica. Desta forma foi possível a identificação / pesquisa de todos os pronomes em cada texto.

Verificou-se que a opção pela utilização do contexto textual é uma estratégia eficaz, não só para identificação do respectivo pronome, como também para evitar o falso reconhecimento de outro pronome idêntico.

5. Avaliação

Neste capítulo pretende-se avaliar o método de resolução de anáforas pronominais implementado (aplicação do *Centering*), assim como, avaliar o impacto dessa resolução na resposta dos sistemas de recuperação de informação, nomeadamente, os sistemas SENTA e *Google Desktop*.

Avaliar correctamente sistemas de resolução de anáforas é um processo exigente, que está dependente de alguns factores. Um factor que limita bastante a avaliação é a qualidade da anotação dos corpora. Numa situação ideal e mais justa de comparação de resultados, a anotação deveria estar isenta de erros. Segundo [Mitkov (2002)] é possível ter um sistema de resolução de anáforas que tenha um mau desempenho e, no entanto, possuir um algoritmo muito eficiente. Há pois que distinguir o todo (pré-processamento e algoritmo) do próprio algoritmo. Neste sentido, sempre que possível, a avaliação recairá sobre o algoritmo e não sobre o sistema em geral, pois de outra forma implicaria uma revisão aprofundada dos corpora utilizados neste estudo.

Relativamente à avaliação do impacto da resolução de anáforas sobre os sistemas de recuperação de informação utilizados, esta pode ser vista sobre dois ângulos diferentes: sob o ponto de vista do sistema ou do ponto vista humano. A avaliação centrada no sistema será feita através da comparação de resultados obtidos, por outro lado, a avaliação dos utilizadores será feita de acordo com a satisfação de uma necessidade de informação proposta.

5.1. Métodos de Avaliação

5.1.1. Método de Avaliação do Processo de Resolução de Anáforas

Para avaliar a performance do algoritmo de resolução de anáforas pronominais implementado, utilizaram-se quatro medidas de avaliação: as medidas de Precisão (*Precision*) e de Cobertura (*Recall*), propostas por [Aone & Bennett (1995)], a Medida-F_1 (F_1-*Measure*) definida por [Rijsbergen (1979)] e a Taxa de Sucesso Crítico definida por [Mitkov (2000)]. As três primeiras medidas de avaliação são medidas clássicas habitualmente utilizadas para avaliar sistemas de resolução de anáforas. Já a última

medida é uma tentativa inovadora de [Mitkov (2000)] para determinar qual a parte do sucesso que se deve aos critérios de escolha do antecedente no algoritmo.

De seguida serão apresentadas as definições de cada uma das medidas.

- Precisão (P) = $\frac{Número\_de\_anáforas\_correctamente\_resolvidas}{Número\_de\_anáforas\_correctamente\_identificadas}$

- Cobertura (C) = $\frac{Número\_de\_anáforas\_correctamente\_resolvidas}{Número\_de\_anáforas\_identificadas}$

- Medida-F_1 = $\frac{2 \times P \times C}{P + C}$, em que P representa a Precisão e o C a Cobertura. Esta medida traduz o desempenho geral do sistema.

- Taxa de Sucesso Crítico (TSC) = $\frac{R}{T}$, em que R representa o número de anáforas correctamente resolvidas, com mais de um candidato e T, o número total de anáforas com mais de um candidato.

Ainda para tentar compreender melhor os critérios de escolha do antecedente, serão apresentados alguns resultados intermédios do processo de resolução de anáforas, bem como uma caracterização dos casos de insucesso (em 5.3.1.).

5.1.2. Método de Avaliação das Respostas dadas pelos SRIs utilizados

Para chegar ao objectivo final deste estudo, é necessário avaliar a influência da resolução de anáforas pronominais em processos de recuperação de informação. Assim sendo, formaram-se dois conjuntos de documentos, um com os documentos originais e outro com os documentos que passaram por substituições. A cada um destes conjuntos de documentos aplicaram-se os dois sistemas de recuperação de informação propostos (SENTA e *Google Desktop*) e posteriormente fez-se uma análise comparativa das duas respostas fornecidas por cada um dos sistemas.

A seguir descrever-se-á a forma como se fez essa análise comparativa em cada um dos sistemas.

5.1.2.1. Comparação das Respostas do Sistema SENTA

Para cada par, documento original e documento com substituições, foi aplicada a ferramenta SENTA e obteve-se como resultado, as expressões relevantes de cada documento, a sua frequência e a força de associação das respectivas expressões relevantes. Veja-se um exemplo deste *output*:

```
Begin Process ...
Mon Oct  1 01:37:33 WEST 2007
======
2grams
======
0.000426388   5       de _____ ,
0.000764088   2       são mais
0.00138925    4       « _____ »
0.00138925    4       « _____ _____ »
0.00152818    2       tão vaidoso
0.00152818    2       primeira dama
0.00154091    11      . _____ ,
0.00191022    2       contexto lúdico
======
3grams
======
0.000764088   2       que _____ . _____ a
0.000848987   2       do _____ _____ , a
0.00109155    2       , _____ a _____ de
0.00109155    2       a _____ _____ , o
0.00114613    2       de _____ . E
0.00191022    2       das _____ _____ jogos _____ _____ .
0.00191022    3       . _____ a _____ .
0.00198981    5       , _____ , _____ ,
0.00234436    3       ) , o
0.00286533    3       contos de fadas
======
4grams
======
0.00191022    2       , a violência _____ _____ .
0.00191022    2       , o sr .
0.00191022    2       sr . padre Victor
0.00286533    3       a dra . Maria
======
5grams
======
======
6grams
======
0.00191022    2       . E , aqui sim ,
End Process ...
Mon Oct  1 01:39:00 WEST 2007
```

Assim para cada par de documentos é possível fazer uma comparação e verificar se houve alterações significativas ou não, provocadas pela resolução de anáforas.

Neste sentido, fez-se a comparação das 10 expressões mais relevantes, em termos da medida de **Expectativa Mútua** (que quantifica a força de associação de cada *n-grama*), em ambos os conjuntos de documentos.

5.1.2.2. Comparação das Respostas do Sistema *Google Desktop*

Tendo em atenção, os dois conjuntos de documentos, com e sem substituições, prepararam-se vinte testes, para avaliar a influência da resolução de anáforas. Estes testes / pesquisas consistiram em expressões nominais que foram pesquisadas em cada um dos conjuntos de documentos, em separado. Para estas pesquisas foram escolhidas as expressões nominais mais frequentemente utilizadas nas substituições dos pronomes.

Após a aplicação das vinte pesquisas em cada um dos conjuntos de documentos, registou-se, a partir da resposta do sistema, o *ranking* dos documentos, mais concretamente os 10 melhores classificados, em termos de relevância. Veja-se o registo de uma pesquisa na tabela abaixo.

Pesquisa nº 12	Ordem de Relevância dos Documentos									
Corpus Original	103	100	94	97	128	95	42	126	123	124
Corpus Modificado	103	100	97	94	95	128	42	126	123	124

Tabela 7 – Comparação de resultados para uma pesquisa no *Google Desktop*.

Com os dois resultados referentes à mesma pesquisa, um por cada conjunto de documentos, é possível fazer uma comparação e verificar se há alterações entre os 10 melhores classificados, em termos de relevância.

Posteriormente, analisaram-se as alterações de *ranking* e pediu-se a um grupo de pessoas que verificasse se, na sua opinião, estas significavam um aumento de precisão (maior satisfação da necessidade de informação inicial).

5.2. Resultados

Os resultados a apresentar serão divididos em duas partes: resultados do processo de resolução de anáforas pronominais (em 5.2.1.) e resultados do processo de recuperação de informação (em 5.2.2.).

5.2.1. Resultados do Processo de Resolução de Anáforas

De seguida, serão apresentadas várias tabelas com os resultados obtidos em cada documento do corpus de teste, assim como no total de documentos desse mesmo corpus.

Documento	Pronomes identificados	Pronomes bem identificados	Pronomes bem resolvidos	Precisão	Cobertura	Medida-F_1
2	32	29	19	0,655	0,594	0,623
3	22	21	12	0,571	0,545	0,558
6	86	84	39	0,464	0,453	0,459
7	14	14	7	0,5	0,5	0,5
8	42	40	29	0,725	0,69	0,707
16	63	62	45	0,726	0,714	0,720
Total	259	250	151	0,604	0,583	0,593

Tabela 8 – Cálculo da Precisão, Cobertura e Medida-F_1 nos documentos do corpus de teste jurídico.

Neste cálculo da Cobertura e como foi sugerido por [Aone & Bennett (1995)], deve entender-se por "Pronomes identificados" todos os pronomes que a análise morfo-sintáctica conseguiu identificar e não os que podem ser identificados manualmente, no documento. Ou seja, os pronomes identificados são a soma dos pronomes reais e dos incorrectamente classificados como pronomes.

Documento	Pronomes com mais de um candidato	Pronomes bem resolvidos com mais de um candidato	Taxa de Sucesso Crítico
2	28	19	0,679
3	21	12	0,571
6	83	35	0,422
7	14	7	0,5
8	40	29	0,725
16	58	41	0,707
Total	244	143	0,586

Tabela 9 – Cálculo da Taxa de Sucesso Crítico nos documentos do corpus de teste jurídico.

Documento	Precisão	Cobertura	Medida-F_1	Taxa de Sucesso Crítico
2	0,655	0,594	0,623	0,679
3	0,571	0,545	0,558	0,571
6	0,464	0,453	0,459	0,422
7	0,5	0,5	0,5	0,5
8	0,725	0,69	0,707	0,725
16	0,726	0,714	0,720	0,707
Total	0,604	0,583	0,593	0,586

Tabela 10 – Medidas de avaliação utilizadas nos documentos do corpus de teste jurídico.

Pronome	Ocorrências	Mal Resolvidas	Taxa de Insucesso
lhe	41	29	70,7%
lhes	26	12	46,2%
ele	41	16	39%
eles	24	6	25%
ela	25	9	36%
elas	9	1	11,1%
o / -o	15	6	40%
os / -os	10	0	0%
a / -a	24	12	50%
as / -as	16	2	12,5%
-lo	5	2	40%
-los	2	0	0%
-la	8	3	37,5%
-las	3	1	33,3%
-no	0	0	0%
-nos	0	0	0%
-na	1	0	0%
-nas	0	0	0%

Tabela 11 – Casos de insucesso nos documentos do corpus de teste jurídico.

Tipo de Insucesso	Pronomes
Antecedente distante	54
Antecedente é o 2º SN compatível anterior	24
Parte do antecedente	14
Referência a forma verbal	2
Referência a duas entidades	1
Referência a uma ideia	2
Catáfora	1
Antecedente com sinal de plural "(s)"	1

Tabela 12 – Tipo de insucesso nos documentos do corpus de teste jurídico.

Corpus / Tipo de Transição	Corpus de Trabalho Jurídico	Corpus de Teste Jurídico	Corpus de Trabalho Jornalístico	Corpus de Teste Jornalístico
Continuing	1,88%	0,79%	0,7%	0,85%
Retainig	2,96%	2,78%	1,54%	2,14%
Shifting	95,16%	96,43%	97,76%	97,01%

Tabela 13 – Percentagem de cada tipo de transição-*Centering* nos vários corpora.

Corpus / Localização do Antecedente	Corpus de Trabalho Jurídico	Corpus de Teste Jurídico	Corpus de Trabalho Jornalístico	Corpus de Teste Jornalístico
Mesmo enunciado	84,8%	86,1%	56,1%	44,2%
Até ao enunciado anterior	98,3%	97,2%	87,6%	82,1%
Até 3 enunciados anteriores	95,2%	99,2%	94,5%	97%
Até 5 enunciados anteriores	99,9%	100%	99,4%	99,5%
Até 8 enunciados anteriores	100%	100%	100%	100%

Tabela 14 – Localização do antecedente nos vários corpora.

5.2.2. Resultados do Processo de Recuperação de Informação

Os resultados a apresentar serão divididos em duas partes: resultados do sistema SENTA (em 5.2.2.1.) e resultados do *Google Desktop* (em 5.2.2.2.).

5.2.2.1. Resultados do SENTA

De seguida, serão apresentadas várias tabelas com os resultados comparativos obtidos (antes e depois das substituições), em cada documento do corpus de teste jornalístico.

Documento	Número de Pronomes	Expressões Relevantes		
		Novas Entradas no Top-10?	Alterações do Top-10?	Novas Entradas?
1	1	Não	Sim	Não
14	1	Não	Não	Não
16	1	Não	Não	Não
31	1	Não	Não	Não
33	2	Sim	Sim	Sim
48	2	Sim	Sim	Sim
29	3	Sim	Sim	Sim
40	3	Sim	Sim	Não
12	4	Não	Não	Sim
37	4	Não	Não	Não
7	5	Não	Sim	Sim
13	6	Sim	Sim	Sim
90	7	Sim	Sim	Sim
42	8	Sim	Sim	Sim
92	9	Sim	Sim	Sim
39	10	Sim	Sim	Sim
94	11	Sim	Sim	Sim
58	12	Não	Sim	Sim
110	15	Sim	Sim	Sim
4	16	Sim	Sim	Sim
103	19	Sim	Sim	Sim
113	22	Sim	Sim	Sim
160	31	Sim	Sim	Sim
97	101	Sim	Sim	Sim
100	174	Sim	Sim	Sim

Tabela 15 – Alteração de expressões relevantes no corpus de teste jornalístico.

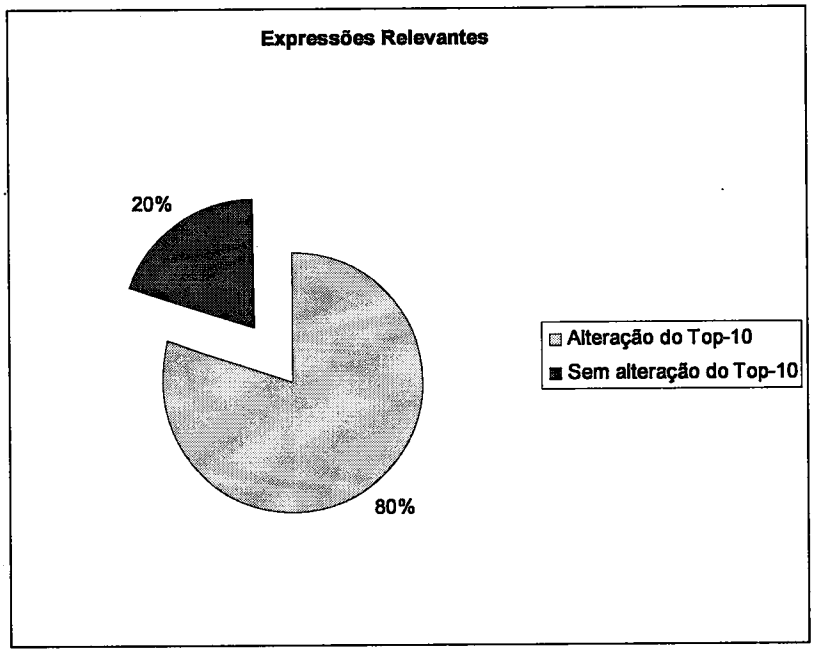

Figura 4 – Resultados na obtenção de expressões relevantes do corpus de teste jornalístico.

Documento	Alterações do Top-10?	Proporção Pronomes / Palavras
31	Não	0,16%
16	Não	0,21%
40	Sim	0,24%
14	Não	0,25%
33	Sim	0,27%
1	Sim	0,31%
7	Sim	0,37%
12	Não	0,42%
48	Sim	0,53%
92	Sim	0,54%
37	Não	0,61%
29	Sim	0,68%
58	Sim	0,69%
13	Sim	0,70%
160	Sim	0,77%
94	Sim	0,79%
90	Sim	0,79%
42	Sim	0,96%
39	Sim	0,99%
4	Sim	1,28%
110	Sim	1,42%
103	Sim	1,42%
113	Sim	1,46%
97	Sim	2,87%
100	Sim	2,99%

Tabela 16 – Alteração do Top-10 de expressões relevantes no corpus de teste jornalístico.

5.2.2.2. Resultados do *Google Desktop*

De seguida, serão apresentadas várias tabelas com os resultados obtidos nos dois corpora de trabalho (jurídico e jornalístico).

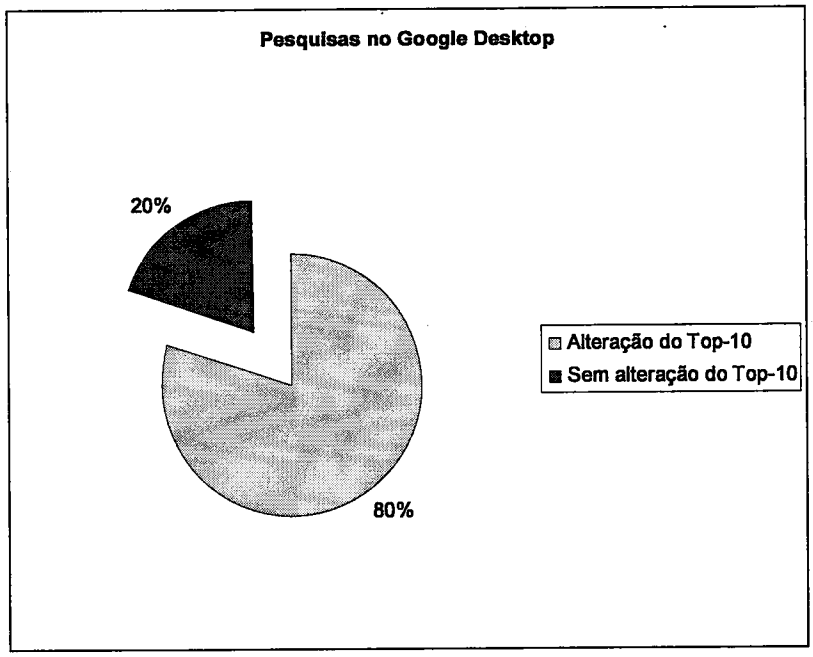

Figura 5 – Resultados das pesquisas efectuadas no corpus de trabalho jornalístico.

			Opinião sobre o documento mais relevante				
Expressão Nominal Pesquisada	**Documento que subiu no *ranking***	**Documento imediatamente ultrapassado**	**Indivíduo 1**	**Indivíduo 2**	**Indivíduo 3**	**Indivíduo 4**	**Geral**
Banesto	156	**154**	154	154	154	154	**154**
construção	90	**52**	52	52	52	52	**52**
dinheiro	108	**160**	108	160	160	160	**160**
facto	**143**	100	143	143	143	100	**143**
governo	**73**	70	73	73	73	73	**73**
José Roquette	**153**	146	146	153	153	153	**153**
ministro	148	**146**	148	146	146	146	**146**
mundo	97	**94**	94	97	94	94	**94**
papel	**42**	157	157	42	42	42	**42**
processo	151	**64**	64	64	64	64	**64**
prova	100	**159**	159	159	159	159	**159**
registo	**29**	154	154	29	29	29	**29**
tempo	117	**113**	113	113	113	113	**113**
Totta	156	**152**	152	152	156	152	**152**
vida	**4**	103	4	4	4	4	**4**
voz	**109**	160	109	109	109	109	**109**

Tabela 17 – Opiniões sobre a mudança de *ranking* em pesquisas efectuadas no corpus jornalístico.

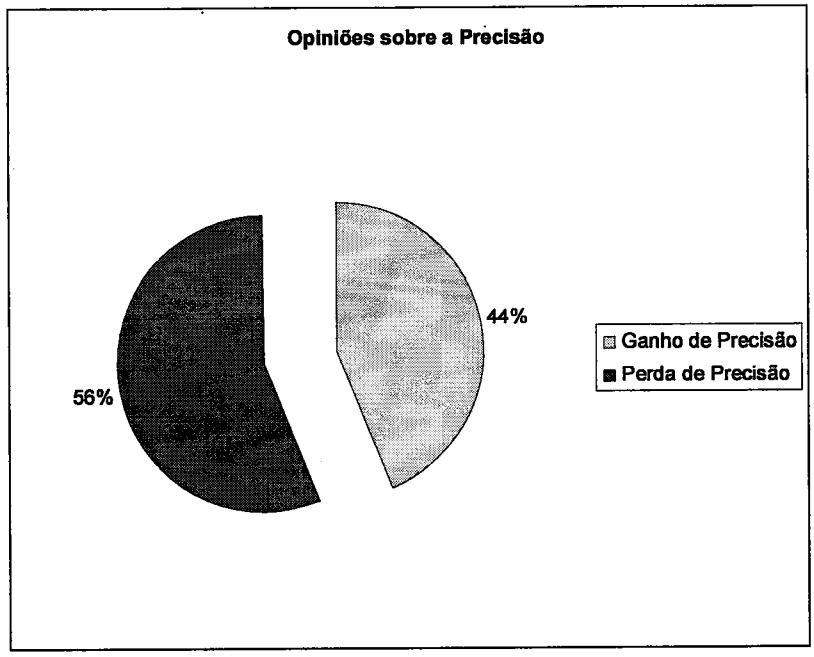

Figura 6 – Opiniões sobre a precisão dos resultados resultantes da integração da RA.

5.3. Análise dos Resultados

A análise dos resultados será dividida em duas partes: resultados do processo de resolução de anáforas pronominais, (em 5.3.1.) e resultados dos sistemas de recuperação de informação utilizados (em 5.3.2.).

5.3.1. Análise dos Resultados da Resolução de Anáforas

Como definido no subcapítulo Métodos de Avaliação, (em 5.1.), optou-se por utilizar quatro medidas de avaliação para o processo de resolução de anáforas. São elas: a Precisão, a Cobertura, a Medida-F_1 e a Taxa de Sucesso Crítico. No entanto, como já foi referido anteriormente, considera-se mais correcto que a avaliação recaia sobre o algoritmo e não sobre o sistema em geral (pré-processamento e algoritmo). Neste sentido, valoriza-se mais as medidas de avaliação: Precisão e Taxa de Sucesso Crítico.

Com a medida de Precisão, acompanhar-se-á a evolução do algoritmo no processo de resolução de anáforas pronominais, até à sua proposta final, apresentada em

(3.1.). Posteriormente, serão analisados outros aspectos considerados relevantes neste processo de resolução; e por fim, é feita uma comparação com resultados obtidos noutros estudos.

Vejam-se então, os resultados alcançados em cada uma das etapas / opções de desenvolvimento, até se chegar ao algoritmo final.

Numa primeira fase de experimentação / implementação da teoria de *Centering* foi utilizado o método de resolução de anáforas proposto por [Brennan *et al* (1987)]. Este difere essencialmente do proposto por [Grosz *et al* (1986)], por possuir mais um estado de transição, *shifting-1*, que permitiria resolver algumas situações de ambiguidade. No entanto, apresentou os seguintes resultados, para os documentos do corpus de teste jurídico:

Documentos	Precisão da Proposta de [Brennan *et al* (1987)]	Precisão da Proposta de [Grosz *et al* (1986)]
2	0,531	0,594
3	0,5	0,545
6	0,326	0,407
7	0,643	0,5
8	0,548	0,69
16	0,603	0,714
Total	0,486	0,568

Tabela 18 – Resultados das propostas de [Brennan *et al* (1987)] e [Grosz *et al* (1986)].

Visto a abordagem de [Brennan *et al* (1987)] apresentar piores resultados globais, esta foi preterida em favor da proposta de [Grosz *et al* (1986)].

Perante estes resultados alcançados pela abordagem de [Grosz *et al* (1986)] e após uma caracterização dos casos de insucesso (conforme a tabela seguinte), tentou-se ainda melhorar um pouco mais os resultados, com a introdução de regras suplementares.

Pronome	Ocorrências	Casos Mal Resolvidos	Taxa de Insucesso
lhe	41	29	70,7%
lhes	26	12	46,2%
ele	41	16	39%
eles	24	6	25%
ela	25	9	36%
elas	9	1	11,1%
o / -o	19	10	52,6%
os / -os	10	0	0%
a / -a	24	12	50%
as / -as	17	3	17,6%
-lo	9	8	88,9%
-los	2	0	0%
-la	8	5	62,5%
-las	3	1	33,3%
-no	0	0	0%
-nos	0	0	0%
-na	1	0	0%
-nas	0	0	0%

Tabela 19 – Casos de insucesso na proposta de [Grosz *et al* (1986)].

À proposta de [Grosz *et al* (1986)] acrescentaram-se quatro tipos de regras que permitiram, nomeadamente, filtrar alguns casos de pronomes, que não referiam expressões nominais; eliminar alguns candidatos a antecedente, que não podiam ser referência; ignorar alguns pronomes (sem antecedentes até três enunciados anteriores), que não podiam ser resolvidos correctamente; e ainda uma alteração na ordem de selecção de candidatos para os pronomes –lo(s) e –la(s). Conseguiu-se desta forma, alcançar os resultados finais que a seguir se apresentam:

Documentos	Precisão da Proposta de [Grosz *et al* (1986)]	Precisão do Algoritmo Final
2	0,594	0,655
3	0,545	0,571
6	0,407	0,464
7	0,5	0,5
8	0,69	0,725
16	0,714	0,726
Total	0,568	0,604

Tabela 20 – Resultados da proposta de [Grosz *et al* (1986)] e do algoritmo proposto.

Tendo em conta os resultados obtidos para o total de pronomes do corpus de teste jornalístico (Precisão: 0,604; Cobertura: 0,583; Medida-F_1: 0.593; Taxa de Sucesso Crítico: 0,586), pode afirmar-se que se encontrou um algoritmo que resolve com sucesso 60% dos casos.

Por outro lado, numa perspectiva de insucesso, há que realçar que o pronome **lhe** obteve a maior taxa de insucesso (70,7%). Este resultado já era esperado, pois o pronome **lhe** não impõe restrições de género aos seus possíveis antecedentes, dificultando a sua escolha. Ainda em relação ao insucesso, refira-se que os pronomes na forma singular obtiveram piores resultados que os mesmos na forma plural. Isto deve-se ao maior número de sintagmas nominais (candidatos), na forma singular, presentes nos documentos. Uma justificação, que pode também ser encontrada para a maior parte dos casos de insucesso, é a ocorrência em simultâneo de uma percentagem elevada de transições *shifting* (96,43%) e de pronomes com antecedentes distantes (54 em 99). Ou seja, a não existência de outras transições (*continuing* e *retaining*), implica que a escolha do candidato a antecedente se faça sobretudo pelo mais próximo, não permitindo alcançar antecedentes distantes.

Por fim, relativamente a resultados de outros estudos, em língua portuguesa, apresentados em (2.1.2.3.), pois não seria leal a comparação com outras línguas, refira-se que aqueles ficam um pouco abaixo dos aqui apresentados (60%).

5.3.1. Análise dos Resultados dos Sistemas de RI utilizados

A análise dos resultados dos dois sistemas de recuperação de informação utilizados será dividida em duas partes: análise de resultados do sistema SENTA, (em 5.3.1.1.) e análise de resultados do *Google Desktop*, (em 5.3.1.2.).

5.3.1.1. Análise de Resultados do SENTA

A partir dos resultados expressos na tabela 15, que referem a alteração de expressões relevantes no corpus de teste jornalístico, pode constatar-se que, para estes documentos (de dimensão entre 1KB e 33KB), nos quais ocorram mais de 5 pronomes pessoais, há uma modificação importante das suas expressões relevantes. Mais concretamente, há uma alteração evidente nas 10 expressões mais relevantes de cada documento. O mesmo se passa, quando se faz a análise em termos da proporção de pronomes, em relação às palavras do documento (ver tabela 16). Neste caso, a partir de proporções superiores a 0,61% há também uma alteração nas 10 expressões mais relevantes de cada documento.

Pode-se então dizer que quando houver um número significativo de pronomes pessoais nos documentos, a sua resolução implicará uma resposta diferente por parte do sistema SENTA.

5.3.1.2. Análise de Resultados do *Google Desktop*

Numa primeira fase de obtenção de resultados, utilizou-se o corpus de trabalho jurídico, para o qual se obteve os resultados expressos na tabela 24. Estes reflectem quase sempre uma mudança na lista ordenada de documentos recuperados, após a substituição dos pronomes pelas respectivas soluções. Mais precisamente, nas 20 pesquisas efectuadas, antes e após as substituições, verificou-se que em 12 das 20 pesquisas, havia uma alteração na posição dos 10 documentos mais relevantes. Mais ainda, nestes 12 casos, 6 implicavam também uma mudança nas primeiras 5 posições. Por outro lado, ocorreram 6 casos em que não houve qualquer alteração das primeiras 10 posições.

Apesar de estes dados confirmarem que a resolução de anáforas pronominais influencia a maior parte das respostas, dadas pelo sistema *Google Desktop*, e perante o facto de este corpus de trabalho jurídico ser de dimensão reduzida (39 documentos), põe-se a questão se estes resultados se manteriam num corpus maior e com características diferentes. Por esta razão, decidiu-se testar novamente este sistema com um corpus de trabalho jornalístico de maior dimensão (160 documentos).

Para este novo corpus de trabalho, foram também efectuadas 20 pesquisas, antes e após as substituições dos pronomes. Verificou-se, neste caso, que para as 20 pesquisas realizadas, 16 implicavam uma mudança nas 10 primeiras posições. Mais ainda, destas 16 mudanças, 15 constituíam também uma mudança nas 5 primeiras posições. No campo oposto, aconteceram 4 casos em que não houve qualquer alteração das primeiras 10 posições mais relevantes (ver tabelas 25 e 26).

Com estes dois resultados, de corpora diferentes, pode-se concluir que a resolução de pronomes afecta as respostas dadas pelo sistema *Google Desktop*.

Contudo, falta verificar se esta alteração de respostas implica verdadeiramente um aumento de precisão, ou seja, uma maior satisfação da necessidade de informação por parte do utilizador deste sistema. Neste sentido, pediu-se a opinião a um grupo de pessoas que confirmassem a alteração de resultados, em termos de relevância (conforme a tabela 17). As opiniões expressas foram algo inconclusivas, pois para os 16 casos em que houve alteração de *ranking*, a maioria das pessoas concordou que em 9 desses casos não havia aumento de precisão, mas o mesmo já não acontecia para os restantes 7 casos.

6. Conclusão

Como conclusão deste trabalho, será feita uma avaliação global da aplicação da teoria de *Centering*, na resolução de referências anafóricas pronominais, em textos de língua portuguesa (em 6.1.), assim como também uma avaliação do seu impacto nos dois sistemas de recuperação de informação utilizados, SENTA e *Google Desktop* (em 6.2.).

6.1. Conclusões da Resolução de Anáforas

A primeira conclusão é a de que a escolha da metodologia do *Centering* foi uma escolha adequada para a análise de textos em língua portuguesa, pois apresenta uma taxa de sucesso bastante razoável (ver tabela 10). A aplicação desta teoria justifica-se pela quantidade mínima de informação sintáctica e semântica que é necessária para perceber o elemento central, em cada momento. Constata-se, portanto, que não é uma teoria específica de uma determinada língua, mas que é perfeitamente aplicável à língua portuguesa.

Outra conclusão a que se pode chegar é a de que, apesar das poucas exigências do método utilizado, as quais facilitam a sua implementação, este modelo pode revelar-se pouco eficiente quando aplicado a determinados corpora. Ou seja, um corpus com uma linguagem específica e frases estruturalmente complexas, de que é exemplo, o caso do corpus jurídico, aqui estudado. Os resultados obtidos, ao nível do insucesso podem ser justificados, em parte, pela natureza desse corpus e também pela qualidade da análise morfo-sintáctica utilizada. Esta análise falhou em alguns aspectos, sendo os mais relevantes, os seguintes:

- **Enunciados longos e complexos**: perante este tipo de enunciados, a análise não foi completa, pois ignorou 387 fragmentos de texto só no corpus de teste jurídico.

- **Marcas e numerações**: encontram-se distribuídas, ao longo dos textos, anotações, em forma de marcas e números, que não respeitam as regras de construção frásica (por exemplo: "*...xl; (22)*").
- **Dificuldade em reconhecer o número do sintagma nominal**: o sinal de plural "(s)" que se pode encontrar nos sintagmas nominais não são reconhecidos pelo *parser* com esse valor (por exemplo: "*A designação da (ou das) autoridade(s) central(is) receptora(s) é de fundamental importância para a implementação do mecanismo que se pretende estabelecer através do Acordo, na medida em que são elas que...*").
- **Dificuldade em agregar sintagmas nominais ligados por conjunções**: a análise não consegue associar sintagmas nominais que estão ligados por conjunções, o que impede que vários sintagmas nominais no singular sejam referidos por um pronome no plural (por exemplo: "*No artigo 3º prevê-se uma dupla garantia: a da origem da transmissão e a garantia da confidencialidade da transmissão. Para as acautelar...*").
- **Dificuldade em reconhecer o valor das vírgulas que coordenam os sintagmas nominais**: o *parser* não conseguiu extrair correctamente os sintagmas nominais que compõem um sintagma nominal composto (por exemplo: "*...a excepção introduzida no artigo 4, nº 1, da Lei nº 9/90, de 1 de Março...*").
- **Dificuldade em reconhecer expressões preposicionais e verbais**: a análise nem sempre reconhece este tipo de expressões que contêm nomes e que, por isso, se tornam candidatos para a solução (por exemplo: *no âmbito de; dar conta*).
- **Citações em língua estrangeira**: ocorrem várias citações que são incorrectamente identificadas como pertencendo à língua portuguesa (por exemplo: "*La Police secrète du Roi Louis XIV...*") e, em menor número, expressões em latim (por exemplo: *ipsis verbis*).

Ainda no que diz respeito ao insucesso, para além dos obstáculos encontrados no pré-processamento, há que realçar outra questão sobre o pronome **lhe(s)**, que constituiu o principal motivo de insucesso na resolução (41 casos no total de 99 que não foram

resolvidos, conforme a tabela 11). Este resultado já era esperado, pois o pronome **lhe(s)** não impõe restrições de género aos seus possíveis antecedentes, dificultando a sua escolha.

Outro caso que deve ser considerado como um aspecto a melhorar é o facto de quase todas as soluções serem classificadas com um tipo de transição: *shifting* (ver tabela 13), o que, na prática, implica que a escolha da melhor solução seja feita muitas vezes pelo primeiro sintagma nominal compatível anterior.

Como trabalhos futuros, poderá ser feita uma análise detalhada sobre os resultados das tabelas 11 e 12, que apresentam a ocorrência de casos mal resolvidos. Há também a possibilidade de melhorar as restrições aplicadas aos pronomes pessoais, de forma a eliminar casos que não interessam. Poder-se-á ainda aperfeiçoar o filtro aplicado aos sintagmas nominais, com a finalidade de reduzir ainda mais o número de candidatos. Outro trabalho possível será encontrar uma melhor solução para seleccionar um de entre dois ou mais candidatos para a mesma referência anafórica pronominal, quando estes têm a mesma classificação.

Existe ainda a possibilidade de estender o domínio dos candidatos a antecedente, que até agora tinham sido apenas sintagmas nominais, de modo a abranger adjectivos, formas verbais e outros.

6.2. Conclusões do Processo de Recuperação de Informação

A conclusão mais imediata que se pode retirar, após a análise dos resultados, é que a integração da resolução de anáforas pronominais em sistemas de recuperação de informação, nomeadamente, o SENTA e o *Google Desktop*, influencia efectivamente as suas respostas.

No entanto, fica por esclarecer se esta alteração de resposta implica um ganho de precisão. Intuitivamente, parece que haveria razões para que isso acontecesse, mas após a análise efectuada por um grupo de 4 pessoas, sobre o possível ganho de relevância nos novos resultados, nada se pode concluir. Estes resultados são inconclusivos e não mostram claramente uma tendência para o aumento da precisão, ou para a sua perda (veja-se a figura 6).

Analisando mais profundamente os documentos lidos por aquele grupo de pessoas, cujas opiniões permitiriam responder à questão de ganho de precisão, foi possível observar que os temas pesquisados não correspondiam aos assuntos mais tratados nesses documentos. Desta forma, aumentou-se bastante a dificuldade de análise ao ponto de, em textos de várias páginas, haver apenas uma ou duas referências localizadas, o que não permite, claramente, obter uma resposta objectiva de ganho ou perda de precisão.

Como trabalhos futuros, poderão fazer-se mais análises de opinião, sobre a variação de precisão nas respostas dadas, pelos dois sistemas de recuperação de informação que aqui integram a resolução de anáforas pronominais, ou mesmo, alargar este estudo a mais sistemas de RI.

Referências Bibliográficas

[Aires *et al* (2004)] Aires, A. M., Coelho, J. C. B., Collovini, S., Quaresma, P. & Vieira, R. (2004). *Avaliação de centering em resolução pronominal da língua portuguesa.* In *Taller de Herramientasy Recursos Linguísticos para el Español y el Português (Iberamia).*

[Aone & Bennett (1995)] Aone, C. & Bennett, S. (1995). *Evaluating automated and manual acquisition of anaphora resolution rules.* In Proceedings of the 33th Annual Meeting of the Association for Computational Linguistics (ACL '95), pages 122–129.

[Baldwin (1997)] Baldwin, F. B. (1997). *CogNIAC High Precision Coreference with Limited Knowledge and Linguistics Resources.* In Proceedings of the ACL'97/EACL'97 WORKSHOP ON Operational Factors in Practical, Robust Anaphora Resolution, 38-45, Madrid, Spain.

[Baeza-Yates & Ribeiro-Neto (1999)] Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval.* ACM Press: New York.

[Bick (2000)] Bick, E. (2000). *The Parsing System PALAVRAS: Automatic Grammatical Analysis of Portuguese. Constraint Grammar Framework.* Ph.D. Thesis, Aarhus University Press.

[Borst (1997)] Borst, W. (1997). *Construction of Engineering Ontologies.* Ph.D. Thesis, University of Tweenty. Enschede, The Netherlands - Centre for Telematica and Information Technology.

[Brennan *et al* (1987)] Brennan, Susan E.; Friedman, Marilyn Walker; & Pollard, Carl J. (1987). *A centering approach to pronouns.* In Proceedings, 25th Annual Meeting of the Association for Computational Linguistics, 155-162, Stanford, California, USA.

[Brin & Page (1998)] Brin S., Page L. (1998). *The anatomy of a large-scale hypertextual web search engine.* WWW7 / Computer Networks, 30(1-7): 107-117.

[Chomsky (1981)] Chomsky, N. (1981). *Lectures on Government and Binding.* The Hague: Mouton.

[Coelho (2005)] Coelho, T. (2005). *Resolução de anáfora pronominal em português utilizando o algoritmo de Lappin e Leass.* Dissertação de Mestrado, Universidade Estadual de Campinas.

[Cunha & Cintra (1991)] Cunha, C. & Cintra, L. (1991). *Nova Gramática do Português Contemporâneo.* 8ª Edição. Edições João Sá da Costa: Lisboa.

[Deoskar (2004)] Deoskar, T. (2004). *Techniques for Anaphora Resolution: A Survey.* CS 674. 5/17/2004.

[Dias *et al* (1999)] Dias, G., Guilloré, S. & Lopes, J. (1999). *Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text Corpora.* In: Pascal Amsili and Phileppe Blache (eds), 6th. Annual conference on Traitement Automatique des Langues Naturelles (TALN99), 333-338. Cargése, France.

[Eckert & Strube (2000)] Eckert M. & Strube, M. (2000). *Dialogue Acts, Synchronising Units and Anaphora Resolution.* Journal of Semantics 17, 51-89.

[Frakes & Baeza-Yates (1992)] Frakes, W. & Baesa-Yates, R. (1992). *Information Retrieval: Data Structures & Algorithms.* Prentice Hall, Upper Saddle River, New Jersey, USA.

[Gasperin *et al* (2003)] Gasperin, C., Vieira, R., Goulart, R. and Quaresma, P. (2003). *Extracting XML Syntactic Chunks from Portuguese Corpora.* Proceedings of the Workshop TALN 2003 Natural Language Processing of Minority Languages and Small Languages. Batz-sur-Mer, France.

[Ge *et al* (1998)] Ge, N., Hale, J., and Charniak, E. (1998). *A statistical approach to anaphora resolution.* In Proceedings of the 6th Workshop on Very Large Corpora, 161-170.

[Gil (2002)] Gil, A. (2002). *Extracção eficiente de padrões textuais utilizando algoritmos e estruturas de dados avançadas.* Dissertação de Mestrado, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa.

[Grosz *et al* (1986)] Grosz, B.J., Joshi, A.K. & Weinstein, S. (1986). *Towards a computational theory of discourse interpretation.* Preliminary draft.

[Grosz *et al* (1995)] Grosz, B. J., Weinstein, S. & Joshi, A. K. (1995). *Centering: A framework for modelling the local coherence of discourse. Association for Computational Linguistics*, 21(2):203–225.

[Halliday & Hasan (1976)] Halliday, M., & Hasan, R. (1976). *Cohesion in English.* Longman English Language Series 9. London: Longman

[Hankamer & Sag (1976)] Hankamer, J. & Sag, I. (1976). *Deep and surface anaphora.* Linguistic Inquiry 7.3:391-428.

[Hirst (1981)] Hirst, G. (1981). *Anaphora in Natural Language Understanding.* Berlin: Springer-Verlag.

[Hobbs (1978)] Hobbs, J. (1978). *Resolving pronoun references.* Lingua, 44:311-338.

[Kibble (2001)] Kibble, R. (2001). *A Reformulation of Rule 2 of Centering Theory.* Computational Linguistics, 27(4), 579-587.

[Lappin & Leass (1994)] Lappin, S., & Leass, H.J. (1994). *An algorithm for pronominal anaphora resolution.* Computational Linguistics, 20(4), 535-561.

[Lappin & McCord (1990a)] Lappin, S. & McCord, M. (1990a). *A Syntactic Filter on Pronominal Anaphora in Slot Grammar.* In 28th Annual Meeting of the Association for Computational Linguistics, 135-142.

[Lappin & McCord (1990b)] Lappin, S. & McCord, M. (1990b). *Anaphora Resolution in Slot Grammar.* Computational Linguistics, 16(4): 197-212.

[Market *et al* (2003)] Market, K., Nissin, M. & Modjeska, N. (2003). *Using the web for anaphora resolution.* In EACL Workshop on the Computational Treatment of Anaphora, Budapest, Hungary.

[Maron & Kuhns (1960)] Maron, M. & Kuhns, J. (1960). *On relevance, probabilistic indexing, and information retrieval.* Journal of the ACM, 7(3):216-244.

[Mitkov (1998)] Mitkov, R. (1998) *Robust pronoun resolution with limited knowledge.* Proceedings of the 18.th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference, 869-875, Montreal, Canada.

[Mitkov (2000)] Mitkov, R. (2000). *Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems.* In Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC2000), pages 96 – 107, Lancaster, UK.

[Mitkov (2002)] Mitkov, R. (2002). *Anaphora Resolution.* London: Longman, Pearson Education.

[Paice (1984)] Paice, C. (1984). *Soft evaluation of boolean search queries in information retrieval systems.* Information Technology: Research and Development, 3(1), 33-42.

[Poesio *et al* (2000)] Poesio, M., Cheng, H., Henschel, R., Hitzeman, J., Kibble, R. & Stevenson, R. (2000). *Specifying the parameters of centering theory: a corpus-based evaluation using text from application-oriented domains.* In ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, 400-407, Morristown, New Jersey, USA.

[Rijsbergen (1979)] Rijsbergen, C. (1979). *Information Retrieval.* Butterworths, London.

[Salton & McGill (1983)] Salton, G. & McGill, M. (1983). *An Introduction to Modern Information Retrieval.* McGraw-Hill: New York.

[Sidner (1979)] Sidner, C. (1979). *A Computational Model of Co-Reference Comprehension in English.* Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

[Strawson (1950)] Strawson, P. (1950). *On referring.* Mind, 59(235):320–344. Oxford University Press.

[Strube (1998)] Strube, M. (1998). *Never look back: an alternative to centering.* In Proceedings of the 17th International Conference on Computational Linguistics (COLING' 98/ACL'98), Montreal, Canada.

[Tetreault (2001)] Tetreault, J. R. (2001). *A Corpus-based Evaluation of Centering and Pronoun Resolution.* Computational Linguistics, 27(4): 507-520.

[Walker (1998)] Walker, M. (1998): *Centering, Anaphora Resolution and Discourse Structure.* In M. Walker, A. Joshi, & E. Prince (eds.), Centering in Discourse. Oxford University Press.

Anexo A – Exemplo do Corpus Jurídico

Documento 25

O exercício de salto em pára-quedas de uma aeronave em voo corresponde a um tipo de actividade com risco agravado enquadrável no nº 4 do artigo 2º, referido ao nº 2 do artigo 1º, ambos do Decreto-Lei nº 43/76, de 20 de Janeiro;
O acidente de que foi vítima o Sargento Pára-quedista Número de Identificação Militar, em 18 de Fevereiro de 1991, verificou-se em circunstâncias subsumíveis ao quadro descrito na conclusão anterior.
SENHOR SECRETÁRIO DE ESTADO DA DEFESA NACIONAL, EXCELÊNCIA:
Para emissão do parecer a que se refere o nº 4 do artigo 2º do Decreto-Lei nº 43/76, de 20 de Janeiro, determinou Vossa Excelência o envio à Procuradoria-Geral da República do processo respeitante ao 1º Sargento Pára-quedista Número de Identificação Militar.
Cumpre satisfazer o solicitado.
A matéria de facto disponível, constante do processo de averiguações por acidente em serviço, pode ser assim condensada:
No dia 18 de Fevereiro de 1991, pelas 15.45 horas, quando efectuava a aterragem de um salto de pára-quedas, na zona de lançamento do Arripiado, o militar em referência sofreu um acidente;
Tal acidente resultou de embate violento no solo, o qual foi provocado por súbitas e fortes rajadas de vento;
O sinistrado estava devida e legalmente nomeado para a missão em que se produziu o acidente;
O acidente foi considerado em serviço, concluindo-se não ter existido responsabilidade do sinistrado ou de outrem na sua ocorrência;
Nas conclusões do parecer técnico a propósito elaborado, pode ler-se, com interesse para o apuramento das circunstâncias em que o acidente ocorreu, o seguinte:
As causas do acidente escaparam ao domínio técnico e físico do sinistrado, antes se enquadrando naquelas situações de risco imputáveis à imprevisibilidade das condições meteorológicas";
O acidente verificado deve-se à instabilidade meteorológica, mais propriamente às súbitas rajadas que naquele momento se fizeram sentir";

As rajadas de vento aumentaram significativamente a velocidade horizontal do conjunto calote - pára-quedista";

O aumento da velocidade horizontal, conjugado com a pesada mochila e saco de armas já suspensos, conduziu a uma atitude pendular de difícil controlo por parte do pára-quedista";

O sinistrado embateu no solo com velocidade superior ao normal";

O embate foi agravado pelas irregularidades e dureza do terreno";

O sinistrado agiu correctamente e a sua acção atempada nas tiras de suspensão evitou a ocorrência de danos pessoais de gravidade muito superior, eventualmente fatais".

Como consequência do acidente resultaram para o requerente lesões integradas num quadro de lombo-ciatalgias por lesões discais, tendo sido operado por duas vezes (em 12 de Setembro de 1991 e em 4 de Fevereiro de 1993) a hérnia discal em L5-S1 ;

Em 15 de Julho de 1993, a Junta de Saúde da Força Aérea considerou que o requerente padece de "hérnia discal recidiva da L5-S1, à esquerda", com afectação das funções da coluna vertebral em grau incompatível com todo o serviço militar, tendo, em consequência, emitido o seguinte parecer:

Incapaz de todo o serviço.

Apto para trabalhar e para angariar meios de subsistência";

Presente a exame de sanidade em 16 de Novembro de 1993, entenderam os peritos médicos que se encontrava clinicamente curado, tendo-lhe resultado do acidente 46 dias de incapacidade total e 405 dias de incapacidade parcial, e sendo-lhe atribuído um coeficiente de desvalorização funcional de 45% (quarenta e cinco por cento) segundo a Tabela Nacional de Incapacidade ATDP;

Analisado o processo em 24 de Fevereiro de 1994, na Direcção de Saúde da Força Aérea, os peritos médicos concordaram com o coeficiente de desvalorização de 0,45, mais entendendo haver relação das lesões com o acidente e o serviço, posição que mereceu despacho concordante do Director de Saúde da Força Aérea Portuguesa, de 28 de Fevereiro;

Em parecer do Serviço de Justiça e Disciplina de 2 de Março de 1994, foi entendido que o acidente deve ser considerado em serviço e as lesões resultantes do acidente.

Dispõe o nº 2 do artigo 1º do Decreto-Lei nº 43/76, de 20 de Janeiro:

É considerado deficiente das forças armadas portuguesas o cidadão que:

No cumprimento do serviço militar e na defesa dos interesses da Pátria adquiriu uma diminuição na capacidade geral de ganho;

Quando em resultado de acidente ocorrido:

Em serviço de campanha ou em circunstâncias directamente relacionadas com o serviço de campanha, ou como prisioneiro de guerra;

Na manutenção da ordem pública;

Na prática de acto humanitário ou de dedicação à causa pública;

No exercício das suas funções e deveres militares e por motivo do seu desempenho, em condições de que resulte, necessariamente, risco agravado equiparável ao definido nas situações previstas nos itens anteriores;

Vem a sofrer, mesmo a posteriori, uma diminuição permanente, causada por lesão ou doença, adquirida ou agravada, consistindo em:

Perda anatómica;

Prejuízo ou perda de qualquer órgão ou função;

Tendo sido, em consequência, declarado, nos termos da legislação em vigor:

Apto para o desempenho de cargos ou funções que dispensem plena validez;

Incapaz do serviço activo;

Incapaz de todo o serviço militar".

E o artigo 2º, nº 1, alínea b):

Para efeitos da definição constante do nº 2 do artigo 1º deste decreto-lei, considera-se que:

É fixado em 30% o grau de incapacidade geral de ganho mínimo para o efeito da definição de deficiente das forças armadas e aplicação do presente decreto-lei».

Os nºs 2, 3 e 4 do mesmo artigo 2º estabelecem:

O "serviço de campanha" ou "campanha" tem lugar no teatro de operações onde se verifiquem operações de guerra, de guerrilha ou de contra-guerrilha e envolve as acções directas do inimigo, os eventos decorrentes de actividade indirecta de inimigo e os eventos determinados no decurso de qualquer outra actividade terrestre, naval ou aérea de natureza operacional.

As "circunstâncias directamente relacionadas com o serviço de campanha" têm lugar no teatro de operações onde ocorram operações de guerra, guerrilha ou de contra-guerrilha e envolvem os eventos directamente relacionados com a actividade operacional que pelas suas características impliquem perigo em circunstâncias de contacto possível com o inimigo e os eventos determinados no decurso de qualquer outra actividade de natureza operacional, ou em actividade directamente relacionada, que pelas suas características próprias possam implicar perigosidade.

O exercício de funções e deveres militares e por motivo do seu desempenho, em condições de que resulte, necessariamente, risco agravado equiparável ao definido nas situações previstas nos itens anteriores", engloba aqueles casos especiais, aí não previstos que, pela sua índole, considerado o quadro de causalidade, circunstâncias e agentes em que se desenrole, seja identificável com o espírito desta lei (redacção rectificada no Diário da República, I Série, 2º Suplemento, de 26/6/76).

A qualificação destes casos compete ao Ministro da Defesa Nacional, após parecer da Procuradoria-Geral da República".

Este corpo consultivo tem interpretado as disposições conjugadas dos artigos 1º, nº 2, e 2º, nº 4, do Decreto-Lei nº 43/76 no sentido de que o regime jurídico dos deficientes das Forças Armadas, para além das situações expressamente contempladas no primeiro preceito - de serviço de campanha ou em circunstâncias com ela relacionadas, de prisioneiros de guerra, de manutenção da ordem pública e de prática de acto humanitário ou de dedicação à causa pública -, só é aplicável aos casos que, «pelo seu circunstancialismo, justifiquem uma equiparação, em termos objectivos, àquelas situações de facto, dado corresponderem a actividades próprias da função militar ou inerentes à defesa de altos interesses públicos, importando sujeição a um risco que excedendo significativamente o que é próprio do comum das actividades castrenses, se mostra agravado em termos de se poder equiparar ao que caracteriza aquelas situações paradigmáticas».

Assim implica esse regime não só que o acidente tenha ocorrido em serviço, mas também que a actividade militar que o gera envolva, por sua natureza, objectiva e necessariamente, um risco agravado em termos de poder equiparar-se ao que decorre em situações de campanha ou a elas por lei igualadas» .

Este Conselho Consultivo tem vindo a entender que o risco inerente ao salto em pára-quedas de uma aeronave surge agravado relativamente ao comum das actividades castrenses, em termos de permitir a sua equiparação abstracta a qualquer das outras actividades directamente contempladas na lei .

Na generalidade dos casos, os acidentes vêm descritos segundo uma tipicidade própria que aponta para a relevância do risco, designadamente porque se mostram observadas as regras técnicas e de segurança, ausência de culpabilidade do sinistrado ou de outrem, intromissão no processo causal de factores condicionantes ou agravantes (fortes e súbitas rajadas de vento, dificuldades na abertura do pára-quedas ou «enganche» noutros).

Estes factores aparecem de tal modo ligados ao processo causal normal, típico, que não podem ser considerados imprevistos ou ocasionais.

É também este o quadro em que se deve situar o caso dos autos que, por isso, não pode deixar de entender-se que configura uma situação de risco agravado.

Conclusão:

Termos em que se extraem as seguintes conclusões:

O exercício de salto em pára-quedas de uma aeronave em voo corresponde a um tipo de actividade com risco agravado enquadrável no nº 4 do artigo 2º, referido ao nº 2 do artigo 1º, ambos do Decreto-Lei nº 43/76, de 20 de Janeiro;

O acidente de que foi vítima o Sargento Pára-quedista Número de Identificação Militar, em 18 de Fevereiro de 1991, verificou-se em circunstâncias subsumíveis ao quadro descrito na conclusão anterior.

Dos pareceres nºs 55/87, de 29 de Julho de 1987, e 80/87, de 19 de Novembro de 1987, homologados mas não publicados, e reflectindo orientação uniforme desta instância consultiva.

Confira também os pareceres nºs 10/89, de 12-04-89, e 89/90, de 06-12-90 . Confira parecer nº 33/86, de 29-07-87, homologado, e outros aí citados, por exemplo, pareceres nºs 4/80, de 07-02-80, 86/81, de 11-06-81, 147/81, de 22-10-81, 219/81, de 04-03-82, 42/82, de 01-04-82, e 6/86, de 27-02-86, não publicados.

Vejam-se ainda, por mais recentes, os seguintes pareceres:

nº 44/89, de 11-05-89;

nº 25/90, de 12-07-90;

nº 89/90, já citado;

nº 89/91, de 30-01-92;

nº 24/92, de 09-07-92;

nº 12/93, de 01-94-93;

e nº 24/93, de 20-04-93 .

Anexo B – Exemplo do Corpus Jornalístico

Documento 113

Guilherme Inês*.

O seu percurso musical começa pelos grupos pop, prolonga-se pelas sessões de estúdio e culmina na produção.

A partir do momento em que um gajo começa a fazer estúdio, o meu interesse passou de um instrumento para a possibilidade de poder ter uma visão mais global e aberta do universo das gravações. A mudança teve início na gravação de «Se Cá Nevasse», dos Salada de Frutas. A partir daí passei a entrar mais na área da produção. «No segundo disco da banda, «Crime Perfeito», entrei um bocadinho ainda mais. Mas continuo a ser músico, a tocar bateria, guitarra, piano. No último disco da Dulce, tocámos praticamente os instrumentos todos.

Enquanto músico e produtor, quais são as suas preferências?.

O meu «background» tem duas vertentes: a música popular portuguesa e o rock, com letra maiúscula. Hoje em dia o que eu gosto de ouvir está um bocado ligado às músicas alternativas e aquilo a que se poderá chamar «world music». Coisas que até há pouco nem sabiam que existiam, música dos pigmeus do Gabão, um basco chamado Tomás San Miguel, um tipo vai chegando à conclusão que neste momento há um planeta, uma série de pessoal que aparentemente não está relacionado com nada mas está no mesmo comprimento de onda, a fazer trabalhos de fusão de culturas. Quando se procura as próprias raízes, vai-se encontrar as raízes dos outros. Quanto mais fundo se vai, mais para cima se vai. Chegando ao Peter Gabriel, para mim o fulano que faz música mais consensual.

Transpõe esses gostos para o trabalho de produção ou aceita todas as solicitações de trabalho, pondo de lado essas mesmas preferências?.

No esquema de produção mais recente, apenas produzi o disco da Dulce Pontes, «Lágrimas». A minha outra área de trabalho é a publicidade. No caso da Dulce, houve à partida uma grande identificação entre os dois, para onde é que queríamos ir. À partida, quando um artista escolhe um produtor, fá-lo porque reconhece no trabalho dele qualquer coisa que lhe diz respeito.

Além de Dulce Pontes, também já produziu um disco da Dora. Para quem diz situar-se perto das músicas alternativas não acha um paradoxo?.

Mas também fiz um trabalho com a Lena d'Água, sobre temas do António Variações, um disco que passou completamente ao lado das pessoas mas onde já havia um desvio para essa área. E quando digo músicas alternativas não estou a dizer que elas não sejam comerciais. O que decididamente não me interessa são coisas como a «house», a música de dança ou, na generalidade, a dos tops. Não a ouço, não tenho discos, não me interessam enquanto área de trabalho. Interessa-me cada vez mais uma área onde possa pesquisar, fazer coisas que ainda não fiz. Por outro lado, tenho neste momento um projecto para um disco a solo, algo que tenho na cabeça há dez anos, sobretudo desde que andei um ano e meio em digressão com José Afonso, uma pessoa para mim decisiva em termos de influência.

Como se processa o seu trabalho enquanto compositor e produtor de «jingles» publicitários?.

Tenho a sorte de estar em estúdio consecutivamente. Neste momento e de há dez anos para cá, todos os dias estou em estúdio. Isto permite-me ir burilando o meu próprio trabalho. Vou ouvindo muita asneira que faço. É uma escola de disciplina e de despojamento muito boa, porque de facto aquilo que vai numa peça publicitária de 30 segundos é o estritamente necessário. Nada a mais nem a menos. Ali não há espaço nem tempo a perder. Desenvolve-se um poder de síntese -- que remédio! -- e de análise grandes. E um poder de microscopia. Divide-se o segundo em 25 partes e cada uma delas é crucial. O cérebro tem uma capacidade limitada de assimilar informação, não se pode sobrecarregá-lo. Por exemplo, num filme publicitário, não pode haver excesso de informação, sob pena que a mensagem não passe. Outra coisa importante nesta área é o sentido de se trabalhar numa equipa, desde os fulanos da agência que concebem a campanha até ao texto, à música e à parte gráfica. Toda a gente trabalha para uma finalidade.

Não existe o perigo de o artista se transformar num simples técnico?.

Esse risco existe. Há vezes em que tenho liberdade de criação quase total e outras em que há grandes restrições. Aí vem o factor disciplina ao de cima. Quando se está numa situação de total liberdade, vamos imaginar um halterofilista que treina com pesos de 40 quilos e de repente lhe atiram com um cinzeiro que pesa 200 gramas. Para o segurar nas mãos, veja lá a agilidade que ele tem! É um pouco isto. É um pouco potenciar toda a energia que está acumulada.

Em estúdio e enquanto produtor, já lhe aconteceu entrar em conflito com os músicos? De que maneira lida com essas situações?.

Lido mal. Para já não gosto de conflitos. Se calhar é por ser um bocado preguiçoso. Geralmente prefiro ceder. Depois, há artistas que são improduzíveis, conheço dois, que se produzem a si próprios. Não vale a pena tentar o nosso contributo. Esses artistas não deviam contratar produtores. O produtor para eles é uma estátua que está ali para pôr o nome no disco: «Produzido por». Ora eu quando ponho «produzido por», gosto de sentir e ouvir que está lá alguma coisa minha. Que há uma responsabilidade minha. Se for mau, é mau; se for bom, é bom. Eu sou o trabalho que faço. Por exemplo, no disco da Dulce, ouço-me lá. Sou uma pessoa com grande tendência para a nostalgia. Não para a tristeza. Nem é saudosismo mas uma certa nostalgia que me liga a coisas como o amanhecer num rio, como o Zêzere, o cheiro dos eucaliptos.

Referiu há pouco que está quase permanentemente em estúdio. Não sente necessidade do silêncio? De parar?.

Sim. Então quando chega o Verão!... Todos os anos, felizmente, há períodos de paragem. Quando eu digo não parar, é sobretudo mentalmente, não ficar desligado. Embora haja alturas em que tenho que desligar e pôr uma folha em branco à frente. Comer um marisco, olhar para o mar, nadar... Costumo fazer isto quando vou para Ferreira do Zêzere. Vou limpando as baterias.

O termo «new age» diz-lhe alguma coisa?.

Diz. Englobo a «new age» na «world music», embora num outro plano, mais sensorial e impressionista.

Diga o nome de produtores que considere revolucionários.

Brian Eno... na criação de sinergias entre a pessoa e o que ela está a fazer. Umas vezes é a pessoa que puxa a criatividade, noutras é aquilo que se faz que puxa a pessoa. É esse o sentido do erro e do aproveitamento desse erro. Ir atrás do erro e interagir com ele. Malcolm McLaren não me diz grande coisa. Phil Spector, um gajo que criou um som. Há coisas que mal se ouvem e vê-se logo que é Phil Spector. O «wall of sound» e aquelas cenas todas. George Martin, com os Beatles. Grande profissional, ainda por cima lutando contra grandes dificuldades tecnológicas. Peter Gabriel, em termos de concepção. Há um tipo que fez completamente discos de produtor que é o Trevor Horn, que trabalhou com os Frankie Goes to Hollywood e foi teclista dos Yes. Em termos de manipulação tecnológica, é um tipo perfeitamente pop.

Os Kraftwerk e a sua noção do estúdio como instrumento musical?.

Interessante. Têm uma perspectiva curiosa que é não rejeitar a tecnologia, assumi-la a cem por cento e humanizá-la, perspectivá-la e dá-las às pessoas no seu lado humano. Eu falo com as máquinas com que trabalho. Não estou a brincar. Vou ter com o «Fairlight» e digo-lhe «hoje estás mal disposto!».

Não sente a angústia de ter que escolher entre infinitas possibilidades de criação postas à sua disposição num estúdio?.

Há o factor da criatividade e sensibilidade próprias. Aí reajo absolutamente por instinto. Em geral, primeiro ouço o som e depois é que vou à procura dele. Imagine que olha para uma parede em branco e «vê» lá um quadro. Depois de «ver» o quadro é que o vai pintar. Não é o contrário. O fundamental é que o que se ouve esteja correcto com o instinto e as emoções do momento. Como nas fotografias. Quando se tira uma fotografia não se pode voltar atrás. É um paralítico no tempo. Esse segundo, essa fracção, não existe mais. Nunca. Há alguns engenheiros de som que dizem que sou uma pessoa um bocado ansiosa, porque acho que determinadas coisas têm que ser feitas depressa. Para se aproveitar o jorro criativo. As máquinas têm de estar ali para nos servir. Como escravas. O que eu procuro é captar as magias, as faíscas que saltam em determinado momento. Isso é que tem de ficar gravado.

Fernando Magalhães.

* Músico e produtor. Fez parte, nos anos 60 e 70, de grupos como os Chinchilas, Objectivo, Zoom e Salada de Frutas. Tocou ao vivo e como músico de estúdio, entre outros, com José Afonso, Vitorino, Fausto e Sérgio Godinho. Recentemente produziu o álbum a solo de Dulce Pontes, «Lágrimas».

Anexo C – Outras Tabelas

Documento	Pronomes	Pronomes bem resolvidos	Precisão
2	29	19	0,655
3	21	12	0,571
6	84	39	0,464
7	14	7	0,5
8	40	29	0,725
16	62	45	0,726
Total	250	151	0,604

Tabela 21 – Cálculo da Precisão nos documentos do corpus de teste.

Documento	Pronomes (correctos e incorrectos)	Pronomes bem resolvidos	Cobertura
2	32	19	0,594
3	22	12	0,545
6	86	39	0,453
7	14	7	0,5
8	42	29	0,69
16	63	45	0,714
Total	259	151	0,583

Tabela 22 – Cálculo da Cobertura nos documentos do corpus de teste.

Documento	Precisão	Cobertura	Medida-F_1
2	0,655	0,594	0,623
3	0,571	0,545	0,558
6	0,464	0,453	0,459
7	0,5	0,5	0,5
8	0,725	0,69	0,707
16	0,726	0,714	0,720
Total	0,604	0,583	0,593

Tabela 23 – Cálculo da Medida-F_1 nos documentos do corpus de teste.

Documentos Jurídicos	Expressão Nominal Pesquisada	Documentos por ordem de relevância									
Originais	associação	3	11	9	31	26	20	8	23		
Após substituições		3	11	9	31	26	20	8	24	23	
Originais	autorização	9	39	23	20	34	30	28	13	21	7
Após substituições		9	39	34	23	20	13	30	28	7	36
Originais	Constituição	9	6	39	38	34	3	30	26	23	21
Após substituições		9	6	39	38	34	30	3	26	23	21
Originais	deliberação	8	4	26	20	3	34	22	21		
Após substituições		8	4	26	20	3	34	22	21		
Originais	Europol	31	30								
Após substituições		31	30								
Originais	lei	9	8	7	6	4	39	37	38	36	35
Após substituições		9	8	7	6	4	39	38	37	36	35
Originais	norma	6	39	37	38	3	21	13	11	7	34
Após substituições		6	39	38	37	3	21	13	11	7	34
Originais	proposta	8	4	39	26	20	13	34	3	30	29
Após substituições		8	4	39	26	20	2	13	34	30	3
Originais	Universidade	34	26								
Após substituições		34	26								
Originais	certidões	22	2	8	20	13					
Após substituições		22	2	8	20	13					
Originais	direito	9	6	38	30	28	22	15	7	4	39
Após substituições		9	6	38	30	28	22	15	7	4	39
Originais	Estado	6	35	31	3	26	7	4	38	36	30
Após substituições		6	35	31	3	7	4	38	36	30	29
Originais	Ministério Público	20	37	35	34	22	38	2	7	36	39
Após substituições		20	37	35	34	22	38	36	2	7	39
Originais	parecer	36	34	29	24	2	14	15	13	8	7
Após substituições		36	34	29	24	2	14	15	35	13	8
Originais	património	6	37	26	20	7	35	21			
Após substituições		6	37	26	24	20	7	35	21		
Originais	processo	35	22	2	8	30	29	39	37	38	36
Após substituições		35	30	22	2	29	39	38	37	36	28
Originais	requerente	9	4	28	2	37	13	35	25	20	31
Após substituições		4	28	2	9	37	22	13	35	25	20
Originais	tribunal	4	39	28	16	10	9	38	35	34	29
Após substituições		4	39	28	16	10	9	38	35	34	29
Originais	agentes públicos	7	3	9	21	31	6	36	12	34	25
Após substituições		7	3	9	21	31	6	36	12	34	25
Originais	Estados-Membros	31	29	28	2	16	10	30	23	8	20
Após substituições		31	29	28	2	16	10	30	23	8	20

Tabela 24 – Resultados das pesquisas efectuadas sobre o corpus de trabalho jurídico.

Documentos Jornalísticos	Expressão Nominal Pesquisada	Documentos por ordem de relevância									
Originais	acções	160	159	152	153	151	143	158	148	29	26
Após substituições		160	159	152	153	151	143	158	148	29	26
Originais	banco	160	159	156	154	152	151	149	146	144	143
Após substituições		160	159	156	154	152	151	149	146	144	143
Originais	Banesto	160	159	158	154	153	152	151	148	149	143
Após substituições		160	159	158	156	154	152	153	150	151	149
Originais	construção	52	58	90	62	51	38	31	139	131	117
Após substituições		90	52	58	62	51	38	31	139	131	117
Originais	dia	100	97	58	127	67	4	148	114	27	153
Após substituições		100	97	58	127	67	4	148	114	27	153
Originais	dinheiro	153	160	108	157	57	12	152	147	97	98
Após substituições		153	108	97	160	157	57	12	152	147	98
Originais	facto	160	100	150	143	127	159	154	151	146	110
Após substituições		160	143	100	150	127	144	159	154	151	146
Originais	governo	160	70	48	159	77	73	50	156	147	89
Após substituições		160	73	70	48	159	77	50	157	156	147
Originais	homem	100	97	7	160	114	103	86	78	123	98
Após substituições		100	97	7	160	114	103	86	78	123	98
Originais	José Roquette	160	146	143	159	153	151	148	152	158	157
Após substituições		160	153	151	146	143	159	157	148	152	158
Originais	ministro	160	154	146	75	148	74	149	142	77	48
Após substituições		160	154	148	146	75	74	149	142	77	48
Originais	mundo	103	100	94	97	128	95	42	126	123	124
Após substituições		103	100	97	94	95	128	42	126	123	124
Originais	papel	100	94	67	88	157	126	122	92	84	82
Após substituições		100	94	67	88	42	157	126	122	92	84
Originais	processo	160	150	146	64	155	82	65	158	154	151
Após substituições		160	150	146	151	64	155	100	82	65	158
Originais	prova	13	159	16	155	100	12	160	149	147	101
Após substituições		13	100	159	16	155	12	160	149	147	101
Originais	registo	160	153	151	154	148	145	143	128	91	67
Após substituições		160	153	151	29	154	148	145	143	128	91
Originais	tempo	100	97	92	4	2	160	127	124	113	94
Após substituições		100	97	92	4	2	160	127	124	117	113
Originais	Totta	160	158	159	157	153	152	150	148	146	147
Após substituições		160	158	159	157	156	152	153	150	148	147
Originais	vida	100	97	103	94	4	98	124	96	93	92
Após substituições		100	97	4	103	94	98	124	96	93	92
Originais	voz	128	103	160	157	141	136	126	124	119	109
Após substituições		128	103	109	160	157	141	136	126	124	119

Tabela 25 – Resultados das pesquisas efectuadas sobre o corpus de trabalho jornalístico.

Número da Pesquisa	Expressão Nominal Pesquisada	Documento	Posição original	Posição após substituições
3	Banesto	156	> 10ª	4ª
4	construção	90	3ª	1ª
6	dinheiro	108	4ª	2ª
7	facto	143	4ª	2ª
8	governo	73	6ª	2ª
10	José Roquette	153	5ª	2ª
11	ministro	148	5ª	3ª
12	mundo	97	4ª	3ª
13	papel	42	> 10ª	5ª
14	processo	151	10ª	4ª
15	prova	100	5ª	2ª
16	registo	29	> 10ª	4ª
17	tempo	117	10ª	9ª
18	Totta	156	> 10ª	5ª
19	vida	4	5ª	3ª
20	voz	109	10ª	3ª

Tabela 26 – Pesquisas efectuadas sobre o corpus de trabalho jornalístico com alteração de *ranking*.