Categorização e Análise de Dados Não Estruturados:
O Caso dos Debates Parlamentares

por

Ana Espírito Santo

Trabalho de Projecto apresentado como requisito parcial para obtenção do grau de

Mestre em Estatística e Gestão de Informação

pelo

Instituto Superior de Estatística e Gestão da Informação
Universidade Nova de Lisboa

ÍNDICE

Resumo

Na presente dissertação, desenvolveu-se um protótipo que recorre a um programa de categorização textual (o *software Teragram TK 240*) para estudar o Diário da Assembleia da República (DAR), 1.ª Série, IX Legislatura (2002-2005). Com base na descrição das emoções dos deputados presente nos DAR, analisaram-se as reacções dos Grupos Parlamentares durante os debates parlamentares, com o intuito de compreender de que modo é que estas reflectem a articulação dos diferentes Grupos Parlamentares entre si e em relação ao Governo. Para contextualizar o modelo desenvolvido, fez-se um breve enquadramento teórico sobre os principais temas implicados, nomeadamente a categorização textual e o *text mining*.

Abstract

In the present dissertation, it was developed a prototype with the help of a Textual Categorization software (*Teragram TK 240)* to study the Portuguese Assembleia da República Diaries (DAR), 1st Series, IX Legislature (2002-2005). Having the descriptions of the reactions present in the DAR as a basis, we have analyzed the emotions of the AR Members and we have tried to understand in which way the AR Members emotions reflect the relation between the different parties represented in the AR. We have also tried to understand the relation that these parties have concerning the Govern. Finally, we have made a theoretical research about the main themes implied in this project, namely textual categorization and text mining.

Índice de Ilustrações

Índice de Tabelas

Lista de siglas e abreviaturas

AMEC	*Association for Measurement and Evaluation of Comunication*
AM	Aprendizagem Máquina
AR	Assembleia da República
ARG	*Automatic Rule Generator*
BE	Bloco de Esquerda
CDS-PP	Centro Democrático Social – Partido Popular
Cf.	confrontar
CT	Categorização de Textos
DAR	Diário da Assembleia da República
DM	*Data Mining*
FIBEP	*International Federation of Media Monitoring Companies*
GC	Governo Constitucional
GP	Grupo(s) Parlamentar(es)
IR	*Information Retrieval*
ISEGI	Instituto Superior de Estatística e Gestão da Informação
KDD	*Knowledge Discovery in Databases*
KDT	*Knowledge Discovery in Textual Databases*
KNN	k-Nearest Neighbour
LLSF	*Linear Least-squares Fit*
NB	*Naïve Bays*
NER	*Named Entity Recognition*
NLP	*Natural Language Processing*
NNet	Abordagem de Redes Neuronais
PCP	Partido Comunista Português
PE	Português Europeu
PEV	Partido Ecologista Os Verdes
PLN	Processamento de Língua Natural
PPD/PSD	Partido Popular Democrático / Partido Social Democrata
PSD	Partido Social Democrata

QREN	Quadro de Referência Estratégico Nacional
RBC	*Rule Based Categorizer*
RI	Recuperação de Informação
SAT	Sumarização Automática de Textos
SC	*Statistical Categorizer*
SCIP	*Society of Competitive Intelligence Professionals*
ss.	seguintes
SVM	*Support Vector Machine*
TDM	*Text Data Mining*
TM	*Text Mining*
TS	*Text Summarisation*
WWW	*World Wide Web*

Agradecimentos

Agradece-se à Divisão de Comunicação e Apoio Audiovisual, em particular ao Dr. Fernando Marques, a cedência dos dados relativos à IX Legislatura, imprescindíveis para a realização do protótipo.

Agradece-se ao SAS, nomeadamente a Jos Van der Velden, não apenas a cedência do *hardware* e do *software* que tornaram possível esta tese, mas também toda a disponibilidade e empenho sempre demonstrados para que a sua realização fosse possível.

Agradece-se ainda ao orientador deste trabalho, o prof. Dr. Miguel Neto, pelo incentivo e apresentação de linhas orientadoras nos momentos de maior dificuldade, e ao co-orientador, o prof. Dr. Fernando Bação, pelas óptimas sugestões que contribuíram sem dúvida para melhorar este projecto.

Por fim, agradeço ao Gonçalo e aos meus pais todo o apoio que sempre me deram.

Nota Preliminar

O presente projecto foi inicialmente desenvolvido com a expectativa de vir a ser integrado no âmbito do *NovaIntell*, projecto de parceria entre a empresa Manchete e o ISEGI[1]. Apesar de o *NovaIntell* ter recebido o financiamento solicitado, o presente projecto não foi convidado a integrá-lo. Tendo em conta estas circunstâncias, através do ISEGI, foi solicitado o apoio do SAS Portugal[2], que se revelou incansável no auxílio prestado e na cedência dos meios (*software* e *hardware*) que tornaram possível a realização deste projecto.

[1]No âmbito do programa de apoio a Projectos de I&D em co-promoção (Quadro de Referência Estratégico Nacional – QREN (http://www.qren.pt/), foi solicitado financiamento, pela empresa Manchete, em parceria com o Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa (ISEGI/UNL), para o projecto NovaIntell. Este é um projecto inovador que visa promover a criação de novos conhecimentos nas áreas do *Text Mining* e da *Competitive Intelligence* por parte da empresa Manchete e do ISEGI, instituição com competências reconhecidas na área de intervenção deste projecto, que nesse âmbito se associaram para potenciarem sinergias, bem como partilhar custos e riscos.

[2] O SAS é líder em *software* analítico e o maior fornecedor independente no mercado de business intelligence. Para mais informações, consultar http://www.sas.com/offices/europe/portugal/index.html

1 Capítulo 1

1.1 Introdução

A facilidade de acesso e de armazenamento de grandes volumes de dados é, actualmente, uma realidade incontornável. Os progressos conhecidos nas tecnologias de recolha, organização e armazenamento da informação digital; a facilidade de troca e de transmissão de dados proporcionada pelo serviço de correio electrónico; a descentralização da informação e dos dados, devido à proliferação de páginas pessoais, *weblogs* e redes sociais na *World Wide Web*, entre outros factores, contribuíram para esta realidade.

Apesar de todos estes progressos, a capacidade humana para processar informação é limitada. Como tal, a facilidade com que se acede a um conjunto de documentos actualizados sobre um determinado tema contribui para a existência de um excesso de dados que pode revelar-se prejudicial, podendo inclusivamente levar o autor de uma pesquisa a ignorar conteúdos relevantes por excesso de dados.

Para fazer face a este problema de "sobrecarga de informação" (mais detalhado no enquadramento teórico do presente trabalho), criaram-se novas áreas de investigação, com contributos de disciplinas que se situam em campos do saber dispersos e que reúnem esforços para ajudar o ser humano a dominar e tirar partido do gigantesco fluxo de informação que ele próprio criou. Estas áreas do saber, para além de serem confrontadas com o desafio representado pela sobrecarga de informação, têm a particularidade de tratarem dados "não estruturados", que, em virtude da sua heterogeneidade e natureza não previsível, dificultam a tarefa de gestão da informação.

Referimo-nos, nomeadamente, à "Recuperação de Informação", que visa recuperar documentos, informação e meta-dados a partir de grandes volumes de dados; ao "Processamento de Língua Natural" e às técnicas computacionais que lhe estão associadas, mas também à "Categorização Textual", desempenhando esta última um papel particularmente relevante neste trabalho, pois será a técnica aplicada no nosso projecto.

Como veremos, a Categorização Textual não é apenas a "atribuição automática de textos em língua natural a um conjunto de categorias pré-definidas com base no seu conteúdo", é uma tecnologia de apoio em muitas outras tarefas relacionadas com a gestão documental (tais como indexação de vocabulário controlado, filtragem textual, automatização de respostas, etc.) e apresenta ainda contributos relevantes para a disciplina que se encontra neste momento na vanguarda desta área do conhecimento, o *Text Mining*.

Não obstante, ao longo deste projecto, testaram-se sobretudo as virtualidades desta tecnologia do ponto de vista do seu automatismo na categorização de grandes conjuntos de documentos. Com recurso a um *software* de Categorização Textual (o *Teragram TK240*), foi criado um modelos de análise de um conjunto

de documentos em português europeu (os debates parlamentares). Tirando partido do formato dos documentos e das características sintáctico linguísticas dos mesmos, tentou-se extrair informação relativa às emoções que perpassam na Assembleia da República (AR).

1.2 Objectivos

O presente trabalho de projecto foi orientado com vista a dar resposta a um conjunto de questões de investigação:

- Como é que os Grupos Parlamentares se unem nas emoções manifestadas?
- Que relação existe entre a coesão na manifestação de emoções de dois Grupos Parlamentares e as suas orientações políticas (por exemplo, os partidos da esquerda e da direita aplaudem/protestam/riem/vozeiam sempre em conjunto)?
- Como é que as emoções transmitidas se articulam com o poder?
- Concretamente em relação à legislatura em análise (a IX), há diferenças significativas entre o governo liderado por Durão Barroso (XV) e o dirigido por Santana Lopes (XVI)?
- Qual o grau de isolamento dos Grupos Parlamentares quando reagem emotivamente?
- Há relação entre o número de deputados representados e a capacidade de demonstrar emoções?

1.3 Relevância deste projecto

A apresentação de uma síntese teórica de uma área de saber inovadora e em constante actualização (nomeadamente com um modesto contributo para a revisão da literatura destas matérias em português europeu, que não passa, no entanto, de um ponto de partida para um trabalho mais aprofundado), e a possibilidade de pôr à prova e explorar um software de categorização textual – o Teragram TK 240 - são dois dos elementos que, na nossa opinião, melhor representam a relevância do presente trabalho.

Por outro lado, este projecto deu lugar não apenas a uma análise sistemática dos Diários da Assembleia da República (DAR), dados pouco explorados e de indiscutível interesse público, como também a uma abordagem dos dados inovadora e nunca levada a cabo, que passa pela "quantificação"/ análise sistematizada de reacções emotivas, e que nos permitiu sondar as emoções que perpassam a AR e compreender de que forma estas reflectem o ambiente político que as enquadra.

Não menos importante, do nosso ponto de vista, é o facto de se terem aberto portas para novos caminhos de investigação, nomeadamente com o alargamento do espectro dos dados analisados, a alteração da perspectiva de análise ou o recurso a softwares mais sofistificados.

1.4 Estrutura

No segundo capítulo faz-se a revisão crítica da literatura consultada. Iniciamos com uma perspectiva global da área, necessariamente geral, debruçando-nos sobre os conceitos de "dados", "informação" e "conhecimento", e dando conta do aparecimento das disciplinas de "Recuperação de Informação", "Processamento de Língua Natural" e "Categorização Textual", contextualizando-as no âmbito da "sobrecarga de informação". Apresentamos uma breve definição e história de "Categorização Textual", numa abordagem genérica, com menção aos métodos utilizados por esta tecnologia e às suas principais aplicações. Referimo-nos ainda ao *Text Mining* como disciplina de vanguarda nesta área, elencando sucintamente as forças para o seu desenvolvimento e algumas das suas aplicações.

No capítulo três, abordamos o caso dos debates parlamentares, apresentando o projecto desenvolvido, com a descrição do software utilizado e dos dados analisados, bem como da metodologia adoptada no seu desenvolvimento.

O capítulo quarto é exclusivamente dedicado à análise dos resultados, iniciando com uma breve reflexão sobre as categorias e a representatividade dos Grupos Parlamentares na legislatura em estudo, seguindo-se depois uma análise detalhada dos resultados monopartidários e da prestação global de cada partido em cada uma das categorias consideradas.

O capítulo cinco apresenta a conclusão, onde damos conta das razões que motivaram a nossa escolha e nos ajudaram a levar este projecto a bom termo, dos principais resultados e aprendizagens retiradas deste projecto, bem como das linhas de investigação que se abrem para o futuro.

2 Capítulo 2

2.1 Enquadramento

A evolução que se tem vindo a conhecer, desde a segunda metade do século XX, no âmbito das telecomunicações, sistemas computacionais e Internet conduziu a alterações significativas em todos os domínios da vida e do conhecimento humano. Uma das componentes da 'Revolução Digital', o progresso nas tecnologias de recolha, organização e armazenamento da informação digital, levou ao aparecimento de enormes bases de dados em todos os contextos da actividade e do conhecimento humano (Bação 2007).

O desenvolvimento da Internet (enquanto conglomerado de redes de milhões de computadores) facilitou o acesso à informação e à transferência de dados digitais, nomeadamente com a disponibilização dos serviços de correio electrónico e com o desenvolvimento da World Wide Web (WWW), que permitiu a descentralização da informação e dos dados, incluindo a criação de páginas pessoais, weblogs e redes sociais[3].

As Tecnologias de Informação superaram a capacidade humana para processar, utilizar e explorar os dados armazenados - se os sistemas computacionais duplicam as potencialidades de 18 em 18 meses, de acordo com a lei de Moore[4], as capacidades de armazenamento de informação digital duplicam com o dobro da velocidade (Fayyad and Uthurusamy 2002, citado em Kloptchenko 2003: 1).

Para processar estes repositórios de dados e deles extrair informação relevante, surgiu um novo paradigma: Descoberta de Conhecimento em Bases de Dados (do inglês *Knowledge Discovery in Data Bases* - KDD) ou *Data Mining* (DM)[5], uma disciplina relativamente recente e com importância crescente devido ao crescimento exponencial dos conjuntos de dados e da necessidade de os agregar e explorar, criando informação de valor acrescentado.

---

[3] Para a clarificação dos conceitos de Internet e World Wide Web (WWW), consultei a Wikipedia.

[4] Em 1965, o co-fundador da Intel, Gordon Moore, referindo-se aos avanços da tecnologia, afirmava que "O número de transístores e resístores existentes num chip duplica a cada 18 meses" ("The number of transistors and resistors on a chip doubles every 18 months.") (http://www.answers.com/Moore%27s+law?cat=technology).

[5] A expressão "Descoberta de Conhecimento em Bases de Dados" tem vindo a ganhar cada vez mais aceitação, especialmente na área académica, como forma de designar todo o processo que medeia entre o acesso aos dados digitais até à aplicação concreta e prática do conhecimento gerado no processo. No entanto, apesar das subtis distinções, "Descoberta do Conhecimento em Bases de Dados" e *Data Mining* são utilizados, por grande parte dos autores, como sinónimos (F.L. Bação (2007). Data Mining . Lisboa, ISEGI-UNL: 4). Neste âmbito, utilizaremos preferencialmente a expressão *Data Mining*.

2.2 Dados, Informação e Conhecimento

Importa talvez fazer uma breve reflexão sobre "dados", "informação" e "conhecimento", uma vez que a clarificação destes conceitos é fundamental no contexto da “sobrecarga de informação” e das tecnologias de gestão documental e processamento automatizado de informação que lhe estão inevitavelmente associadas.

Os dados podem ser considerados "realidades factuais dispersas, que descrevem acontecimentos sem juízo de valor prévio ou desprovidos de sentido" (Santos 2004: 31), ou podem ser definidos como o "veículo de conhecimento e informação, isto é, a forma como quer o conhecimento quer a informação podem ser armazenados e transferidos" (Bação 2007: 35). Os dados passam a "informação" quando são inseridos num contexto e lhes é atribuído um significado, isto é, quando são interpretados por um receptor. Recorrendo ao exemplo apresentado em Santos (2004: 31), numa empresa, os dados avulsos de um relatório de contas transformam-se em informação quando são inseridos num contexto e lhes é atribuído um significado.

Pode considerar-se que a dicotomia informação/conhecimento se baseia no facto de o conhecimento ser mais "substancial e completo", implicando, assim, uma reflexão crítica sobre a informação. O conhecimento pode ser considerado "um patamar superior na compreensão do mundo, ao ponto de nos tornar capazes de agir sobre ele" (Santos 2004: 31).

Uma definição que ilustra esta visão de conhecimento é a apresentada por Huseman and Goodman:

> *O conhecimento é informação carregada de experiência, verdade, juízo, intuição e valores; uma combinação única que permite aos indivíduos avaliar novas situações e gerir a mudança.*
>
> (Huseman and Goodman 1999, tradução de M.I.G. Santos 2004: 31-32)

De um outro ponto de vista, pode considerar-se que esta dicotomia se baseia no facto de a informação ser descritiva (relacionando-se com o passado e o presente) e o conhecimento eminentemente preditivo (proporcionando as bases para a predição do futuro, com determinado grau de certeza, baseado na informação referente ao passado e ao presente) (Bação 2007: 35).

Duma ou doutra perspectiva, é consensual a existência de uma hierarquia entre dados, informação e conhecimento, estando o conhecimento no topo desta hierarquia.

A disponibilização de conteúdos em formato digital e a necessidade de extrair “informação” com valor acrescentado a partir de grandes conjuntos de “dados”, e dela gerar “conhecimento”, levou a que as tarefas de gestão documental, classificadas sob a designação mais genérica de Recuperação de Informação (RI), do inglês “Information Retrieval”, desempenhem um papel cada vez mais importante no campo dos sistemas de informação (Sebastiani 2002: 1).

2.3 Data Mining e Processamento de Dados Não Estruturados

Para melhor compreender as disciplinas que se focam em processamento automatizado de informação textual (e, por isso, não estruturada, como adiante veremos), detenhamo-nos brevemente no DM. Esta disciplina tem vindo a desenvolver-se com contributos de áreas disciplinares diversas[6], e, como tal, há várias definições e abordagens possíveis para esta área de conhecimento, reflectindo os interesses e a proveniência dos investigadores implicados. Uma das definições mais populares na literatura é a de Fayyad et al. 1996, segundo a qual: "*O DM é o processo não trivial de identificação de padrões válidos, inovadores, potencialmente úteis e compreensíveis nos dados*[7]*"*.

De acordo com Bação, importa sobretudo reter que "*o Data Mining, 'a extracção de informação escondida e de carácter eminentemente preditivo de grandes bases de dados' constitui uma poderosa tecnologia, com enorme potencial de crescimento, que procura traduzir dados em informação, e informação em conhecimento, que por sua vez proporciona oportunidade de agir, sobre o real, racionalmente e com propriedade."* (Bação 2007: 5). Esta disciplina conheceu um grande impulso na última década do século XX, assistindo-se a uma implementação generalizada do ponto de vista empresarial por volta de 1994 (Kloptchenko 2003: 8).

Os processos e metodologias de DM aplicam-se a dados estruturados, ou seja, a números, tabelas, linhas, colunas, atributos, etc. São dados de natureza numérica, quantificável, repetitiva e previsível. Ao contrário da análise estatística "tradicional" que efectua voluntariamente o levantamento de dados ditos "primários" (que são recolhidos com o objectivo de serem alvo de análise estatística), os dados utilizados em DM têm a particularidade de serem "secundários", o que significa que são recolhidos para outros efeitos e depois "reaproveitados" para a análise de DM (este é, por exemplo, o caso dos dados solicitados aos clientes por empresas de crédito ao consumo).

No entanto, a forma mais popular e conveniente de transmissão da informação é através de dados não estruturados (dados textuais ou não textuais, tais como imagens, cores, sons e formas), verificando-se uma tendência de crescimento de conteúdos disponíveis em formato digital, irreversível, e acentuada pelo desenvolvimento da WWW. Com efeito, estes dados desempenham um papel fundamental nas empresas.

---

[6] Contam-se, entre estas disciplinas, as Bases de Dados, a Estatística, a Visualização, a Aprendizagem Máquina ou as Ciências de Informação (F.L. Bação 2007: 3).

[7] Esta definição é citada em F. L. Bação 2007: 5: "DM is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data".

Em 1999, estimava-se que representavam cerca de 80% da informação empresarial (Tan 1999, citado em Kloptchenko 2003: 8), mas, de acordo com os mais recentes estudos de McKnight (McKnight 2005: 80), em 2005 estes valores atingiam já os 85%/90%.

As fontes dos dados não estruturados são as mais variadas, tais como mensagens de correio electrónico, relatórios técnicos ou financeiros, documentos Word, folhas de cálculo Excel, etc. Vejam-se, a título de exemplo, três fontes de dados não estruturados presentes nas organizações, que nos permitem aferir da heterogeneidade e diversidade de formatos deste tipo de dados: mensagens de correio electrónico, relatórios ou contratos. Em qualquer um dos casos, estamos perante informação de dimensão variável (tendendo as mensagens de correio electrónico a ser mais curtas), com vocabulário específico e relacionado com as áreas concretas a que se referem (um relatório médico conterá certamente informação e vocabulário muito diferenciado de um relatório que faça um estudo de mercado no ramo imobiliário, por exemplo).

Outra diferença entre dados estruturados e não estruturados é a actualização. Os dados integrados num ambiente estruturado são actualizados regularmente (sempre que é depositado um cheque ou é feito um levantamento numa caixa ATM, a conta bancária do utilizador é actualizada, por exemplo). Em contrapartida, a generalidade dos dados não estruturados não sofre alterações após ter sido criada: depois de um contrato ter sido redigido e assinado, pode sofrer correcções ou acréscimos, mas a versão original não pode ser alterada; depois de enviado, um e-mail pode ser respondido ou reencaminhado, mas a mensagem original mantém-se. Da mesma forma, depois de publicado, um artigo de uma revista ou de um jornal não pode ser alterado.

Tradicionalmente, devido à sua natureza previsível (resultado de transacções/ de operações repetitivas) e numérica (e por isso mais facilmente manipulável), os dados estruturados são alvo de estudo. Por seu lado, os dados não estruturados apresentam uma série de desafios *a priori* para a sua análise, nomeadamente:

- Diversidade de formatos (doc, html, xls, pst, etc);
- Palavras polissémicas: uma grafia pode ter mais do que um significado. Veja-se, por exemplo, a representação gráfica “banco”. Esta pode ser a primeira pessoa do singular do verbo “bancar” ou o substantivo masculino “banco”. Por sua vez, este substantivo apresenta quinze significados distintos: 1. Assento estreito e comprido; 2. Mocho; 3. Assento dos remadores; 4. Pranchão elevado em que trabalham os carpinteiros, marceneiros, etc.; 5. Balcão de comércio; 6. Cepo de ferrador; 7. Sala de hospital, onde se recebem os consulentes externos; 8. Porção de mar em que a água tem pouca altura; 9. Baixio; 10. Grande cardume de peixe; 11. Grande massa de gelo flutuante (nos mares glaciais); 12. *Com.* Estabelecimento para transacções pecuniárias; 13. Camada de pedra,

numa pedreira; 14. *Geol.* Alta e extensa aglomeração de conchas fósseis, detritos de rochas, etc; 15. *Heráld.* Banco que serve de distintivo às armas de infante ou de príncipe[8];

- Palavras homónimas: que se pronunciam do mesmo modo, mas diferem na ortografia, como "sinto" e "cinto", "laço" e "lasso", por exemplo;
- Palavras sinónimas: por ex., mágoa, tristeza, dor de alma;
- Palavras compostas: quando é necessário utilizar mais do que uma palavra para designar um conceito: guarda-chuva, couve-flor, chapéu-de-sol, primeiro-ministro, etc.;
- Língua natural: os documentos escritos em mais de uma língua levantam sérios problemas à análise;
- Volume dos dados: dado o enorme volume dos dados, os recursos necessários para a análise poderão ser excessivos e desmotivantes;
- Hierarquia de dados não estruturados: alguns dados podem ser extremamente relevantes, carregados do ponto de vista semântico, ao passo que outros constituem meros elementos formais e gramaticais (nomeadamente preposições, artigos, etc.);
- Possibilidades de pesquisa: neste caso, o maior desafio consiste em fazer uma pesquisa que tenha como resultados os conceitos pesquisados e todos aqueles que se encontram semanticamente associados;
- Custo das infra-estruturas necessárias para suportar o ambiente não estruturado;
- Segurança: nem todos os dados não estruturados estão seguros, sendo por isso necessário assumir de antemão que todos os que têm acesso à mesma rede podem aceder a semelhantes infra-estruturas.

Não sendo repetitivos ou previsíveis, os dados não estruturados representam simultaneamente um desafio e uma oportunidade para as organizações que os pretendam utilizar no processo de decisão.

À heterogeneidade dos dados não estruturados, soma-se a "sobrecarga de informação" que pode representar um dos grandes bloqueios à inovação e competitividade das empresas.

2.4 Sobrecarga de Informação e Recuperação de Informação

O conceito "sobrecarga de informação" resulta da descoberta de que a capacidade de tomada de decisão de um indivíduo e a quantidade de informação a que este está exposto só estão positivamente correlacionadas

[8] Definição apresentada no Dicionário Priberam da Língua Portuguesa: http://www.priberam.pt/DLPO/default.aspx?pal=banco

até um determinado ponto. Se for fornecida mais informação do que aquela que o indivíduo pode processar, o desempenho do mesmo entra em declínio, e a informação que se encontra para lá desse ponto não será integrada no processo de tomada de decisão (Eppler and Mengis 2004 : 326).

A figura que se segue (adaptada de Eppler and Mengis 2004 : 326) apresenta a curva-U invertida que ilustra esta descoberta[9]

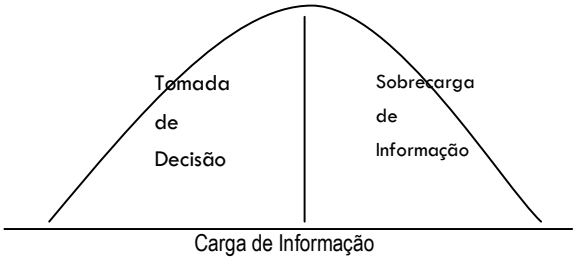


Ilustração 1 - Sobrecarga de Informação como uma curva em U invertida

De facto, se considerarmos os custos e o esforço necessariamente envolvidos no tratamento de dados não estruturados, o principal argumento que pode levar uma empresa/ instituição a fazer tamanho investimento é o enorme potencial da informação contida nos dados não estruturados, e as possibilidades abertas pelo conhecimento que a partir desta se pode gerar.

Contemplar apenas dados estruturados é escamotear um conjunto de informações potencialmente interessantes, que apoiam o processo de decisão. Vejam-se alguns exemplos do tipo de informação que pode estar "escondida" nos volumes de dados não estruturados:

- Feedback do consumidor
- Compromissos contratuais
- Garantias
- Informações médicas
- Segurança
- Marketing *buzz*: que impacto tem uma dada campanha na comunidade dos consumidores?
- Concorrência
- Recursos Humanos

[9] A partir de H.M. Schroder, M.J. Driver et al. (1967). Human information processing - Individuals and groups functioning in complex social situations. New York, Holt, Rinehart & Winston, referido em M. J. Eppler and Mengis 2004: 326.

Para fazer face ao problema da sobrecarga de informação surgem disciplinas de gestão documental, que se podem albergar sob o conceito mais vasto de Recuperação de Informação (RI)[10], e entre as quais se contam a "Categorização Textual"[11] (CT). Será este o processo aplicado através da utilização do *software Teragram TK240* para desenvolver o protótipo que constitui o cerne do presente projecto (cf. Capítulo 3 e Capítulo 4).

A Recuperação de Informação é uma área interdisciplinar (com contributos da ciência computacional, matemática, gestão documental, ciências de informação, arquitectura de informação, psicologia cognitiva, linguística, estatística e física) que visa pesquisar documentos, informação dentro dos documentos e meta dados dentro dos documentos, partindo de bases de dados relacionais e da WWW[12].

No âmbito da gestão documental e do processamento automatizado de dados não estruturados, importa ainda mencionar o Processamento de Língua Natural (PLN)[13]. O PLN reúne a ciência computacional e a linguística com vista a compreender as interacções entre linguagens naturais (humanas) e computacionais[14]. As técnicas de PLN são utilizadas para melhorar e impulsionar a RI e as disciplinas relacionadas.

A *Teragram* (empresa detentora do *software* utilizado para o protótipo desenvolvido no presente projecto) recorre a tecnologias de PLN no âmbito da pesquisa corporativa, utilizando-as para fazer buscas em bases de dados, com dados estruturados e não estruturados (incluindo relatórios de texto e páginas Web), visando assim fornecer respostas abrangentes a partir de múltiplas fontes de informação[15].

2.5 Categorização Textual

A CT (também designada, em inglês, como "text classification" ou "topic spotting", cf. Sebastiani 2002: 1) surgiu no início da década de 60 do século passado, mas só nos anos 90 passou a desempenhar um papel mais relevante no âmbito dos sistemas da informação, graças ao aumento do interesse neste tipo de soluções e ao desenvolvimento de *hardware* com melhor desempenho. A CT pode ser aplicada em diversos

---

[10] Cf. http://en.wikipedia.org/wiki/Information_retrieval) e Fabrizzzio Sebastiani 2002: 1 "In the last 10 years content-based document management tasks (collectively known as information retrieval – IR) have gained a prominent status in the information systems field."

[11] Do inglês "Text Categorization".

[12] Cf. http://en.wikipedia.org/wiki/Information_retrieval

[13] Do inglês "Natural Language Processing"

[14] "Natural language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages" Wikipedia http://en.wikipedia.org/wiki/Natural_language_processing.

[15] Cf. http://www.sas.com/news/preleases/031708/acq.html: "With today's multinational companies and distributed workforces, as well as tremendous amounts of data in disparate systems and formats, it's more important than ever to get quick and accurate answers to key business questions.
Enterprise search is a competitive weapon for tapping an organization's existing data resources. Combining SAS' business intelligence, data integration and advanced analytics with Teragram's NLP technologies will deliver answers to search queries in seconds".

contextos, desde indexação de documentos com vocabulário controlado, filtragem documental, geração automática de meta dados, desambiguação semântica, catálogos hierárquicos de recursos *Web* e qualquer tipo de aplicação que requeira organização documental ou selecção e adaptação de documentos. Até ao final dos anos 80, a abordagem mais comum a esta disciplina era do âmbito da "engenharia do conhecimento"[16], e passava pela definição manual de um conjunto de regras que codificavam conhecimento especializado sobre como classificar os documentos nas categorias respectivas. Nos anos 90, esta tendência foi perdendo popularidade entre a comunidade científica, em prol do paradigma Aprendizagem Máquina[17] (AM), de acordo com o qual um processo indutivo geral constrói automaticamente um classificador de texto automático, através da aprendizagem das características das categorias de interesse, a partir de um conjunto pré-definido de documentos. A grande vantagem desta abordagem é o alcance de uma precisão equiparável à obtida por especialistas, e uma redução significativa da intervenção de peritos, uma vez que não é necessária a participação de engenheiros ou de especialistas da área para a construção do classificador.

A CT aparece definida na literatura como "atribuição automática de textos em língua natural a um conjunto de categorias pré-definidas com base no seu conteúdo"[18] (Lewis, Yang, Rose and Li 2004; Sebastiani 2002). No entanto, esta expressão também é utilizada na literatura para designar (i) o processo de definir, de forma concisa, a informação principal contida num dado documento, ou, por outras palavras, o principal tópico/assunto de um determinado texto[19]; (ii) a identificação automática de um conjunto de categorias (por exemplo, Borko and Bernick 1963); (iii) a identificação automática de um conjunto de categorias *e* o agrupamento dos documentos sob estas categorias, sendo esta actividade normalmente designada como "text clustering" (por exemplo Merkl 1998); (iv) qualquer actividade de colocar itens textuais em grupos, que tem a CT e o "Text Clustering" como instâncias principais; (v) a ferramenta utilizada para classificar automaticamente um conjunto de documentos numa ou mais categorias pré-existentes, não tendo outra finalidade senão recuperar informação (Peixoto e tal. s.d.: 4).

No presente trabalho, entendemos a CT na primeira acepção apresentada, ou seja, como um processo de atribuição automática de documentos a uma categoria pré-estabelecida, com base no conteúdo dos

---

[16] Do inglês "Knowledge Engineering".

[17] Do inglês "Machine Learning"

[18] "Text Categorization is the automated assignment of natural language texts to predefined categories based on their content" Lewis, Yang, Rose and Li (2004). "RCV1: A New Benchmark Collection for Text Categorization Research", Journal of Machine Learning Research 5 (2004), 361.

[19] "Categorization is the process of concisely defining the information contained within a particular document; in other words, the major topic or subject of the text" (*Teragram TK240 User's Guide* Version 5.1.: 149).

documentos. Esta atribuição é conseguida através de uma ferramenta informática (neste caso o *software Teragram TK240*), e é determinada por regras linguísticas que permitem detectar automaticamente a pertença/ ausência de um determinado documento à/ da categoria em causa. À estrutura de classificação organizada resultante da descoberta de classes, chamamos "taxonomia"[20].

Este processo permite a organização e compreensão de grandes volumes de dados textuais, com vista a identificar e agrupar documentos relacionados. O seu principal objectivo é, sem dúvida, a ordenação de enorme quantidade de dados, com vista a deles extrair informação e sobre esta formar conhecimento útil e relevante[21].

Como adiante veremos, as tecnologias de categorização disponibilizadas pela *Teragram* possibilitaram a classificação de documentos de acordo com critérios pré-definidos, permitindo um acesso mais rápido e com maior exactidão aos documentos de *input* analisados, de acordo com tópicos específicos estabelecidos em função das necessidades definidas por um dado utilizador, independentemente do original.

2.6 Métodos de Categorização Textual

Embora no protótipo desenvolvido se recorra a um *software* para o processo de CT (sem intervenção no método utilizado por este *software*), importa mencionar a existência de diversos métodos de CT, e a aplicação de vários tipos de abordagem à aprendizagem, tais como:

- Modelos de regressão (N. Fuhr, S. Hartmanna, G. Lustig, M. Schwantner & K. Tzeras 1991; Y. Yang & C.G. Chute 1994);
- Classificador k-Nearest Neighbour (kNN) (B. Masand, G. Linoff & D. Waltz 1992; Yang 1994; Yang & Pederson 1997; Yang 1999; W. Lam & C.Y. Ho. 1998);
- Abordagens probabilísticas bayesianas (K. Tzeras & S. Hartman 1993; D. D. Lewis & M. Ringuette 1994; I. Moulinier 1997; D. Koller & M. Sahami 1997; Thorsten Joachims 1998; A. McCallum & K. Nigam 1998; L. Douglas Baker & Andrew K. McCallum 1998); árvores de decisão (N. Fuhr, S. Hartmanna, G. Lustig, M. Schwantner & K. Tzeras 1991; D. D. Lewis & M. Ringuette 1994; I. Moulinier 1997; C. Apte, F. Damerau & S. Weiss 1998; Thorsten Joachims 1998);
- Aprendizagem indutiva de regras (C. Apte, F. Damerau & S. Weiss 1994; William W. Cohen 1995; William W. Cohen & Yoram Singer 1996; I. Moulinier, G. Raskinis & J. Ganascia 1996);

[20] "A taxonomy is an organized classification structure that facilitates information retrieval according to the language and text of original documents" (*Teragram TK240 User's Guide* Version 5.1.: 149)

[21] Ver supra distinção dados, informação e conhecimento, Capítulo 2.2.

- Redes neuronais (NNet) (E. Wiener, J.O. Pedersen & A.S. Weigend 1995; H.T. Ng, W.B. Goh & K.L. Low 1997);
- Aprendizagem *on-line* (William W. Cohen & Yoram Singer 1996; D. D. Lewis, Robert E. Schapire, James P. Callan & Ron Papka 1996) e *Support vector machines* (SVM) (Thorsten Joachims 1998)[22].

Em Yming Yang & Xi Lu 1999, comparam-se cinco métodos 1) SVM; 2) classificador kNN; 3) abordagem redes neuronais; 4) Mapeamento Linear Least-squares Fit (LLSF) e o 5) Classificador *Naive Bays* (NB), concluindo-se que os resultados obtidos com SVM, kNN e LLSF superam claramente os métodos NNet e NB quando o número de instâncias de treino positivas por categoria é pequeno (menos de dez); e que todos os métodos apresentam um desempenho análogo quando as categorias têm mais de 300 instâncias comuns.

2.7 Aplicações de Categorização Textual

Descrevemos em seguida as principais aplicações da CT, baseando-nos em Sebastiani 2002, desde o primeiro trabalho de Maron (1961) sobre classificação estatística de textos[23], nomeadamente:

2.7.1 Indexação Automática para Sistemas de Recuperação de Informação com Operadores Boleanos

A aplicação que produziu as primeiras investigações nesta área (Borcko & Bernick 1963; Field 1975; Gray & Harley 1971; Heaps 1973; Maron 1961) foi a indexação automática de documentos para sistemas de RI com base num dicionário controlado, sendo o melhor exemplo os sistemas de operadores boleanos. Neste caso, são atribuídas uma ou mais palavras ou frases chave a cada documento, descrevendo o seu conteúdo, sendo que estas palavras e frases chave pertencem a um dicionário controlado, normalmente constituído por um *thesaurus* hierárquico. Normalmente, a atribuição é feita por mão-de-obra humana, sendo por isso uma actividade onerosa. Se considerarmos as entradas no vocabulário controlado como categorias, a indexação textual surge como uma instância da CT.

A indexação automática com dicionários controlados está relacionada com a geração automática de meta-dados. Nas bibliotecas digitais, os documentos são normalmente classificados com *tags* com meta-dados, que os descrevem sob vários aspectos: data da criação, tipo ou formato do documento, disponibilidade, etc. Alguns destes meta-dados são temáticos, isto é, o seu papel é descrever a semântica do documento através

[22] As referências bibliográficas a cada um dos métodos são feitas com base em Yming Yang & Xin Liu 1999.
[23] M. Maron (1961). Automatic indexing: na experimental inquiry, *J. Assoc. Comput. Mach. 8, 3,* 404-417.

de códigos bibliográficos, palavras-chave ou frases chave. A geração dos meta-dados pode assim ser encarada como um problema da indexação documental e abordada através de técnicas de CT.

2.7.2 Organização de Documentos

A indexação com vocabulário controlado é um exemplo do principal problema que se coloca à organização de base documental. Genericamente, muitos dos problemas relativos à organização de documentos podem ser abordados com recurso a técnicas de CT. Por exemplo, os anúncios de classificados de um jornal têm de ser previamente atribuídos a categorias, tais como "emprego", "imobiliário comprar", "imobiliário arrendar", "imobiliário vender" "compra-se", "vende-se", etc. Para lidar com grandes volumes de anúncios classificados, os jornais podem beneficiar de um sistema automatizado, que classifique automaticamente um dado anúncio na categoria mais adequada. Outras aplicações possíveis são a organização das patentes em categorias para facilitar a sua pesquisa (Larkey 1999), a atribuição automática de artigos de jornal nas secções adequadas (por exemplo Política, Economia, Cultura, Desporto, etc.) ou o agrupamento automático de artigos de conferências nas respectivas sessões.

2.7.3 Filtragem Textual

A Filtragem Textual FT[24] é a actividade de classificar um conjunto de documentos organizados de modo assíncrono, por um produtor de informação, para um consumidor de informação (Belkin & Croft 1992). O exemplo típico é um *feed* de notícias, em que o produtor é uma agência noticiosa, e o consumidor é um jornal (Hayes e tal. 1990). Neste caso, o sistema de filtragem textual bloqueia os documentos que não interessam ao consumidor da informação – por exemplo, todas as notícias não relacionadas com desporto, se considerarmos um jornal desportivo. A FT pode ser considerada como um caso de CT de etiquetagem única, ou seja, a classificação de documentos de *input* em duas categorias distintas, a relevante e a irrelevante. Adicionalmente, um sistema de filtragem pode classificar os documentos considerados relevantes para o consumidor em categorias temáticas. No exemplo do jornal desportivo acima referido, todos os artigos sobre desporto deveriam ser posteriormente classificados em função do desporto a que se referem. Do mesmo modo, um filtro de e-mails pode ser treinado para filtrar *junk e-mail* (Abdroutsopoulos et al. 2000; Drucker et al. 1999) e posteriormente classificar os restantes e-mails de acordo com categorias de interesse para o utilizador.

[24] Do inglês "Text Filtering"

Um sistema de filtros pode ser instalado na óptica do produtor de informação, em cujo caso deve orientar os documentos apenas para os consumidores interessados, construindo e actualizando um perfil de cada consumidor; ou do ponto de vista do consumidor, devendo, nesta situação, bloquear a informação considerada desinteressante para o utilizador (se assim for, só é necessário um perfil). Esta última situação é a mais frequente.

O perfil pode ser inicialmente decidido pelo utilizador, e actualizado pelo sistema com *feedback* fornecido pelo utilizador sobre a relevância (ou não) das mensagens recebidas. Este tipo de filtragem é denominado *adaptative filtering*, por oposição às situações em que não é especificado um perfil de utilizador, designadas como *Routing* ou *Batch Filtering,* dependendo se os documentos têm de ser filtrados em *rankings* de importância decrescente ou apenas aceites/ rejeitados. Assim, o *Batch Filtering* assemelha-se à CT de etiquetagem única com duas categorias.

A explosão de informação disponível em formato digital aumentou a importância destes sistemas, que são actualmente utilizados em páginas Web de jornais, bloqueio de *junk* e-mail, etc.

2.7.4 Desambiguação do Sentido de Palavras

A Desambiguação do Sentido de Palavras (DSP)[25] é a actividade de definir o sentido da ocorrência de uma dada palavra, num contexto de palavras ambíguas (polissémicas ou homónimas). No exemplo da palavra "banco" acima citada, pode constatar-se que em português europeu esta tem pelo menos 17 significados distintos (cf. *supra*, 2.3). É assim uma tarefa de Desambiguação do Sentido de Palavras decidir qual o sentido que "banco" tem, por exemplo, na frase: "O Pedro pediu dinheiro ao *banco* para comprar uma casa". Esta actividade pode ser considerada uma tarefa de CT, uma vez que os contextos de ocorrência das palavras podem ser vistos como documentos, e o sentido das palavras como categorias.

Este é apenas um exemplo do tipo de tarefas implicadas na resolução de problemas relacionados com a ambiguidade das línguas naturais, um dos maiores problemas na linguística computacional (ver *supra*, desafios provocados pelos dados não estruturados – ponto 2.3). Outros exemplos que recorrem a tecnologias de CT são correcção de ortografia de acordo com o contexto[26], anexação de frases preposicionais[27], etiquetagem de partes do discurso[28] e selecção de escolha de palavras[29].

---

[25] Do inglês "Word Sense Desambiguation"
[26] Do inglês "context sensitive spelling correction"
[27] Do inglês "prepositional phrase attachment"
[28] Do inglês "parts of speech tagging"
[29] Do inglês "Word choice selection"

2.7.5 Categorização Hierárquica de Páginas Web

A CT despertou recentemente o interesse de muitos investigadores dada a sua aplicação possível na classificação automática de sítios ou de páginas Web, nos catálogos hierárquicos alojados em portais da Internet. Quando os documentos são catalogados dessa forma, é mais fácil navegar primeiro dentro da hierarquia de categorias e depois restringir a pesquisa a uma dada categoria de interesse, em vez de publicar uma pesquisa num motor de busca genérico.

A classificação automática de páginas Web apresenta vantagens óbvias, uma vez que a categorização manual, neste caso, não é fiável. Relativamente aos casos anteriores, importa salientar que a hierarquização automática de páginas Web apresenta duas particularidades: 1) natureza hiper-textual dos documentos; 2) estrutura hierárquica do conjunto de categorias.

Para além das aplicações anteriormente referidas, a CT foi utilizada em categorização de partes do discurso através da combinação de reconhecimento do discurso[30] com CT (Myers et al. 2000; Schapire & Singer 2000); categorização de documentos multimédia através da análise de legendas (Sable & Hatzivassiloglou 2000); identificação de autores de textos literários de autoria desconhecida ou polémica (Forsyth 1999); identificação de línguas em textos de língua desconhecida (Cavnar & Trenkle 1994), identificação automática do género do texto (Kessler et. al. 1997).

2.8 Próximos Passos: Text Mining

A CT é uma disciplina com fronteiras difusas, que se situa entre a AM e a RI, partilhando ainda algumas características com o Text Mining (TM). A definição de fronteiras entre CT e TM é alvo de debate, sendo que a terminologia, nesta área, ainda se encontra em desenvolvimento. A tendência generalizada é para que o TM refira todas as tarefas que, através da análise de grandes quantidades de texto e da detecção de padrões de utilização, procuram extrair informação provavelmente útil. Neste sentido, a CT é uma instância do TM, uma parte de uma área mais completa que representa, sem dúvida, a vanguarda da tecnologia e da investigação.

O TM resulta da confluência de diversas disciplinas, como a Linguística Computacional, PLN, RI, Estatística, Bases de Dados e Ciências da Informação. Algumas destas disciplinas são comuns ao DM e ao TM, já que

[30] Do inglês "speech recognition"

o TM pode ser visto como uma sub-parte do DM que lida com uma forma particular de documento: os textos em língua natural (Kloptchenko 2003: 5).

2.8.1 Definição de Text Mining

Genericamente, o TM pode ser definido como um conjunto de técnicas ou processos aos quais se recorre para fazer face ao problema da "sobrecarga de informação" com a utilização de técnicas de DM, Aprendizagem Máquina, Processamento de Língua Natural, Recuperação de Informação e Gestão do Conhecimento (GC).

De entre as várias definições disponíveis, destacamos a oferecida por Delen and Crossland:

> *O TM pode ser visto como um processo de extracção de informação inovadora[31], previamente desconhecida e potencialmente útil, a partir de um conjunto de fontes de dados não estruturados, tais como documentos empresariais, comentários de clientes, páginas Web e ficheiros XML.*
>
> (Delen and Crossland 2008: 1710)[32].

Esta definição está muito próxima de uma das definições de DM mais populares na literatura (Fayyad 1996, cf. nota 16). Com efeito, o TM coincide com o DM no objectivo fundamental de extracção de informação útil, através da identificação e exploração de padrões relevantes. No entanto, as duas disciplinas diferenciam-se num elemento essencial: o tipo de dados analisados. O DM opera sobre dados numéricos armazenados em grandes bases de dados, ao passo que o TM tem como objecto de investigação documentos de texto, cujo único requisito à análise é estarem em formato digital.

Durante o processo de TM, o utilizador interage com um conjunto de documentos textuais[33], recorrendo para tal a uma série de ferramentas analíticas.

O TM implica o pré-processamento dos conjuntos de documentos (onde também entra a CT, extracção de informação, extracção de termos), o armazenamento das representações intermédias, as técnicas para

---

[31] Por "informação inovadora" os autores referem as associações, hipóteses ou tendências que não se encontram explicitamente presentes nas fontes textuais em análise e que serão uma mais valia proporcionada pela aplicação de técnicas de TM.

[32] "Text mining is the process of discovering new, previously unknown, potentially useful information from a variety of unstructured data sources including business documents, custommer comments, Web pages and XML files."

[33] Também denominados "corpus" (D. Delen and M. Crossland (2008). "Seeding the survey and analysis of research literature with text mining." Expert Systems With Applications **34**: 1710).

analisar as representações intermédias (tais como análise de distribuição, *clustering*, análise de tendências, e regras de associação), e a visualização dos resultados (Feldman and Sanger 2007 : x)[34].

Em virtude da natureza dos dados, na aplicação de uma metodologia de TM é imprescindível uma fase de pré-processamento linguístico dos mesmos, inexistente em DM (Chen 2001: 15). Por outro lado, a natureza não estruturada ou semi-estruturada do objecto da investigação obriga a que as aplicações de TM lidem com maior diversidade de formatos (mensagens de correio electrónico, páginas web, documentos de texto, etc.).

Os objectivos do DM podem ser preditivos (estimar os resultados de situações futuras) ou descritivos (analisar as razões que afectam o resultado esperado, visualizar as relações entre os dados). Os objectivos do TM passam pela descoberta de informação relevante num conjunto de textos, mas também pela categorização de conteúdos e pela comparação e descoberta de relações entre textos.

Os métodos usados no DM são entre outros, árvores de decisão, algoritmos genéticos, redes neuronais ou regressão multilinear (Kloptchenko 2003: 7; Bação 2007). O TM utiliza técnicas de indexação, redes neuronais, algoritmos de *clustering* e de categorização, análises linguísticas e ontologias (Kloptchenko 2003: 7). Tanto o DM como o TM adoptam métodos analíticos, obtendo resultados visuais e gráficos. As técnicas de visualização de dados e de visualização de informação visam criar uma interface adequada enquanto sistema de apoio à decisão (Chen 2001 : 15).

2.8.2 Forças para o desenvolvimento do Text Mining

Em 1999, Hearst lamentava o facto de o TM não ter muitos seguidores[35]. Seis anos depois, na abertura de um livro dedicado ao TM, Zanasi dá conta do dinamismo e da vitalidade do mercado do TM, como reflexo do interesse que esta área tem vindo a angariar em diferentes sectores de actividade, a nível mundial: "*o mercado do* text mining *está agora a nascer, e demonstra uma vitalidade inesperada"*[36] (Zanasi 2005 : xxvii).

Actualmente, a investigação e as propostas de aplicações de TM estão a conhecer um crescimento exponencial. Zanasi (2005) identifica três forças determinantes que orientam este crescimento.

---

[34] " Text mining involves the preprocessing of document collections (text categorization, information extraction, term extraction), the storage of the intermediate representations, the techniques to analyze these intermediate representations (such as distribution analysis, clustering, trend analysis, and association rules), and visualization of the results."

[35] "O recém-nascido campo do TDM tem a peculiaridade de já ter um nome e um grande impacto, mas, até ao momento, quase nenhum praticante." ("The nascent field of text data mining (TDM) has the peculiar distinction of having a name and a fair amount of hype but as yet almost no practitioners.") (Hearst 1999: 1)

[36] "The text mining market has just been born, and is showing unexpected vitality."

A primeira consiste no aumento significativo do fluxo de informação textual. O processamento de toda a informação disponível constitui um desafio cada vez maior, que se coloca no âmbito empresarial, no campo da investigação científica, mas também no quotidiano, e que é visível quando realizamos uma pesquisa simples num motor de busca[37]. A necessidade de mecanismos de exploração automática cresce de dia para dia pois, com a profusão de informação, despende-se cada vez mais tempo no processo de selecção da mesma, com o risco de, ainda assim, numa dada pesquisa, não nos apercebermos de documentos relevantes para a mesma. O TM pode facilitar a tarefa de determinar objectivamente quais as referências textuais relevantes num determinado contexto, levando o empresário, investigador ou o utilizador comum a concentrar-se nelas. No campo empresarial esta necessidade é tanto mais premente, uma vez que a informação textual, quando convenientemente manipulada, pode tornar-se numa poderosa ferramenta de negócio (Zanasi 2005 : xxix).

O empenho governamental e empresarial na detecção de ameaças e oportunidades, no campo da segurança nacional ou na actividade empresarial, respectivamente, é o segundo dos incentivos ao desenvolvimento do TM identificados por Zanasi. Os progressos conhecidos neste âmbito, na última década, simplificaram a gestão dos documentos de texto em ambiente empresarial e contribuíram para avanços nas tecnologias nos motores de busca (Zanasi 2005 : xxx)

Os avanços na investigação da Aprendizagem Máquina são a terceira das forças apontadas por Zanasi para o desenvolvimento do TM. Os progressos tecnológicos resultantes da investigação levada a cabo por grupos de trabalho, em contexto universitário ou empresarial, antecipam potenciais novas aplicações no âmbito governamental e empresarial. De acordo com o mesmo autor, estes avanços podem apresentar-se em três categorias: (Delen and Crossland) Reconhecimento Inteligente de Texto (possibilidade de compreender o contexto gramatical e as relações lógicas entre conceitos dentro de um texto); (ii) Classificação Inteligente (capacidade de organizar os documentos em categorias pré-definidas ou geradas automaticamente[38]); (iii) Trabalho com várias línguas naturais (possibilidade de trabalhar em simultâneo com documentos escritos em várias línguas diferentes ou que contenham várias línguas diferentes) (Zanasi 2005 : xxxii). Tais desenvolvimentos permitirão o aparecimento de uma série de aplicações em ambiente empresarial, em áreas que ainda não são abrangidas pelas tecnologias de análise textual actualmente existentes.

---

[37] A título ilustrativo, a pesquisa por "*Text Mining*", efectuada num dos mais populares motores de busca (Google) em Abril de 2008, teve como resultado cerca de 3 210 000 resultados.

[38] O que no presente projecto referimos como "Categorização Textual".

2.8.3 Aplicações de Text Mining

As aplicações de TM são variadas e têm interesse em diversos ramos de actividade. De entre as referidas na literatura, destacam-se as seguintes:

- Avaliação da evolução das tendências nas reclamações e garantias (Mcknight 2005: 80);
- Possibilidade de recolher mais informação com a aplicação de programas de marketing ou de *focus groups* - a análise automatizada/ semiautomatizada de documentos não estruturados permite processar individualmente declarações personalizadas dos sujeitos em análise (Mcknight 2005: 80);
- Pesquisa de dados em áreas de negócio que tradicionalmente manipulam grandes volumes de documentos em formato textual, como a indústria farmacêutica, a área dos cuidados de saúde ou o direito (Mcknight 2005: 80);
- Introdução de melhorias na personalização de aplicações de *e-commerce* B2C (Zhang and Jiao 2007: 357 e ss.);
- Construção automática de hiper-textos (Yang and Lee 2005: 723 e ss.);
- Elaboração de sumários automáticos de documentos (Delen and Crossland 2008: 1710);
- Estabelecimento de "ligação entre conceitos"[39]: estabelecer a relação entre documentos através da identificação dos conceitos partilhados pelos mesmos, permitindo assim aos utilizadores encontrar informação à qual, de outra forma, poderiam não ter acesso (Delen and Crossland 2008: 1710);
- *Clustering*: agrupamento de documentos semelhantes sem ter um conjunto pré-determinado de categorias (Weng and Lin 2003: 355 e ss.; Delen and Crossland 2008: 1710);
- Gestão do correio electrónico, com classificação e filtragem de mensagens de correio electrónico, e criação de mecanismos de resposta automática (Weng and Liu 2004: 529 e ss.)

[39] Do inglês "Concept Linking"

3 Capítulo 3

3.1 O Caso dos debates parlamentares

Este projecto nasceu do desejo de aplicar um programa de processamento automatizado de dados não estruturados ao português europeu, tendo por base as descrições dos debates parlamentares que nos são fornecidas nos Diários da Assembleia da República (DAR).

Tirámos partido, em primeiro lugar, da possibilidade de utilizar o software Teragram TK 240 pela primeira vez em português europeu, que nos foi gentilmente concedida pelo SAS. Por outro lado, procurámos colmatar algumas limitações de pesquisa detectadas no sítio da AR (cuja análise é detalhada no ponto 7.2.). Criou-se assim uma "solução de compromisso" que beneficia das características sintáctico linguísticas dos dados, e, concomitantemente, aproveita as potencialidades de automatismo oferecidas pelo *software*, proporcionando uma análise diferenciada dos conteúdos dos debates parlamentares, não possível com a simples consulta do sítio.

Após análise do sítio da AR, constataram-se algumas limitações de pesquisa no que respeita aos dados que seriam alvo da nossa análise:

- Está limitada a um Grupo Parlamentar (GP) e a um orador, o que impossibilita a consulta das intervenções de todos os deputados do mesmo GP;
- Os resultados da pesquisa feita no sítio da AR não são exactos: por exemplo, após pesquisa por Euro 2004, foram elencados resultados relacionados com o Parlamento Europeu, provavelmente devido à coincidência gráfica nas palavras "Euro"/ "Europeu" (ver 7.2.);
- São apenas pesquisadas as intervenções em "discurso directo". As "indicações cénicas" (correspondente ao que, em linguagem teatral, se designa como "didascália") não são consideradas, perdendo-se informação sobre a reacção dos deputados às intervenções.

De um outro ponto de vista, considerando os dados que seriam alvo da nossa análise, uma das potencialidades mais aliciante, devido ao factor de inovação introduzido na análise, era a possibilidade de auscultar as emoções vividas na AR. Com efeito, estas são-nos remotamente transmitidas nos DAR através da descrição das reacções dos Grupos Parlamentares às intervenções, com um de quatro substantivos: *aplausos, protestos, risos* ou *vozes* e estão sempre identificadas com o/os grupo(s) parlamentar(es) que as assumiram.

Assim sendo, com recurso ao programa TK240, desenvolveu-se um modelo de análise que permite sondar as emoções da Assembleia e os entendimentos (ou desentendimentos) dos Grupos Parlamentares

relativamente aos intervenientes. O protótipo elaborado permite categorizar automaticamente os documentos de *input* em função das reacções dos Grupos Parlamentares, pela utilização do programa *Teragram TK 240* e da sua ferramenta *Teragram Categorizer* (cf. *infra* 3.2.).

O software, os dados e a metodologia adoptada na concretização deste projecto serão descritos com maior detalhe ao longo deste capítulo.

3.1.1 Objectivo Inicial

Como descrevemos anteriormente, o presente protótipo resultou da adopção de uma solução de compromisso. Com efeito, o intuito inicial era um pouco mais ambicioso, visando aplicar a tecnologia de *Text Mining* aos mesmos relatos integrais das reuniões plenárias da AR. O objectivo era analisar a evolução dos temas debatidos entre 1976 e 2005, numa perspectiva longitudinal, com o recurso a um *software* de *Text Mining.* Inicialmente, previa-se a utilização do *SAS Text Miner 3.1*., disponível e comercializado em português europeu, para fazer o estudo longitudinal acima referido, analisando todos os debates parlamentares sem a necessidade de seleccionar previamente uma amostra, com a aplicação de uma ferramenta automática de processamento da informação.

Com o apoio do SAS Portugal, o *software SAS Text Miner 3.1.* foi instalado no computador pessoal utilizado para realizar este projecto e foram realizadas várias tentativas de análise dos dados cedidos pela AR. Infelizmente, a utilização do *SAS Text Miner 3.1.* não foi bem sucedida, possivelmente devido ao facto de existirem grandes lacunas no que respeita ao dicionário de português europeu. Outro obstáculo foi o formato dos dados propriamente dito, inicialmente descarregados do sítio da AR em PDF imagem (único formato disponível em livre acesso nesse momento, apesar de actualmente já ser possível descarregar os mesmos dados em PDF editável), e depois disponibilizados pela AR em html. Mesmo com a cooperação do SAS, não foi possível transformar os dados em tabelas SAS, etapa imprescindível para passar à fase de análise dos mesmos. Considerando estas condicionantes, por sugestão do SAS Portugal, recorreu-se à utilização do *software Teragram TK 240,* instalado num computador da SAS gentilmente dispensado por esta empresa durante o período de realização do presente projecto.

3.2 Software Teragram TK 240

Dadas as limitações na utilização do *software* de *Text Miner* acima mencionadas, o SAS sugeriu, como alternativa, a utilização do programa *Teragram TK 240*. A disponibilidade deste programa associou-se à aquisição recente, pelo SAS, da empresa *Teragram*, líder em Processamento de Línguas Naturais e em

tecnologias linguísticas, anunciada em 17 de Março de 2008, no SAS *Global Forum*, em Santo António, Texas. Com esta aquisição, o SAS reforçou as áreas de conhecimento de *Business Intelligence* e *Text Mining*, complementando e potenciando a oferta já disponibilizada, na qual se destacava o S*AS Text Miner* (http://www.sas.com/news/preleases/031708/acq.html).

A *Teragram* é especializada em tecnologias de processamento de línguas naturais, que permitem a extracção de informação de grandes conjuntos de dados. Fundada em 1997 por investigadores do ramo da linguística computacional, a *Teragram* afirma oferecer a velocidade, a exactidão e o apoio linguístico necessários para que clientes e parceiros pesquisem e organizem volumes crescentes de informação digital. A *Teragram* possibilita pesquisas e organização da informação em mais de 30 línguas, permitindo que os seus clientes atinjam novos mercados e apoiando-os na tomada de decisão. Entre os clientes da *Teragram*, contam-se empresas como: *Ariba, Ask.com, Associated Press, CNN, Factiva, EBSCO Publishing, FAST Search & Transfer, Forbes.com, InfoSpace, NYTimes Digital, OneSource, Reed Business Information, Ricoh, Sony, WashingtonPost.com, Wolters Kluwer,* o Banco Mundial e a *Yahoo!* (ver mais em http://www.teragram.com/info).

O *Teragram TK 240* é constituído por duas ferramentas, o *Teragram Categorizer* e o *Teragram Concepts Extractor*, que permitem organizar, de forma sistemática, enormes conjuntos de documentos e extrair conceitos chave de grandes volumes de informação[40]. Estas tecnologias linguísticas facilitam o controlo do fluxo de informação nas organizações e permitem uma melhor organização, acesso e detecção de dados.

O *Teragram Categorizer* permite a classificação de documentos e organização de informação em dois tipos de taxonomia[41]:

1) Taxonomia Hierárquica: este tipo de taxonomia estabelece categorias e subcategorias do tipo pai/ filho. A informação contida numa categoria mais abrangente ("pai") é subdividida em subcategorias separadas ("filhos"), de acordo com as regras de subcategorização.

2) Taxonomia Plana: esta taxonomia não apresenta subcategorias ("filhos"). Neste tipo de taxonomia, as categorias contêm todos os documentos relevantes, sem outro tipo de subdivisões.

O *Teragram Concepts Extractor* possibilita a extracção de conceitos chave (tais como nomes de pessoas, empresas ou topónimos) a partir de um documento de *input*. Estes conceitos podem ser:

---

[40] No protótipo aqui apresentado, utilizamos apenas o *Teragram Categorizer*.

[41] Cf. capítulo 2.5. Por "taxonomia" entenda-se uma estrutura de classificação organizada, que facilita a pesquisa de informação, tendo em conta a língua e o texto original dos documentos.

1) Conceitos Simples: dados isolados ou facilmente reconhecíveis, tais como "José Sócrates", "Partido Socialista", "Secretário-Geral".

2) Conceitos Relacionais: entidades que têm uma relação com outras também podem ser identificadas, com o intuito de reunir mais informação sobre dados que, de outra forma, estariam isolados. Por exemplo, se "José Sócrates, Secretário-Geral do Partido Socialista", forem associados como conceitos relacionais, o utilizador fica a conhecer informação adicional sobre José Sócrates.

Estas duas ferramentas podem ser utilizadas isoladamente ou em conjunto. O *Teragram Categorizer* pode, por exemplo, ser utilizado para recorrer a conceitos no momento de definir regras de categorização.

O *Teragram TK 240* pode ser usado para criar, definir, testar e compilar as categorias nas quais se pretende que os documentos sejam organizados, e/ou os conceitos que são extraídos do conjunto de documentos relevantes para satisfazer a pesquisa, estando assim disponível em três configurações distintas: (1) apenas categorização; (2) apenas extracção de conceitos; (3) categorização e extracção de conceitos.

Este *software* pretende fazer face a alguns desafios da actual sociedade de informação (cf. capítulo 2.4. e ss.), facilitando, nomeadamente, os processos de (1) classificação de informação (reúne documentos relacionados por assunto, e, simultaneamente, separa documentos não relacionados, facilitando a localização dos dados); (2) organização de documentos (torna os documentos mais facilmente acessíveis, facilitando a localização e descoberta da informação); (3) extracção de informação chave (permite filtrar grandes quantidades de informação e reduzi-la a um acervo mais facilmente analisável, através da detecção de conceitos chave que permitem aos utilizadores uma apreensão mais rápida da informação de que necessitam); (4) identificação de conceitos relacionados (a informação relacional permite aos utilizadores localizar e apreender rapidamente o conhecimento fundamental necessário especializado numa dada área).

3.3 Dados

Os dados em análise foram os relatos integrais das reuniões plenárias que decorreram na AR, para a nona legislatura (composta por três sessões legislativas). Estes encontram-se compilados no *Diário da Assembleia da República* (DAR), o jornal oficial da AR. Embora o DAR contenha 2 séries independentes[42] no presente projecto, analisou-se apenas a 1.ª série. O DAR é publicado em formato electrónico na página

[42] Cf. Capítulo 7.2.3

da internet com o endereço: http://www.parlamento.pt/DAR/Paginas/DAR1Serie.aspx, e está acessível desde o início da Primeira Legislatura (com início em 1976), até à actualidade[43].

Como foi anteriormente referido, durante o período de realização deste projecto (Maio 2008- Agosto 2009), a página da AR sofreu diversas alterações e melhoramentos. Inicialmente, o DAR podia ser lido directamente na página da AR, em formato *html*; ou impresso, a partir desta página, para PDF (imagem).

Para a concretização do presente trabalho, era imprescindível ter acesso imediato e facilitado aos documentos em formato texto. Por este motivo, foram solicitados à AR os DAR de todas as sessões legislativas, desde 1974.

Embora a intenção inicial fosse concretizar um estudo longitudinal que permitisse aferir a evolução das reacções dos GP em Assembleia, infelizmente, a Divisão de Redacção e de Apoio Audiovisual da AR não pôde disponibilizar os ficheiros html relativos a este período.

Como tal, optou-se pelo desenvolvimento de um protótipo (que servirá de base a uma eventual análise longitudinal futura), tendo como base apenas uma legislatura (período cronológico considerado razoável pela Divisão de Redacção e de Apoio Audiovisual, relativamente ao fornecimento dos dados).

Foi seleccionada a legislatura, concluída, mais recente em relação à data de realização do nosso trabalho (a IX), uma vez que os originais dos documentos mais antigos (em particular os referentes ao período entre a primeira e a sétima legislaturas) não existem em formato electrónico, tendo sido recuperados com o auxílio de ferramentas de *software* e, como tal, apresentando menor fiabilidade do que os documentos que não foram alvo desta intervenção[44]. Foram-nos gentilmente cedidos em formato *html* os DAR 1.ª série, das três sessões legislativas da nona legislatura (2002-2005)[45].

Os DAR foram utilizados no formato disponibilizado pela Divisão de Redacção e Apoio Audiovisual da AR, isto é, em *html*. Um DAR corresponde a vários documentos *html*, representando cada ficheiro uma visualização do texto na página da internet. No total, foram utilizados 13520 ficheiros *html* (com um total de 94,8 MB), correspondentes a 278 DAR, classificados de forma automática pelo RBC do *Teragram TK 240*.

---

[43] Os textos referentes aos Diários entre a 1ª e a 7ª Legislaturas foram recuperados recorrendo a ferramentas de *software* por já não haver os originais em formato electrónico. Este processo de recuperação de texto oferece, actualmente, um grau de fiabilidade superior a 95%. Por este valor não ser considerado suficiente, os textos foram todos corrigidos manualmente, página por página. Infelizmente, e apesar dos esforços despendidos, persistem alguns erros, quer de sintaxe quer semânticos.

[44] Ver nota 43.

[45] No início de 2009, os dados passaram a estar disponíveis na renovada página da AR, em PDF editável, formato que facilitaria em grande medida este projecto. No entanto, uma vez que o trabalho já tinha sido previamente iniciado com base nos dados inicialmente cedidos, não se tirou partido desse formato editável.

O facto de não haver uma correspondência unívoca entre um ficheiro *html* e um DAR (ou seja, um DAR = vários documentos *html*) não teve qualquer implicação na análise efectuada, uma vez que esta levou em conta as reacções e a forma como os Grupos Parlamentares se associaram (ou não) nestas reacções, independentemente da sessão parlamentar em que estas ocorreram.

3.3.1 IX Legislatura

A legislatura[46] seleccionada tem a particularidade de abranger dois Governos Constitucionais (GC) - o XV e o XVI. O XV GC (2002-2004), liderado por Durão Barroso, foi formado por um acordo de incidência parlamentar entre o Partido Social Democrata (PSD) e o Partido Popular (dirigido por Paulo Portas, que ocupou o cargo de Ministro da Defesa). Em 2004, na sequência do pedido de demissão de Durão Barroso, que assumiu o cargo de Presidente da Comissão Europeia, assistiu-se à dissolução do XV GC e à nomeação, por Jorge Sampaio (Presidente da República de então), de Pedro Santana Lopes, para a presidência do XVI GC (2004-2005). Tal como o anterior, este resultou de um acordo de incidência parlamentar entre o PSD e o CDS-PP. Em Dezembro de 2004, o Presidente da República dissolveu o Parlamento e convocou eleições legislativas antecipadas, determinando dessa forma a demissão do XVI GC.

A eleição da AR ocorreu em 17.03.02. Embora habitualmente uma legislatura seja composta por 4 sessões legislativas[47], dada a dissolução da assembleia em Dezembro de 2004, na IX Legislatura houve apenas 3 sessões legislativas:

- 1.ª Sessão Legislativa – início a 05.04.02
- 2.ª Sessão Legislativa – início a 15.09.03
- 3.ª Sessão Legislativa – início a 15.09.04

[46] Uma legislatura corresponde ao período do mandato de cada Assembleia eleita. Em princípio tem a duração de 4 anos, designados por sessões legislativas. No entanto uma legislatura pode não completar os 4 anos se a Assembleia da República for dissolvida. Neste caso, a nova Assembleia irá iniciar uma nova legislatura cuja duração será acrescida, no seu início, do período correspondente à sessão legislativa em curso à data da eleição (cf. http://pt.wikipedia.org/wiki/Legislatura_(Portugal).

[47] Uma Sessão Legislativa corresponde ao período anual de funcionamento da Assembleia da República e inicia-se a 15 de Setembro (cf. http://pt.wikipedia.org/wiki/Sess%C3%A3o_Legislativa_(Portugal).

A distribuição dos Grupos Parlamentares na AR foi a que se segue:

partido	deputados	votos	percentagem
BE	3	149.966	2,74% a)
PCP	10	b)	b)
PEV	2	b)	b)
PS	96	2.068.584	37,76%
PPD/PSD	105	2.200.765	40,21%
CDS-PP	14	477.350	8,72%

TABELA 1 - DEPUTADOS POR GP DURANTE A IX LEGISLATURA, COM INDICAÇÃO DOS VOTOS RECEBIDOS POR CADA GP E A RESPECTIVA PERCENTAGEM REPRESENTADA EM AR[48]

a) O BE concorreu também em coligação com a UDP no círculo eleitoral da Madeira, tendo obtido 3.911 votos (0,07%);
b) PCP e PEV concorreram juntos na coligação PCP/PEV, tendo obtido o total de 379.670 votos (6,94%).

O facto de esta legislatura ter sido constituída por dois Governos Constitucionais (GC) permitiu-nos efectuar uma análise comparativa dos resultados em cada governo. Assim sendo, o modelo de categorias criado neste projecto foi aplicado aos DAR dos dois GC em separado:

	Sessão Legislativa	DAR	Ficheiros html
XV Governo Constitucional	1.ª	001 a 146 (05-05-2002 a 03-09-2003)	6125
	2.ª (até à tomada de posse do XVI Gov. Constitucional)	001 a 105 (17-09-2003 a 08-07-2004)	5677
XVI Governo Constitucional	2.ª (final - depois da tomada de posse do XVI Gov Constitucional)	106 a 108 (27-07-2004 a 02-09-2004)	245
	3.ª	001 a 024 (2004-09-15 a 2005-03-10)	1473
Total		**278 DAR**	**13.520 html**

TABELA 2 - CORRESPONDÊNCIA ENTRE SESSÃO LEGISLATIVA, DAR E FICHEIRO HTML UTILIZADO

3.3.2 Diário da Assembleia da República

Os DAR são descrições integrais das sessões que decorrem na AR. Apresentam uma estrutura homogénea, nos moldes que se seguem (cf. Ilustração 2 - Imagem da folha de rosto do DAR e Ilustração 3 - Imagem do interior de um DAR. que mostram um DAR, neste caso o número 9 da I.ª Série, IX Legislatura, 1.ª Sessão Legislativa):

[48] Adaptado de http://www.parlamento.pt/DeputadoGP/Paginas/resultadoseleitorais.aspx

3.3.3 Estrutura do DAR

Página de Rosto

- Data, Indicação da Série e Número do Diário
- Legislatura e Sessão Legislativa
- Indicação da Data em que foi realizada a Reunião Plenária
- Presidente da AR e Secretários
- Sumário
- Hora de início da sessão: *O Sr. Presidente declarou aberta a sessão às X horas e X minutos*.
- São resumidas as actividades que tiveram lugar *Antes da ordem do dia* e na *Ordem do Dia*

Primeira página

- Abertura da sessão pelo presidente da assembleia em funções
- Horas (em itálico)
- Listagem dos Deputados Presentes à Sessão
    - Partido Social Democrata
    - Partido Socialista
    - Partido Popular
    - Partido Comunista Português
    - Bloco de Esquerda
    - Partido Ecologista «Os Verdes»
- Intervenções ocorridas, com indicação do nome do deputado e do GP a que este pertence, e a transcrição de todas as intervenções em discurso directo
- Descrição das reacções dos Grupos Parlamentares (em itálico)
- Horas
- Deputados que entraram durante a sessão
- Deputados que faltaram à sessão

Sexta-feira, 10 de Maio de 2002 — I Série — Número 9

DIÁRIO da Assembleia da República

IX LEGISLATURA — 1.ª SESSÃO LEGISLATIVA (2002-2003)

REUNIÃO PLENÁRIA DE 9 DE MAIO DE 2002

Presidente: Ex.mo Sr. João Bosco Soares Mota Amaral

Secretários: Ex.mas Srs. Duarte Rogério Matos Ventura Pacheco
Ascenso Luís Seixas Simões
Isabel Maria de Sousa Gonçalves dos Santos
António João Rodeia Machado

SUMÁRIO

O Sr. Presidente declarou aberta a sessão às 15 horas e 15 minutos.

Deu-se conta da aprovação dos projectos de lei n.os 14 a 21/IX.

Em interpelação à Mesa, a Sr.ª Deputada Helena Neves (PS) deu conhecimento à Câmara da realização, na Sala do Senado, de uma conferência de imprensa, de sua iniciativa, para prestar informação pública acerca de dois quadros de asserta de Natália Correia desaparecidos e encontrados pela Polícia Judiciária.

Ao abrigo do artigo 76.º, n.º 2, do Regimento, procedeu-se a um debate de relevante relevante, requerido pelo Grupo Parlamentar do PCP, sobre a situação no Médio Oriente, tendo-se pronunciado a diversos título, além do Sr. Ministro dos Negócios Estrangeiros e dos Comunidades Portuguesas e dos Comunidades Portuguesas (António Martins da Cruz), os Srs. Deputados António Filipe (PCP), Luís Fazenda (BE), António Nazaré Pereira (PSD), Basílio Horta (CDS-PP), Isabel Castro (Os Verdes), Teresa Patrício Gouveia (PSD), Telmo Correia (CDS-PP) e Francisco da Assis (PS).

A encerrar o debate, proferiram intervenções, além do Sr. Ministro dos Negócios Estrangeiros e das Comunidades Portuguesas, os Srs. Deputados Isabel Castro (Os Verdes), Francisco Louçã (BE), Basílio Horta (CDS-PP), Odaisseu Ferreira (PS), Antero Nazaré Pereira (PSD) e António Filipe (PCP).

Em interpelação à Mesa, o Sr. Deputado Francisco Louçã (BE) lembrou a mensagem do Sr. Presidente da República e solicitou a correcção dos votos relativamente a todas as votações de diplomas sujeitos a promulgação do Sr. Presidente da República que, a partir de agora, se realizam.

A Câmara aprovou dois pareceres da Comissão de Ética, autorizando um Deputado do PSD a depor em tribunal como testemunha.

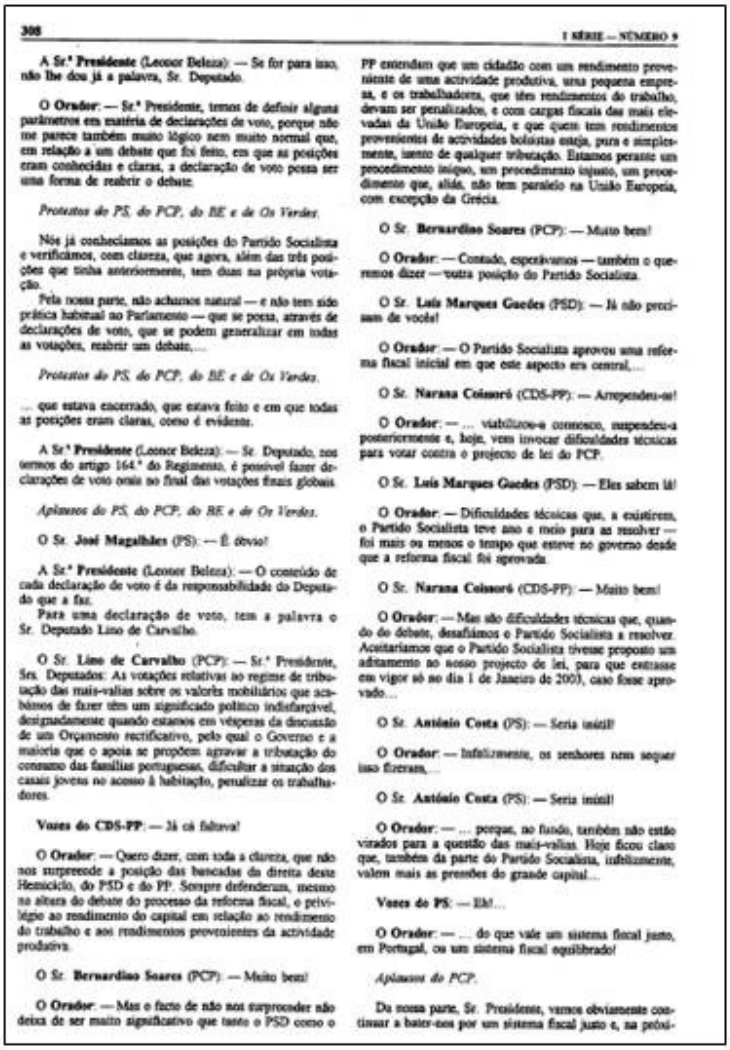
O projecto de resolução n.º 1/IX — Constituição de uma comissão eventual para a análise e a fiscalização dos recursos públicos envolvidos na organização do EURO 2004 (PSD e CDS-PP) mereceu aprovação.

O projecto de lei n.º 2 /IX — Reposição da tributação dos mais-valias nos impostos sobre o rendimento (PCP), foi rejeitado na generalidade, e o projecto de lei n.º 3/IX — Justiça e Cidadania a alterar o Código do IRS, aprovado pelo Decreto-Lei n.º 442-A/88, de 30 de Novembro, revendo o regime jurídico das mais-valias estabelecido pela Lei n.º 30-G/2000, de 29 de Dezembro, bem como a alterar o Estatuto dos Benefícios Fiscais, aprovado pelo Decreto-Lei n.º 215/89, de 1 de Julho, revendo o regime aplicável aos rendimentos dos fundos de investimento, foi aprovada na generalidade, na especialidade e em votação final global. Proferiram declarações de voto os Srs. Deputados António Costa (PS), Jorge Vaz (PSD), Telmo Correia (CDS-PP), Lino de Carvalho (PCP) e Francisco Louçã (BE).

Por último, a Câmara aprovou o projecto de resolução n.º 6/IX — Constituição de uma comissão eventual para a reforma do sistema político (Presidente da AR).

O Sr. Presidente encerrou a sessão eram 18 horas e 40 minutos.

ILUSTRAÇÃO 2 - IMAGEM DA FOLHA DE ROSTO DO DAR



308 I SÉRIE — NÚMERO 9

A Sr.ª **Presidente** (Leonor Beleza): — Se for para isso, não lhe dou já a palavra, Sr. Deputado.

O **Orador**: — Sr.ª Presidente, temos de definir alguns parâmetros em matéria de declarações de voto, porque não me parece também muito lógico nem muito normal que, em relação a um debate que foi feito, em que as posições eram conhecidas e claras, a declaração de voto possa ser uma forma de reabrir o debate.

Protestos do PS, do PCP, do BE e de Os Verdes.

Nós já conhecíamos as posições do Partido Socialista e verificámos, com clareza, que agora, além das três posições que tinha anteriormente, tem duas na própria votação.

Pela nossa parte, não achamos natural — e não tem sido prática habitual no Parlamento — que se possa, através de declarações de voto, que se podem generalizar em todas as votações, reabrir um debate,...

Protestos do PS, do PCP, do BE e de Os Verdes.

... que estava encerrado, que estava feito e em que todas as posições eram claras, como é evidente.

A Sr.ª **Presidente** (Leonor Beleza): — Sr. Deputado, nos termos do artigo 164.º do Regimento, é possível fazer declarações de voto orais no final das votações finais globais.

Aplausos do PS, do PCP, do BE e de Os Verdes.

O Sr. **José Magalhães** (PS): — É óbvio!

A Sr.ª **Presidente** (Leonor Beleza): — O conteúdo de cada declaração de voto é da responsabilidade do Deputado que a faz.

Para uma declaração de voto, tem a palavra o Sr. Deputado Lino de Carvalho.

O Sr. **Lino de Carvalho** (PCP): — Sr.ª Presidente, Srs. Deputados: As votações relativas ao regime de tributação das mais-valias sobre os valores mobiliários que acabámos de fazer têm um significado político indisfarçável, designadamente quando estamos em vésperas da discussão de um Orçamento rectificativo, pelo qual o Governo e a maioria que o apoia se propõem agravar a tributação do consumo das famílias portuguesas, dificultar a situação dos casais jovens no acesso à habitação, penalizar os trabalhadores.

Vozes do CDS-PP: — Já cá faltava!

O **Orador**: — Quero dizer, com toda a clareza, que não nos surpreende a posição das bancadas da direita desta Hemiciclo, do PSD e do PP. Sempre defenderam, mesmo na altura do debate do processo da reforma fiscal, o privilégio ao rendimento do capital em relação ao rendimento do trabalho e aos rendimentos provenientes da actividade produtiva.

O Sr. **Bernardino Soares** (PCP): — Muito bem!

O **Orador**: — Mas o facto de não nos surpreender não deixa de ser muito significativo que tanto o PSD como o PP entendam que um cidadão com um rendimento proveniente de uma actividade produtiva, uma pequena empresa, e os trabalhadores, que têm rendimentos do trabalho, devam ser penalizados, e com cargas fiscais das mais elevadas da União Europeia, e que quem tem rendimentos provenientes de actividades bolsistas esteja, pura e simplesmente, isento de qualquer tributação. Estamos perante um procedimento iníquo, um procedimento injusto, um procedimento que, aliás, não tem paralelo na União Europeia, com excepção da Grécia.

O Sr. **Bernardino Soares** (PCP): — Muito bem!

O **Orador**: — Contudo, esperávamos — também o queremos dizer — outra posição do Partido Socialista.

O Sr. **Luís Marques Guedes** (PSD): — Já não precisam de vocês!

O **Orador**: — O Partido Socialista aprovou uma reforma fiscal inicial em que este aspecto era central,...

O Sr. **Narana Coissoró** (CDS-PP): — Arrependeu-se!

O **Orador**: — ... viabilizou-a connosco, suspendeu-a posteriormente e, hoje, vem invocar dificuldades técnicas para votar contra o projecto de lei do PCP.

O Sr. **Luís Marques Guedes** (PSD): — Eles sabem lá!

O **Orador**: — Dificuldades técnicas que, a existirem, o Partido Socialista teve ano e meio para as resolver — foi mais ou menos o tempo que esteve no governo desde que a reforma fiscal foi aprovada.

O Sr. **Narana Coissoró** (CDS-PP): — Muito bem!

O **Orador**: — Mas são dificuldades técnicas que, quando do debate, desafiámos o Partido Socialista a resolver. Aceitaríamos que o Partido Socialista tivesse proposto um aditamento ao nosso projecto de lei, para que entrasse em vigor só no dia 1 de Janeiro de 2003, caso fosse aprovado...

O Sr. **António Costa** (PS): — Seria inútil!

O **Orador**: — Infelizmente, os senhores nem sequer isso fizeram,...

O Sr. **António Costa** (PS): — Seria inútil!

O **Orador**: — ... porque, no fundo, também não estão virados para a questão das mais-valias. Hoje ficou claro que, também da parte do Partido Socialista, infelizmente, valem mais as pressões do grande capital...

Vozes do PS: — Eh!...

O **Orador**: — ... do que vale um sistema fiscal justo, em Portugal, ou um sistema fiscal equilibrado!

Aplausos do PCP.

Da nossa parte, Sr. Presidente, vamos obviamente continuar a bater-nos por um sistema fiscal justo e, na próxi-

ILUSTRAÇÃO 3 - IMAGEM DO INTERIOR DE UM DAR

3.4 Metodologia

No desenvolvimento deste modelo, privilegiou-se uma abordagem pós-positivista (quantitativa), com a medição de resultados depois da aplicação de um processo automático de categorização textual a um conjunto pré-definido de dados.

Elaborou-se um estudo de caso exploratório, com o intuito de desenvolver hipóteses e questões para uma análise futura (Yin 2003: 6). Com este estudo de caso (sendo o nosso "caso" constituído pelos próprios DAR), o intuito era dar resposta às seguintes questões de investigação:

- Como é que os Grupos Parlamentares se unem nas emoções manifestadas?
- Que relação existe entre a coesão na manifestação de emoções de dois Grupos Parlamentares e as suas orientações políticas (por exemplo, os partidos da esquerda e da direita aplaudem/protestam/riem/vozeiam sempre em conjunto)?
- Como é que as emoções transmitidas se articulam com o poder?
- Concretamente em relação à legislatura em análise (a IX), que diferenças existem entre o governo liderado por Durão Barroso (XV) e o dirigido por Santana Lopes (XVI)?
- Qual o grau de isolamento dos Grupos Parlamentares quando reagem emotivamente?
- Que relação existe entre o número de deputados representados e a capacidade de demonstrar emoções?

O estudo de caso implica uma "investigação empírica de um fenómeno contemporâneo particular inserido seu contexto real, nomeadamente quando as fronteiras entre o fenómeno e o contexto não são claramente evidentes" (Yin 2003: 13)[49]. O estudo de caso revela ser uma estratégia de investigação particularmente útil quando o intuito é dar resposta a questões do tipo "Porquê", "Que/Qual" e "Como" (Yin 2003:13), sendo por isso uma opção que se adequa perfeitamente às nossas questões de investigação.

3.4.1 Desenvolvimento do projecto

De acordo com o *Teragram TK240 User's Guide* (p. 165), são recomendáveis os seguintes passos para a criação de um *Rule Based Categorizer*:

a) Planificar uma Taxonomia para o projecto

[49] «A case study is an empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident».

b) Seleccionar o tipo de *Categorizer* (*Statistical* ou *Rule-Based*)

c) Criar as categorias

d) Seleccionar um conjunto de documentos de teste

e) Escrever as regras linguísticas

Descrevem-se, de seguida, a aplicação destas diferentes fases ao nosso protótipo.

3.4.2 Planificação da Taxonomia

Adoptou-se uma estrutura taxonómica hierarquizada (por oposição à taxonomia plana, ver ponto 3.2), com quatro categorias principais: 1) aplausos, 2) protestos, 3) risos e 4) vozes. Dentro de cada uma destas categorias, foram criadas 22 subcategorias. Estas correspondem à soma de:

- 15 Combinações possíveis entre os GP dois a dois, obtidas através da matriz combinatória ilustrada na figura 19[50];
- 6 Categorias (cada uma representando um GP);
- Uma última subcategoria que dá conta das reacções em bloco ("gerais").

	BE	CDS-PP	Os Verdes	PCP	PS	PSD
BE	BE/BE	BE/CDS-PP	BE/Os Verdes	BE/PCP	BE/PS	BE/PSD
CDS-PP	CDS-PP/BE	CDS-PP/CDS-PP	CDS-PP/Os Verdes	CDS-PP/PCP	CDS-PP/PS	CDS-PP/PSD
Os Verdes	Os Verdes/BE	Os Verdes/CDS-PP	Os Verdes/Os Verdes	Os Verdes/PCP	Os Verdes/PS	Os Verdes/PSD
PCP	PCP/BE	PCP/CDS-PP	PCP/Os Verdes	PCP/PCP	PCP/PS	PCP/PSD
PS	PS/BE	PS/CDS-PP	PS/Os Verdes	PS/PCP	PS/PS	PS/PSD
PSD	PSD/BE	PSD/CDS-PP	PSD/Os Verdes	PSD/PCP	PSD/PS	PSD/PSD

TABELA 3 - MATRIZ COMBINATÓRIA DOS GRUPOS PARLAMENTARES REPRESENTADOS EM AR, DOIS A DOIS

[50] Dado o tempo disponível para desenvolver este projecto (sensivelmente 12 meses), foram apenas tidas em conta as combinações dos Grupos Parlamentares dois a dois. A extensão desta análise a combinações de Grupos Parlamentares (três a três, quatro a quatro ou cinco a cinco) teria todo o interesse, mas prolongaria necessariamente o tempo de realização do projecto. Por uma questão de gestão de tempo, cingimo-nos à possibilidade apresentada.

Como resultado, obteve-se a taxonomia que se segue:

	Aplausos	Protestos	Risos	Vozes
1	Aplausos BE	Protestos BE	Risos BE	Vozes BE
2	Aplausos BE & CDS-PP	Protestos BE & CDS-PP	Risos BE & CDS-PP	Vozes BE & CDS-PP
3	Aplausos BE & Os Verdes	Protestos BE & Os Verdes	Risos BE & Os Verdes	Vozes BE & Os Verdes
4	Aplausos BE & PCP	Protestos BE & PCP	Risos BE & PCP	Vozes BE & PCP
5	Aplausos BE & PS	Protestos BE & PS	Risos BE & PSD	Vozes BE & PS
6	Aplausos BE & PSD	Protestos BE & PSD	Risos BE &PS	Vozes BE & PSD
7	Aplausos CDS-PP	Protestos CDS-PP & PS	Risos CDS-PP	Vozes CDS-PP
8	Aplausos CDS-PP&PS	Protestos Os Verdes	Risos CDS-PP&PS	Vozes CDS-PP & PS
9	Aplausos Gerais	Protestos Os Verdes & CDS-PP	Risos Gerais	Vozes Gerais
10	Aplausos Os Verdes	Protestos Os Verdes & PS	Risos Os Verdes	Vozes Os Verdes
11	Aplausos Os Verdes & CDS-PP	Protestos Os Verdes & PSD	Risos Os Verdes & CDS-PP	Vozes Os Verdes & CDS-PP
12	Aplausos Os Verdes & PS	Protestos CDS-PP	Risos Os Verdes & PS	Vozes Os Verdes & PS
13	Aplausos Os Verdes & PSD	Protestos PCP	Risos Os Verdes & PSD	Vozes Os Verdes & PSD
14	Aplausos PCP	Protestos PCP & CDS-PP	Risos PCP	Vozes PCP
15	Aplausos PCP & CDS-PP	Protestos PCP & Os Verdes	Risos PCP & CDS-PP	Vozes PCP & CDS-PP
16	Aplausos PCP & Os Verdes	Protestos PCP & PS	Risos PCP & Os Verdes	Vozes PCP & Os Verdes
17	Aplausos PCP & PS	Protestos PCP & PSD	Risos PCP & PS	Vozes PCP & PS
18	Aplausos PCP & PSD	Protestos PS	Risos PCP & PSD	Vozes PCP & PSD
19	Aplausos PS	Protestos PS & PSD	Risos PS	Vozes PS
20	Aplausos PS & PSD	Protestos PSD	Risos PS & PSD	Vozes PS & PSD
21	Aplausos PSD	Protestos PSD & CDS-PP	Risos PSD	Vozes PSD
22	Aplausos PSD & CDS-PP	Protestos Gerais	Risos PSD & CDS-PP	Vozes PSD & CDS-PP

TABELA 4 - TAXONOMIA CONSTITUÍDA POR QUATRO CATEGORIAS PRINCIPAIS, CADA UMA COM 22 SUBCATEGORIAS

3.4.3 Selecção do tipo de Categorizer

A escolha entre o *Statistical Categorizer* e o *Rule Based Categorizer* determina o modo como o programa constrói as categorias. O *Statistical Categorizer* (SC) é completamente automatizado. Depois de "treinado" com o conjunto de documentos fornecidos pelo utilizador, atribui automaticamente cada documento a uma categoria, com base na informação extraída do documento. Esta solução é recomendada para categorias não relacionadas (cf. *Teragram TK240 User's Guide*, p. 158).

Com o *Rule Based Categorizer* (RBC), são manualmente especificadas as regras que determinam a atribuição de um documento a uma dada categoria ou subcategoria. Este tipo de *categorizer* supõe um

maior controlo na construção das categorias e das suas regras e possibilita a alteração das regras de uma categoria, sem afectar as restantes.

O *Teragram TK240* apresenta ainda uma terceira solução, a ferramenta *Automatic Rule Generator* (ARG), que pretende ser uma solução intermédia entre os dois *categorizers* acima indicados, ao desenvolver automaticamente as regras linguísticas que criam as categorias na taxonomia, mas permitindo, simultaneamente, que estas sejam manualmente editadas.

Tendo em conta o esquema de categorias acima apresentado, optámos pelo RBC, uma vez que as categorias definidas estão relacionadas e é desejável um controlo mais eficaz e autónomo de cada categoria.

Ao utilizar o RBC, o projecto fica salvaguardado em termos de:

- Precisão: a capacidade de o RBC classificar documentos nas categorias esperadas é controlada pelo utilizador, pois é ele quem determina a pertença a uma categoria, através da escrita das regras;
- Restrição das regras: o utilizador pode optar por afinar as regras linguísticas, tornando-as mais restritas, com o intuito de reduzir as duplicações e de ganhar em precisão.

As regras podem ser restringidas se o utilizador optar por construir uma categoria de cada vez (o que não é possível com o SC, que obriga o utilizador a construir a totalidade da taxonomia antes da definição das regras linguísticas e do teste). Este método dá a possibilidade de se desenvolverem regras mais aprofundadas e restritas para cada categoria que integra a taxonomia.

As regras podem ainda ser afinadas se o utilizador optar por testar cada categoria, à medida que as vai criando. Não é necessário criar toda a taxonomia antes de testar as categorias. Esta possibilidade permite ao utilizador ter uma visão mais aprofundada dos resultados de teste para cada categoria e para localizar problemas que possam ocorrer nos requisitos de pertença a uma dada categoria, ou entre categorias, durante o processo de construção do projecto (cf. pp. 272 e 273, *Teragram TK240 User's Guide*).

3.4.4 Criação das Categorias

Cada categoria, secção constituída por um grupo de documentos que integra um esquema de classificação mais vasto (a taxonomia[51]), foi manualmente criada, de acordo com a taxonomia previamente desenvolvida

[51] Cf. *supra*, definição de taxonomia, capítulos 2.5. e 3.1.2.

e seguindo as recomendações apresentadas no manual de utilizador do programa *Teragram TK240*, relativamente à criação de categorias (p. 152):

1) Foram analisados os documentos para compreender o assunto, conteúdo, ou outros atributos que os documentos tinham em comum, tendo sido tirado partido da descrição sistemática e homogénea das reacções dos GP às intervenções dos deputados em AR;

2) Foram consideradas as necessidades dos utilizadores, tendo-se criado um modelo de categorização diferenciado da oferta já disponibilizada pelo sítio da AR, em livre acesso na página da Assembleia;

3) Os nomes das categorias foram criados tendo em conta a informação categorizada e procurando ser intuitivos e compreensíveis para qualquer utilizador interessado.

3.4.5 Constituição das Regras Linguísticas

Cada categoria tem de ser identificada de forma singular e limitada por um conjunto único de regras, integrando-se na totalidade da taxonomia, de modo a que a pertença dos documentos numa dada categoria seja precisa.

Deste modo, foram elaboradas regras (conjuntos de regras, palavras ou conjuntos de palavras, que definem cada categoria de forma única) que possibilitaram a atribuição dos documentos de *input* às categorias correspondentes. Estas regras foram criadas tendo em linha de conta as três funções fundamentais que lhes são atribuídas no *Teragram User's Guide* (pp. 282-283):

> 1) Descrever a categoria: as regras e modificadores são um conjunto de termos relacionais que identificam as ideias principais de cada categoria;
>
> 2) Localizar «identificadores únicos»: as regras que descrevem adequadamente uma categoria também devem definir e limitar com rigor a categoria sem, simultaneamente, excluir membros. «Identificadores únicos» são termos específicos de uma dada parte do sistema total de classificação. Estes termos descritivos separam uma categoria de todas as outras, quer sejam comparados numa base individual ou colectiva;
>
> 3) Limitar as regras das categorias: os termos únicos (descritivos e relacionais) também devem ser exclusivos por natureza. As regras têm de ser tão abrangentes quanto possível incluindo todas as características que definem uma categoria, mas suficientemente restritas para excluir membros não adequados da categoria respectiva.

Apresentando-se as reacções dos deputados invariavelmente descritas da mesma forma nos documentos em análise (*aplausos de X, protestos de X, risos de X* ou *vozes de X*), tirou-se partido desta homogeneidade sintáctico-linguística, e as regras foram criadas com base na mesma uniformidade.

Dentro de cada subcategoria elaboraram-se regras simples, constituídas pelo substantivo descritivo da emoção (aplausos/protestos/risos/vozes) + o(s) partidos políticos que identificam a respectiva subcategoria. Estas regras reproduzem exactamente o texto patente nos DAR, apresentando as duas possibilidades de ordem dos partidos políticos - por exemplo, se queremos testar a consonância de PCP e PSD, temos de considerar as duas hipóteses de ocorrências de "PCP", antes e depois da conjunção copulativa "e": *Aplausos (1) do PCP* ***e*** *(2)do PSD* ou *Aplausos (2)do PSD* ***e*** *(1)do PCP*.

As duas ordenações possíveis são assim repetidas quatro vezes, apenas com modificação do sinal de pontuação, sendo considerados os casos com ponto final (.), ponto e vírgula (;), dois pontos (:) e vírgula (,). Assegura-se, desta forma, que o programa não contabiliza os casos em que os GP têm reacções três a três, quatro a quatro ou cinco cinco, afinando-se a qualidade da análise.

A título de exemplo, vejam-se as regras criadas para as subcategorias indicadas na tabela em baixo:

Categoria	Aplausos	Protestos	Risos	Vozes
Subcat.	BE & Os Verdes	PS & PSD	PCP & PSD	CDS-PP&PS
Regras	Aplausos do BE e de Os Verdes. Aplausos de Os Verdes e do BE.	Protestos do PS e do PSD. Protestos do PSD e do PS.	Risos do PCP e do PSD. Risos do PSD e do PCP.	Vozes do CDS-PP e do PS. Vozes do PS e do CDS-PP.
	Aplausos do BE e de Os Verdes; Aplausos de Os Verdes e do BE;	Protestos do PS e do PSD; Protestos do PSD e do PS;	Risos do PCP e do PSD; Risos do PSD e do PCP;	Vozes do CDS-PP e do PS; Vozes do PS e do CDS-PP;
	Aplausos do BE e de Os Verdes: Aplausos de Os Verdes e do BE:	Protestos do PS e do PSD: Protestos do PSD e do PS:	Risos do PCP e do PSD: Risos do PSD e do PCP:	Vozes do CDS-PP e do PS: Vozes do PS e do CDS-PP:
	Aplausos do BE e de Os Verdes, Aplausos de Os Verdes e do BE,	Protestos do PS e do PSD, Protestos do PSD e do PS,	Risos do PCP e do PSD, Risos do PSD e do PCP,	Vozes do CDS-PP e do PS, Vozes do PS e do CDS-PP,

TABELA 5 - EXEMPLO DAS REGRAS CRIADAS PARA CADA UMA DAS SUCATEGORIAS DO MODELO

No caso das subcategorias mono partidárias, optou-se pela repetição da mesma regra, retirando-se apenas as ocorrências com vírgula, para assegurar que só seriam extraídas as ocorrências dos partidos isolodamente, pois doutra forma seriam contabilizados casos como *Aplausos do BE, Os Verdes e CDS-PP* (note-se que, no caso das combinações dos partidos 2 a 2, a vírgula pode manter-se, pois os partidos estão sempre unidos entre si pela conjunção copulativa "e").

Categoria	Aplausos	Protestos	Risos	Vozes
Subcat.	**BE**	**PS**	**PCP & PSD**	**CDS-PP&PS**
Regras	Aplausos do BE.	Protestos do PS.	Risos do PCP.	Vozes do CDS-PP.
	Aplausos do BE;	Protestos do PS;	Risos do PCP;	Vozes do CDS-PP;
	Aplausos do BE:	Protestos do PS:	Risos do PCP:	Vozes do CDS-PP:

TABELA 6 - EXEMPLO DE REGRAS CRIADAS PARA CADA UMA DAS SUBCATEGORIAS MONOPARTIDÁRIAS

3.4.6 Selecção dos Documentos de Teste

Para testar a fiabilidade das regras criadas e manualmente introduzidas em cada uma das 88 subcategorias, foram seleccionados aleatoriamente 10 documentos *html*:

- S1L9SL1N2 -0019
- S1L9SL1N3 -0052
- S1L9SL1N40 -1646
- S1L9SL1N54 -2235
- S1L9SL2N11 -0532
- S1L9SL3N1 -0048
- S1L9SL3N2 -0094
- S1L9SL3N11 -0575
- S1L9SL3N23 -1432
- S1L9SL3N23 -1440

3.4.7 Teste das Regras Linguísticas

Ao criar regras que reproduzem exactamente o texto presente nos documentos, a expectativa foi a de que o RBC criado com o *Teragram TK240* incluísse, numa dada categoria, documentos que fossem ao encontro dos critérios estabelecidos essa categoria e, simultaneamente, excluísse documentos que satisfizessem os critérios de outras categorias. Retomando os exemplos anteriores, esperava-se que, dentro da subcategoria "BE&Verdes", na categoria "Aplausos", fossem classificados apenas os documentos *html* onde estivesse presente pelo menos uma ocorrência textual "Aplausos do BE e de Os Verdes" ou "Aplausos de Os Verdes e do BE".

Os dez documentos de teste foram manualmente classificados em cada categoria, com o auxílio da ferramenta *find and replace* do Microsoft Word, antes de se realizar o teste da taxonomia criada com o RBC

do *Teragram TK240*. Os resultados obtidos com a classificação manual foram depois cotejados com os frutos do processamento automático destes documentos com o *Teragram TK240*, revelando-se exactamente coincidentes, de acordo com a distribuição que pode ser consultada na tabela seguinte:

		Categorias		
Ficheiros Teste	S1L9SL1N2-0019	Aplausos Gerais		
	S1L9SL1N3-0052	Aplausos PSD & CDS-PP		
	S1L9SL1N40-1646	Aplausos PSD & CDS-PP		
	S1L9SL1N54-2235	Aplausos PSD & CDS-PP	Protestos PS	Vozes PS
	S1L9SL2N11-0532	Aplausos PSD & CDS-PP	Vozes PSD	
	S1L9SL3N1-0048	Aplausos PSD & CDS-PP		
	S1L9SL3N11-0575	Aplausos PCP & OsVerdes	Vozes BE	Vozes PCP
	S1L9SL3N2-0094	Aplausos PS	Vozes CDS-PP	
	S1L9SL3N23-1432	Protestos PSD	Vozes PCP	
	S1L9SL3N23-1440	Aplausos Gerais	Risos PCP & OsVerdes	Vozes PCP

TABELA 7 - RESULTADOS DA CATEGORIZAÇÃO AUTOMATIZADA DOS DOCUMENTOS DE TESTE COM O PROGRAMA TK240

A categorização correcta, pelo programa *Teragram TK240*, dos dez documentos de teste, permitiu-nos aferir a eficácia das regras linguísticas criadas. Com efeito, de acordo com o manual do utilizador do *Teragram TK 240*, os resultados são tanto mais precisos quanto maior for a percentagem de documentos bem categorizados. A obtenção de 100% de sucesso na classificação dos documentos de teste foi possível devido à simplicidade das regras, ao facto de estas reproduzirem exactamente o conteúdo dos ficheiros, e à homogeneidade dos documentos analisados, tendo sido potenciada pela utilização do *Rule Based Categorizer*, que assegura maior controlo ao permitir a escrita individual das regras linguísticas.

4 Capítulo 4

4.1 Resultados

Os 13520 ficheiros html foram processados em duas fases distintas, correspondentes a cada um dos governos constitucionais, de acordo com a divisão estabelecida na Tabela 1 (ver *supra,* p. 41).

Após processamento dos dados no programa Teragram TK240, seguindo a estrutura taxonómica previamente descrita, obtiveram-se os resultados apresentados nas páginas que se seguem, primeiro considerando a totalidade da IX Legislatura e, de seguida, individualizados em termos de governos constitucionais.

IX Legislatura

Aplausos	
PSD&CDS-PP	5167
PS	2495
PCP	897
BE	353
PSD	326
CDS-PP	194
Gerais	119
PCP&OsVerdes	107
BE&PCP	64
BE&PS	53
PS&PSD	22
PCP&PS	18
OsVerdes	7
BE&OsVerdes	4
CDS-PP&PS	3
PCP&PSD	2
OsVerdes&PS	1
BE&CDS-PP	0
BE&PSD	0
OsVerdes&CDS-PP	0
OsVerdes&PSD	0
PCP&CDS-PP	0
Total	**9832**

Protestos	
PS	653
PSD&CDS-PP	422
PSD	382
PCP	194
CDS-PP	103
PCP&PS	78
BE&PCP	28
BE	19
PCP&OsVerdes	14
OsVerdes	8
BE&PS	6
OsVerdes&PS	3
BE&CDS-PP	0
BE&OsVerdes	0
BE&PSD	0
CDS-PP&PS	0
Gerais	0
OsVerdes&CDS-PP	0
OsVerdes&PSD	0
PCP&CDS-PP	0
PCP&PSD	0
PS&PSD	0
Total	**1910**

Risos	
PS	300
PSD&CDS-PP	298
PCP	121
PSD	109
PCP&PS	66
CDS-PP	62
BE&PCP	31
BE	28
BE&PS	13
PCP&OsVerdes	13
Gerais	5
OsVerdes	2
BE&OsVerdes	1
OsVerdes&PS	1
PS&PSD	1
BE&CDS-PP	0
BE&PSD	0
CDS-PP&PS	0
OsVerdes&CDS-PP	0
OsVerdes&PSD	0
PCP&CDS-PP	0
PCP&PSD	0
Total	**1051**

Vozes	
PS	2208
PSD	2203
PCP	1569
CDS-PP	1358
PSD&CDS-PP	1331
BE	335
PCP&OsVerdes	112
BE&PCP	111
PCP&PS	97
OsVerdes	42
BE&PS	26
PS&PSD	18
BE&OsVerdes	8
CDS-PP&PS	6
PCP&PSD	4
OsVerdes&PS	1
PCP&CDS-PP	1
BE&CDS-PP	0
BE&PSD	0
Gerais	0
OsVerdes&CDS-PP	0
OsVerdes&PSD	0
Total	**9430**

TABELA 8 - RESULTADOS DO PROCESSAMENTO AUTOMÁTICO DOS 13520 FICHEIROS HTML

Aplausos	
PSD&CDS-PP	53%
PS	25%
PCP	9%
BE	4%
PSD	3%
CDS-PP	2%
Gerais	1%
PCP&OsVerdes	1%
BE&PCP	1%
BE&PS	1%
PS&PSD	0%
PCP&PS	0%
OsVerdes	0%
BE&OsVerdes	0%
CDS-PP&PS	0%
PCP&PSD	0%
OsVerdes&PS	0%
BE&CDS-PP	0%
BE&PSD	0%
OsVerdes&CDS-PP	0%
OsVerdes&PSD	0%
PCP&CDS-PP	0%
Total	**100%**

Protestos	
PS	34%
PSD&CDS-PP	22%
PSD	20%
PCP	10%
CDS-PP	5%
PCP&PS	4%
BE&PCP	1%
BE	1%
PCP&OsVerdes	1%
OsVerdes	0%
BE&PS	0%
OsVerdes&PS	0%
BE&CDS-PP	0%
BE&OsVerdes	0%
BE&PSD	0%
CDS-PP&PS	0%
Gerais	0%
OsVerdes&CDS-PP	0%
OsVerdes&PSD	0%
PCP&CDS-PP	0%
PCP&PSD	0%
PS&PSD	0%
Total	**100%**

Risos	
PS	29%
PSD&CDS-PP	28%
PCP	12%
PSD	10%
PCP&PS	6%
CDS-PP	6%
BE&PCP	3%
BE	3%
BE&PS	1%
PCP&OsVerdes	1%
Gerais	0%
OsVerdes	0%
BE&OsVerdes	0%
OsVerdes&PS	0%
PS&PSD	0%
BE&CDS-PP	0%
BE&PSD	0%
CDS-PP&PS	0%
OsVerdes&CDS-PP	0%
OsVerdes&PSD	0%
PCP&CDS-PP	0%
PCP&PSD	0%
Total	**100%**

Vozes	
PS	23%
PSD	23%
PCP	17%
CDS-PP	14%
PSD&CDS-PP	14%
BE	4%
PCP&OsVerdes	1%
BE&PCP	1%
PCP&PS	1%
OsVerdes	0%
BE&PS	0%
PS&PSD	0%
BE&OsVerdes	0%
CDS-PP&PS	0%
PCP&PSD	0%
OsVerdes&PS	0%
PCP&CDS-PP	0%
BE&CDS-PP	0%
BE&PSD	0%
Gerais	0%
OsVerdes&CDS-PP	0%
OsVerdes&PSD	0%
Total	**100%**

TABELA 9 - APRESENTAÇÃO PERCENTUAL DOS RESULTADOS DA TABELA 8.

Para o governo liderado por Durão Barroso, os valores foram os que se seguem:

XV Governo Constitucional

Aplausos	
PSD&CDS-PP	4539
PS	2213
PCP	804
BE	318
PSD	298
CDS-PP	172
Gerais	106
PCP&OsVerdes	86
BE&PCP	62
BE&PS	51
PS&PSD	21
PCP&PS	18
OsVerdes	7
BE&OsVerdes	3
CDS-PP&PS	3
PCP&PSD	1
BE&CDS-PP	0
BE&PSD	0
OsVerdes&CDS-PP	0
OsVerdes&PS	0
OsVerdes&PSD	0
PCP&CDS-PP	0
Total	**8702**

Protestos	
PS	595
PSD&CDS-PP	364
PSD	334
PCP	178
CDS-PP	98
PCP&PS	73
BE&PCP	26
BE	17
PCP&OsVerdes	13
OsVerdes	7
BE&PS	6
OsVerdes&PS	3
BE&CDS-PP	0
BE&OsVerdes	0
BE&PSD	0
CDS-PP&PS	0
Gerais	0
OsVerdes&CDS-PP	0
OsVerdes&PSD	0
PCP&CDS-PP	0
PCP&PSD	0
PS&PSD	0
Total	**1714**

Risos	
PS	270
PSD&CDS-PP	255
PCP	106
PSD	98
PCP&PS	61
CDS-PP	55
BE&PCP	23
BE	21
BE&PS	10
PCP&OsVerdes	10
Gerais	2
OsVerdes	2
BE&OsVerdes	1
OsVerdes&PS	1
PS&PSD	1
BE&CDS-PP	0
BE&PSD	0
CDS-PP&PS	0
OsVerdes&CDS-PP	0
OsVerdes&PSD	0
PCP&CDS-PP	0
PCP&PSD	0
Total	**916**

Vozes	
PSD	1993
PS	1987
PCP	1424
CDS-PP	1227
PSD&CDS-PP	1223
BE	302
BE&PCP	102
PCP&PS	93
PCP&OsVerdes	90
OsVerdes	37
BE&PS	24
PS&PSD	17
BE&OsVerdes	7
CDS-PP&PS	6
PCP&PSD	4
OsVerdes&PS	1
PCP&CDS-PP	1
BE&CDS-PP	0
BE&PSD	0
Gerais	0
OsVerdes&CDS-PP	0
OsVerdes&PSD	0
Total	**8538**

TABELA 10 - RESULTADOS DO PROCESSAMENTO DOS FICHEIROS HTML CORRESPONDENTES À TOTALIDADE DO XV GC

Aplausos	
PSD&CDS-PP	52%
PS	25%
PCP	9%
BE	4%
PSD	3%
CDS-PP	2%
Gerais	1%
PCP&OsVerdes	1%
BE&PCP	1%
BE&PS	1%
PS&PSD	0%
PCP&PS	0%
OsVerdes	0%
BE&OsVerdes	0%
CDS-PP&PS	0%
PCP&PSD	0%
BE&CDS-PP	0%
BE&PSD	0%
OsVerdes&CDS-PP	0%
OsVerdes&PS	0%
OsVerdes&PSD	0%
PCP&CDS-PP	0%
Total	**100%**

Protestos	
PS	35%
PSD&CDS-PP	21%
PSD	19%
PCP	10%
CDS-PP	6%
PCP&PS	4%
BE&PCP	2%
BE	1%
PCP&OsVerdes	1%
OsVerdes	0%
BE&PS	0%
OsVerdes&PS	0%
BE&CDS-PP	0%
BE&OsVerdes	0%
BE&PSD	0%
CDS-PP&PS	0%
Gerais	0%
OsVerdes&CDS-PP	0%
OsVerdes&PSD	0%
PCP&CDS-PP	0%
PCP&PSD	0%
PS&PSD	0%
Total	**100%**

Risos	
PS	29%
PSD&CDS-PP	28%
PCP	12%
PSD	11%
PCP&PS	7%
CDS-PP	6%
BE&PCP	3%
BE	2%
BE&PS	1%
PCP&OsVerdes	1%
Gerais	0%
OsVerdes	0%
BE&OsVerdes	0%
OsVerdes&PS	0%
PS&PSD	0%
BE&CDS-PP	0%
BE&PSD	0%
CDS-PP&PS	0%
OsVerdes&CDS-PP	0%
OsVerdes&PSD	0%
PCP&CDS-PP	0%
PCP&PSD	0%
Total	**100%**

Vozes	
PSD	23%
PS	23%
PCP	17%
CDS-PP	14%
PSD&CDS-PP	14%
BE	4%
BE&PCP	1%
PCP&PS	1%
PCP&OsVerdes	1%
OsVerdes	0%
BE&PS	0%
PS&PSD	0%
BE&OsVerdes	0%
CDS-PP&PS	0%
PCP&PSD	0%
OsVerdes&PS	0%
PCP&CDS-PP	0%
BE&CDS-PP	0%
BE&PSD	0%
Gerais	0%
OsVerdes&CDS-PP	0%
OsVerdes&PSD	0%
Total	**100%**

TABELA 11 - APRESENTAÇÃO PERCENTUAL DOS RESULTADOS DA TABELA 10.

E, por último, estes foram os valores resultantes do processamento dos dados relativos ao governo liderado por Pedro Santana Lopes:

XVI Governo Constitucional

Aplausos	
PSD & CDS-PP	628
PS	282
PCP	93
BE	35
PSD	28
CDS-PP	22
PCP & OsVerdes	21
Gerais	13
BE & PCP	2
BE & PS	2
BE & OsVerdes	1
OsVerdes & PS	1
PCP & PSD	1
PS & PSD	1
BE & CDS-PP	0
BE & PSD	0
CDS-PP & PS	0
OsVerdes	0
OsVerdes & CDS-PP	0
OsVerdes & PSD	0
PCP & CDS-PP	0
PCP & PS	0
Total	**1130**

Protestos	
PS	58
PSD & CDS-PP	58
PSD	48
PCP	16
CDS-PP	5
PCP & PS	5
BE	2
BE & PCP	2
OsVerdes	1
PCP & OsVerdes	1
BE & CDS-PP	0
BE & OsVerdes	0
BE & PS	0
BE & PSD	0
CDS-PP & PS	0
Gerais	0
OsVerdes & CDS-PP	0
OsVerdes & PS	0
OsVerdes & PSD	0
PCP & CDS-PP	0
PCP & PSD	0
PS & PSD	0
Total	**196**

Risos	
PSD & CDS-PP	43
PS	30
PCP	15
PSD	11
BE & PCP	8
BE	7
CDS-PP	7
PCP&PS	5
BE & PS	3
Gerais	3
PCP & OsVerdes	3
BE & CDS-PP	0
BE & OsVerdes	0
BE & PSD	0
CDS-PP & PS	0
OsVerdes	0
OsVerdes & CDS-PP	0
OsVerdes & PS	0
OsVerdes & PSD	0
PCP & CDS-PP	0
PCP & PSD	0
PS & PSD	0
Total	**135**

Vozes	
PS	221
PSD	210
PCP	145
CDS-PP	131
PSD & CDS-PP	108
BE	33
PCP & OsVerdes	22
BE & PCP	9
OsVerdes	5
PCP & PS	4
BE & PS	2
BE & OsVerdes	1
PS & PSD	1
BE & CDS-PP	0
BE & PSD	0
CDS-PP & PS	0
Gerais	0
OsVerdes & CDS-PP	0
OsVerdes & PS	0
OsVerdes & PSD	0
PCP & CDS-PP	0
PCP & PSD	0
Total	**892**

TABELA 12 - RESULTADOS OBTIDOS APÓS O PROCESSAMENTO DOS FICHEIROS HTML, CORRESPONDENTES À TOTALIDADE DOS DAR DO XVI GC.

Aplausos	
PSD & CDS-PP	56%
PS	25%
PCP	8%
BE	3%
PSD	2%
CDS-PP	2%
PCP & OsVerdes	2%
Gerais	1%
BE & PCP	0%
BE & PS	0%
BE & OsVerdes	0%
OsVerdes & PS	0%
PCP & PSD	0%
PS & PSD	0%
BE & CDS-PP	0%
BE & PSD	0%
CDS-PP & PS	0%
OsVerdes	0%
OsVerdes & CDS-PP	0%
OsVerdes & PSD	0%
PCP & CDS-PP	0%
PCP & PS	0%
Total	**100 %**

Protestos	
PS	30%
PSD & CDS-PP	30%
PSD	24%
PCP	8%
CDS-PP	3%
PCP & PS	3%
BE	1%
BE & PCP	1%
OsVerdes	1%
PCP & OsVerdes	1%
BE & CDS-PP	0%
BE & OsVerdes	0%
BE & PS	0%
BE & PSD	0%
CDS-PP & PS	0%
Gerais	0%
OsVerdes & CDS-PP	0%
OsVerdes & PS	0%
OsVerdes & PSD	0%
PCP & CDS-PP	0%
PCP & PSD	0%
PS & PSD	0%
Total	**100 %**

Risos	
PSD & CDS-PP	32%
PS	22%
PCP	11%
PSD	8%
BE & PCP	6%
BE	5%
CDS-PP	5%
PCP&PS	4%
BE & PS	2%
Gerais	2%
PCP & OsVerdes	2%
BE & CDS-PP	0%
BE & OsVerdes	0%
BE & PSD	0%
CDS-PP & PS	0%
OsVerdes	0%
OsVerdes & CDS-PP	0%
OsVerdes & PS	0%
OsVerdes & PSD	0%
PCP & CDS-PP	0%
PCP & PSD	0%
PS & PSD	0%
Total	**100 %**

Vozes	
PS	25%
PSD	24%
PCP	16%
CDS-PP	15%
PSD & CDS-PP	12%
BE	4%
PCP & OsVerdes	2%
BE & PCP	1%
OsVerdes	1%
PCP & PS	0%
BE & PS	0%
BE & OsVerdes	0%
PS & PSD	0%
BE & CDS-PP	0%
BE & PSD	0%
CDS-PP & PS	0%
Gerais	0%
OsVerdes & CDS-PP	0%
OsVerdes & PS	0%
OsVerdes & PSD	0%
PCP & CDS-PP	0%
PCP & PSD	0%
Total	**100 %**

TABELA 13 - APRESENTAÇÃO PERCENTUAL DOS RESULTADOS APRESENTADOS NA TABELA 12.

4.2 Categorias

Vejamos, em primeiro lugar, as quatro categorias consideradas na análise dos resultados (aplausos, protestos, risos e vozes), ao que é que poderão corresponder estas descrições presentes nos DAR, e de que forma devem ser consideradas na nossa análise.

Com efeito, os substantivos "aplausos" e "protestos" correspondem a uma descrição clara de uma reacção (de agrado ou desagrado) por parte de um ou mais GP, na sequência da intervenção de um deputado.

Já os nomes "risos" e "vozes" apontam para descrições mais vagas. Por exemplo, os “risos” de um GP podem ocorrer na sequência de uma proposta política, como forma de ridicularização da mesma, mas também podem derivar de uma gafe na fala ou de um gesto menos apropriado. Por seu lado, o substantivo "vozes" também é ambíguo. É verdade que assinala, indubitavelmente, a reacção de um ou mais GP – se há “vozes”, significa que os GP não ficaram indiferentes a determinado facto ou intervenção – mas fica por esclarecer se estas vozes são de apoio ou de contestação.

Sendo assim, na análise, consideramos estas categorias duas a duas: aplausos e protestos, por um lado; risos e vozes, por outro.

4.3 Representatividade dos grupos parlamentares

Ainda antes da análise dos resultados individuais, atentemos na distribuição dos GP na AR, considerando a representatividade de cada um dos GP:

Grupo Parlamentar	Número de Deputados
PSD	105
PS	96
CDS-PP	14
PCP	10
BE	3
OsVerdes	2
Total	**230**

TABELA 14 - DEPUTADOS EM AR NA IX LEGISLATURA

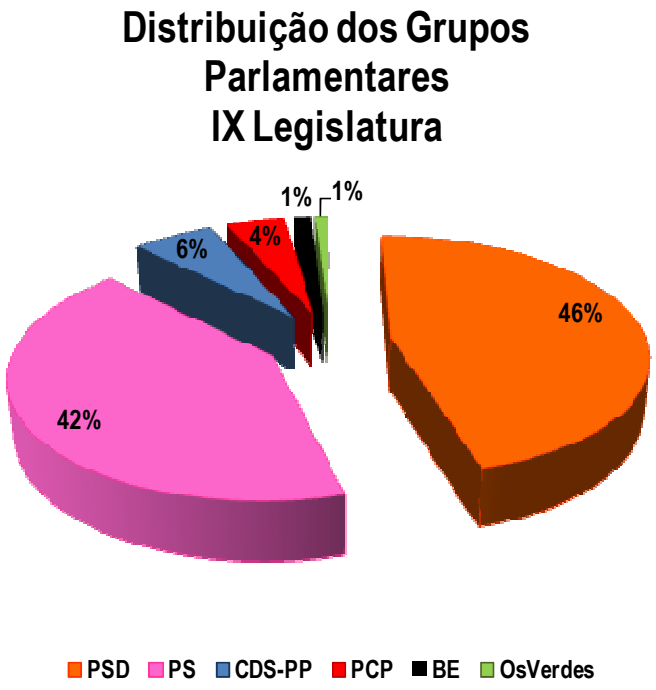


ILUSTRAÇÃO 4 - GRÁFICO REPRESENTATIVO DOS RESULTADOS ANTERIORES

4.4 Análise de Resultados Monopartidários

Os resultados obtidos com a análise individual dos partidos devem ser vistos com alguma prudência, uma vez que a ferramenta utilizada não permite contabilizar os deputados que efectivamente estiveram presentes em cada sessão e correlacionar o número de deputados com a acção demonstrada pelo respectivo GP. Assim, não é possível dar uma resposta segura à última das questões de investigação acima apresentadas (ver 3.4.): há relação entre o número de deputados representados em Assembleia e a capacidade de demonstrar emoções?

Com efeito, esta análise parte do princípio que os deputados dos GP beneficiaram da possibilidade que lhes foi conferida para estarem presentes em AR.

4.4.1 Aplausos & Protestos

As manifestações de agrado/ desagrado são consistentes nos dois governos da nona legislatura.

Tendo como base a totalidade dos resultados (reacções individualmente e reacções em conjunto, Tabela 10, Tabela 11, Tabela 12 e Tabela 13), analisemos isoladamente os aplausos / protestos de cada GP nos dois governos constitucionais[52]:

XV Governo Constitucional

Aplausos	
PS	25%
PCP	9%
BE	4%
PSD	3%
CDS-PP	2%
OsVerdes	0%

Protestos	
PS	35%
PSD	19%
PCP	10%
CDS-PP	6%
BE	1%
OsVerdes	0%

TABELA 15 - APRESENTAÇÃO PERCENTUAL DAS REACÇÕES INDIVIDUAIS DE CADA GP NO XV GC, CONSIDERANDO O UNIVERSO TOTAL DAS REACÇÕES

XVI Governo Constitucional

Aplausos	
PS	25%
PCP	8%
BE	3%
PSD	2%
CDS-PP	2%
OsVerdes	0%

Protestos	
PS	30%
PSD	24%
PCP	8%
CDS-PP	3%
BE	1%
OsVerdes	1%

TABELA 16 - APRESENTAÇÃO PERCENTUAL DAS REACÇÕES INDIVIDUAIS DE CADA GP NO XV GC, CONSIDERANDO O UNIVERSO TOTAL DAS REACÇÕES

O PS lidera este "ranking", em aplausos e em protestos, demonstrando uma participação activa. Podemos constatar que a ordenação destas categorias se mantém em ambos os Governos Constitucionais, com ligeiras oscilações em termos de valores percentuais. Na categoria "aplausos", PS é seguido de PCP e de

[52] Estas tabelas apresentam apenas os valores de cada GP, retirados das Tabela 11 e Tabela 13, onde podem ser consultados os resultados totais.

BE, com o PSD e o CDS-PP a fechar a tabela, não tendo Os Verdes uma participação significativamente representativa. Já no caso dos "protestos", o PSD sucede imediatamente o PS, com uma percentagem ligeiramente mais elevada no XVI GC, seguido pelo PCP, o CDS-PP, o BE e Os Verdes.

Relativamente a esta distribuição, saliente-se:

1. Nos dois Governos Constitucionais destaca-se a presença do PS, demonstrando dinamismo interventivo enquanto líder da oposição, pois a esta presença marcante corresponde uma elevada percentagem de deputados em assembleia (42%);
2. Com a possibilidade de apresentar apenas 1% dos deputados em AR, a percentagem de aplausos representada pelo BE foi ligeiramente superior à do PSD isoladamente, com 46% dos deputados;
3. É de salientar ainda a presença do PCP, pois apesar de contar com apenas 5% de deputados na IX Legislatura, sucede imediatamente o PS na categoria "aplausos" e ocupa uma posição igualmente interessante na categoria protestos.

Tendo a liderança dos Governos abrangidos por esta legislatura sido assumida, precisamente, pelo PSD, em coligação com o CDS-PP, a fraca percentagem de aplausos do PSD isoladamente não pode ser separada do elevado valor que a coligação PSD / CDS-PP assume nesta categoria, com valores superiores ao PS nos dois Governos Constitucionais (52% no XV e 56% no XVI, ver Tabela 11 e Tabela 13).

4.4.2 Risos & Vozes

XV Governo Constitucional

Risos	
PS	29%
PCP	12%
PSD	11%
CDS-PP	6%
BE	2%
OsVerdes	0%

Vozes	
PSD	23%
PS	23%
PCP	17%
CDS-PP	14%
BE	4%
OsVerdes	0%

TABELA 17 - APRESENTAÇÃO PERCENTUAL DAS REACÇÕES INDIVIDUAIS DE CADA GP NO XV GC, CONSIDERANDO O UNIVERSO TOTAL DAS REACÇÕES

XVI Governo Constitucional

Risos	
PS	22%
PCP	11%
PSD	8%
BE	5%
CDS-PP	5%
OsVerdes	0%

Vozes	
PS	25%
PSD	24%
PCP	16%
CDS-PP	15%
BE	4%
OsVerdes	1%

TABELA 18 - APRESENTAÇÃO PERCENTUAL DAS REACÇÕES INDIVIDUAIS DE CADA GP NO XVI GC, CONSIDERANDO O UNIVERSO TOTAL DAS REACÇÕES

Na categoria "Risos", a ordenação é praticamente idêntica nos dois Governos Constitucionais, com o PS a apresentar a maior fatia de reacções individuais, seguido pelo PCP e depois pelo PSD. Em quarto lugar temos o CDS-PP, e depois o BE, no XV GC; já no XVI, BE e CDS-PP ocupam ex-aequo a quarta posição. Os Verdes não têm representação individual na categoria "risos" em nenhum dos casos.

Já na categoria "Vozes", PSD e PS apresentam valores muito similares nos dois Governos Constitucionais, liderando o PSD no XV e o PS no XVI. A ordem e os valores que os sucedem são coincidentes em ambos os governos, seguindo-se o PCP, o CDS-PP, o BE e Os Verdes, este último, mais uma vez, com uma participação residual.

Desta análise, destaca-se mais uma vez a forte presença e dinamismo do PS, mas também a elevada representação do PCP, tendo em vista a baixa representatividade parlamentar. Mais uma vez, os valores mais baixos apresentados pelo PSD para a categoria "risos" não podem ser separados dos valores que este partido apresenta em coligação com o CDS-PP – 28% no XV GC, um ponto percentual abaixo do PS, e 32% no XVI, com mais 10% do que o principal partido da oposição.

4.5 Análise da Prestação Global de Cada Grupo Parlamentar

Vejamos agora qual a prestação global dos GP em cada uma das quatro categorias, em ambos os Governos Constitucionais[53]. Para este efeito, os resultados foram trabalhados por forma a facilitar a sua compreensão e leitura.

Deste modo, nas tabelas que se seguem, consideram-se vários indicadores:

- Por "**Total do GP**" entende-se o conjunto de todas as reacções do GP, ou seja, é apresentada a percentagem correspondente à soma das reacções do GP isolado com todas as reacções desse mesmo GP com os restantes partidos, em relação ao valor global das reacções da categoria respectiva. Vejamos o exemplo para o GP Z: GPZ Isolado + (GPZ + GP1) + (GPZ + GP2) + (GPZ +GP3) + (GPZ + GP4) + (GPZ + GP5) / Total das reacções.
- Em "**GP isolado/ total das reacções**" é apresentado o valor percentual do GP isolado / universo total das reacções (ou seja, GP isolados, GP 2 a 2 e reacções gerais), procurando-se aferir a capacidade reactiva de um dado GP isolado no universo total das reacções consideradas.
- O "**Indicador de Isolamento**" foi obtido a partir do valor absoluto das participações de um dado GP/ Total do GP. Deste modo, quanto maior for o índice de isolamento, menos são as situações em que

[53] As tabelas com os valores absolutos destes resultados podem consultar-se no ponto 7.5

um GP se associa a outros partidos nas reacções manifestadas. Podemos considerá-lo um índice do "carisma" ou do "carácter" emotivo de um GP.

- Por fim, considerando o número de situações em que um GP se associa a outros partidos, analisa-se ainda que percentagem corresponde a uma **união à esquerda** e que percentagem equivale a uma associação **à direita**. Como GPs de esquerda considerámos o PS, o PCP, o BE e Os Verdes; enquanto GP de direita o PSD e o CDS-PP.

4.5.1 Aplausos

XV Governo Constitucional

		PSD	PS	CDS-PP	PCP	BE	Os Verdes
Total GP		56%	26%	54%	11%	5%	1%
GP isolado/ total das reacções		3%	25%	2%	9%	4%	0%
Indicador de Isolamento		6%	96%	4%	83%	73%	7%
União	**Esquerda**	0%	74%	0%	99%	100%	100%
	Direita	100%	26%	100%	1%	0%	0%

XVI Governo Constitucional

		PSD	PS	CDS-PP	PCP	BE	Os Verdes
Total GP		58%	25%	58%	10%	4%	2%
GP isolado/ total das reacções		2%	25%	2%	8%	3%	0%
Indicador de Isolamento		4%	99%	3%	79%	88%	0%
União	**Esquerda**	0%	75%	0%	96%	100%	100%
	Direita	100%	25%	100%	4%	0%	0%

Na categoria aplausos, saliente-se a forte capacidade participativa da coligação com incidência parlamentar, constituída por PSD e CDS-PP, nos dois Governos Constitucionais – o PSD apresenta 56% e 58% e o CDS-PP 54% e 58%, no XV e XVI Governos, respectivamente. No entanto, esta corresponde a uma fraca aptidão para aplaudir isoladamente, por parte de cada um dos partidos que a constitui – o PSD apresenta valores de 3% no XV GC e de 2% no XVI; e o CDS-PP não vai além dos 2% nos dois governos. Se, a este facto, somarmos o fraco indicador de isolamento de cada um dos partidos (6% e 4% para o PSD; 4% e 3% para o CDS-PP) e os 100% de união à direita (o que, neste caso, equivale a uma união bilateral destes dois GP, pois são os únicos considerados de direita), compreendemos que na grande maioria dos casos em que aplaudiram, PSD e CDS-PP o fizeram em conjunto, sem diferenças significativas no governo liderado por Durão Barroso e no governo liderado por Santana Lopes.

Observando agora os resultados do PS, constatamos que a percentagem total de aplausos deste GP é considerável (o segundo GP a seguir a PSD e a CDS-PP), sobretudo se tivermos em conta o elevado indicador de isolamento deste GP nos dois governos constitucionais: 96% e 99%, respectivamente. Ou seja, na grande maioria das situações em que aplaudiu, o GP do PS fê-lo sozinho. Quando acompanhado, teve

prevalentemente o apoio dos GP à esquerda, embora também tenha tido alguns momentos de coincidência com a direita.

Considerando a pequena representativade do PCP na AR, notem-se os 10% e 11% de aplausos em conjunto com os outros GP e os 9% e 8% isoladamente. Apesar de não ter um perfil tão "individualista" como o PS, o GP do PCP aplaude prevalentemente sozinho (com 83% no XV GC e 79% no XVI GC); quando aplaude em conjunto, fá-lo quase exclusivamente com a esquerda.

BE e Os Verdes têm fraca presença (com 5% e 4% de aplausos totais para o BE e com 1% e 2% para Os Verdes, no XV e XVI GC). No entanto, o BE distingue-se de Os Verdes pelo seu "carisma", pois nos dois GC o número de vezes que este GP aplaudiu isoladamente superou o número de vezes em que o fez em conjunto com outros GP, passando-se precisamente o oposto com Os Verdes, o que faz aliás todo o sentido, no contexto da coligação parlamentar existente entre PCP e Os Verdes (CDU).

4.5.2 Protestos

XV Governo Constitucional

		PSD	PS	CDS-PP	PCP	BE	Os Verdes
Total GP		41%	39%	27%	17%	3%	1%
GP isolado/ total das reacções		19%	35%	6%	10%	1%	0%
Indicador de Isolamento		48%	88%	21%	61%	35%	30%
União	**Esquerda**	0%	100%	0%	100%	100%	100%
	Direita	100%	0%	100%	0%	0%	0%

XVI Governo Constitucional

		PSD	PS	CDS-PP	PCP	BE	Os Verdes
Total GP		54%	32%	32%	12%	2%	1%
GP/ total das reacções		24%	30%	3%	8%	1%	1%
Indicador de Isolamento		45%	92%	8%	67%	50%	50%
União	**Esq**	0%	100%	0%	100%	100%	100%
	Dta	100%	0%	100%	0%	0%	0%

Na categoria "protestos", a coesão da coligação de incidência parlamentar não é tão marcada como na categoria "aplausos". Com efeito, o GP do PSD protesta mais vezes isoladamente (com 19% e 24% relativamente ao total das reacções, no XV e XVI GC, respectivamente), apresentando um índice de isolamento substancialmente superior (48% e 45% em cada um dos governos constitucionais). Como consequência desta "demarcação" do PSD, a percentagem total de protestos do CDS-PP é mais baixa do que na categoria aplausos (27% no XV GC e 32% no XVI GC). Os protestos deste GP isoladamente apresentam valores baixos (6% e 3% em cada um dos governos), tendo no entanto sido mais representativos no governo liderado por Durão Barroso (com 21% de índice de isolamento) do que no de Santana Lopes (com um indicador de isolamento de 8%).

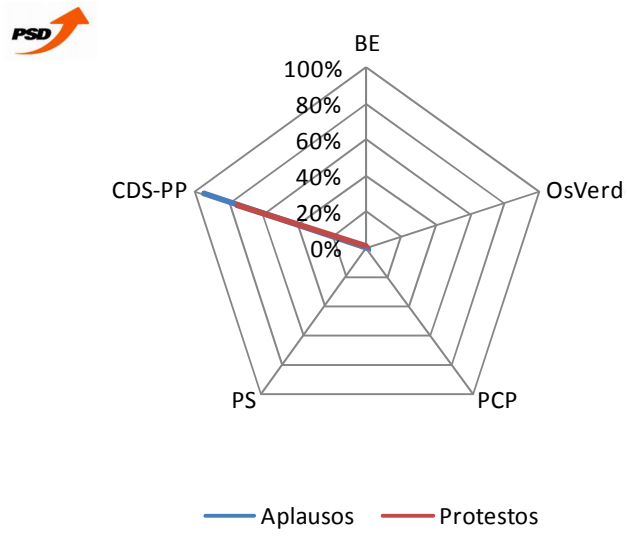
O comportamento do PS é consistente com o analisado nos aplausos, embora com maior participação total deste partido – 39% no XV e 32% no XVI GC – e uma participação individual ligeiramente acima da manifestada nos aplausos (35% e 30%). O PS também parece unir-se mais a outros partidos para protestar, com um indicador de isolamento ligeiramente inferior ao apresentado na categoria aplausos (88% no XV GC e 92% no XVI GC), com associação exclusiva aos GP de esquerda.

Nesta categoria, também o PCP apresenta um comportamento semelhante ao da anterior, com uma percentagem ligeiramente mais elevada no total de protestos (17%) no governo de Durão Barroso do que no de Santana Lopes (12%). É ainda de assinalar o facto de, para protestar, o PCP se aliar mais a outros partidos políticos, com um índice de isolamento inferior (61% e 67% no XV e XVI GC, respectivamente), associando-se sempre a partidos de esquerda.

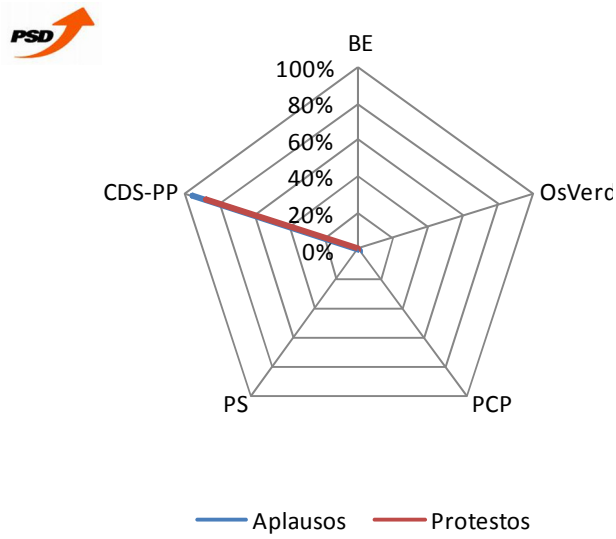
A participação de BE e de Os Verdes é igualmente pouco significativa, sendo de relevar o facto de na categoria protestos Os Verdes assumirem um pouco mais de individualismo (com um índice de 30% no XV GC e de 50% no XVI GC) e, pelo contrário, o BE o perder, com uma prestação muito idêntica à dos Verdes (com 35% no XV GC e 50% no XVI GC). Em qualquer um dos casos, quando se uniram a outros GP para protestar, esta associação foi à esquerda.

Nos gráficos apresentados em seguida é possível visualizar mais facilmente os resultados acima descritos, sendo também bastante notória a homogeneidade de comportamento dos GP relativamente aos dois GC.

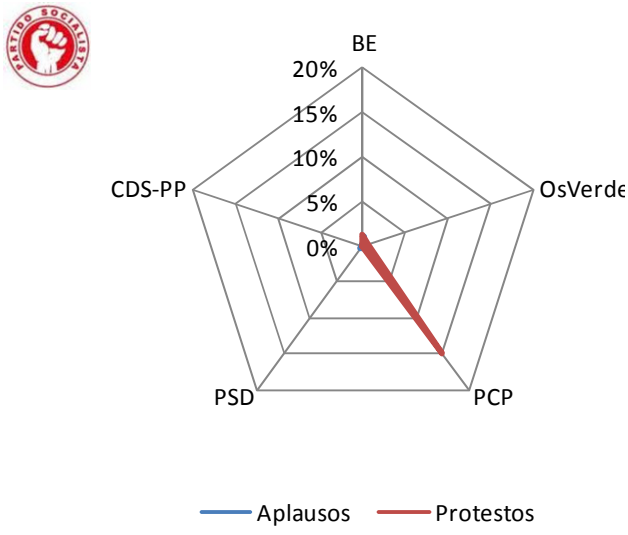
XV Governo Constitucional



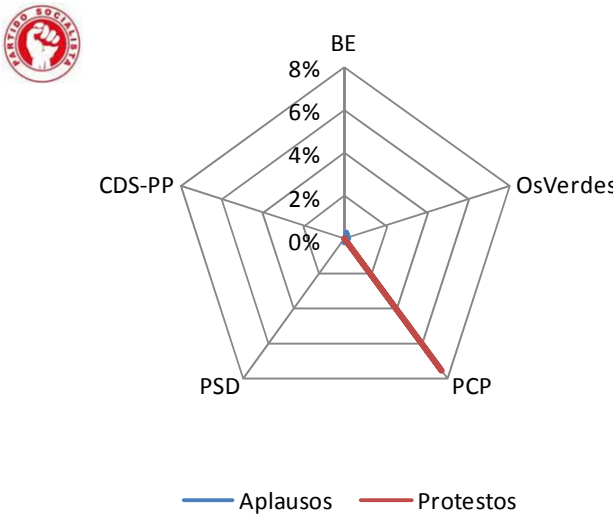
XVI Governo Constitucional



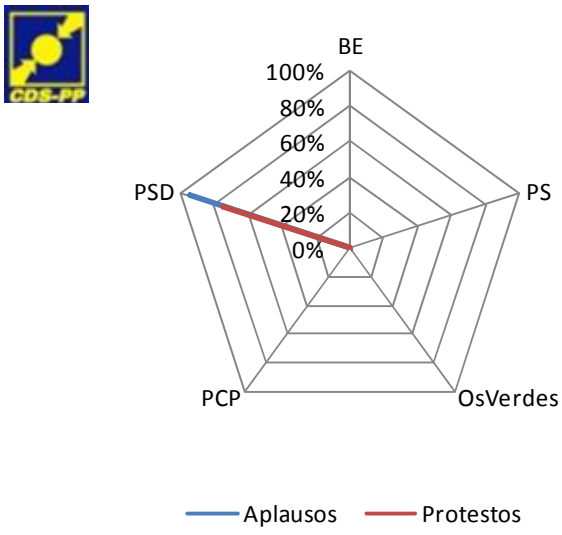
XV Governo Constitucional



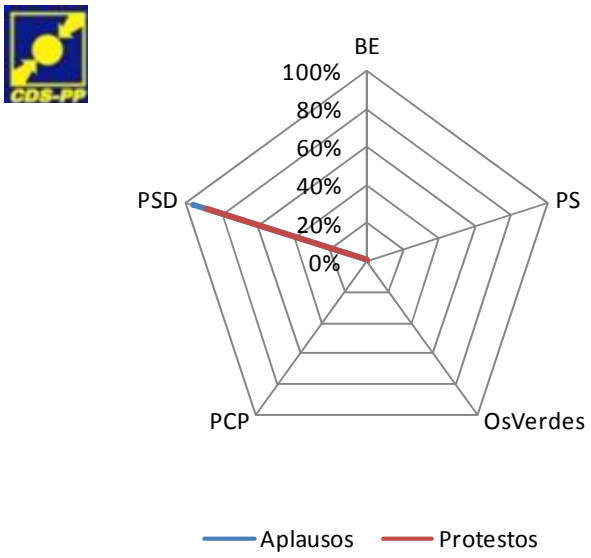
XVI Governo Constitucional



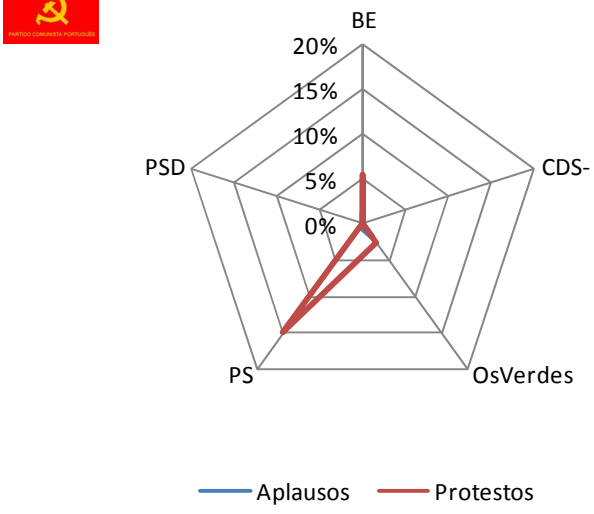
XV Governo Constitucional



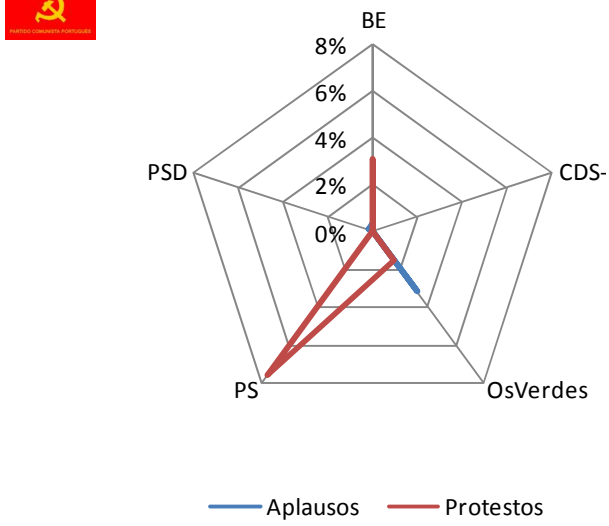
XVI Governo Constitucional



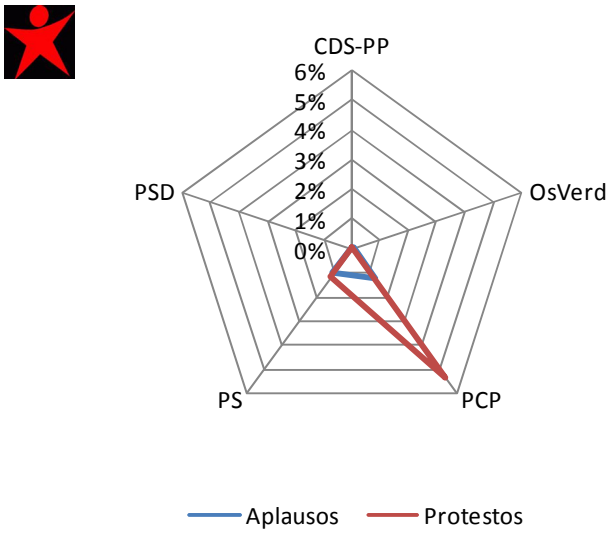
XV Governo Constitucional



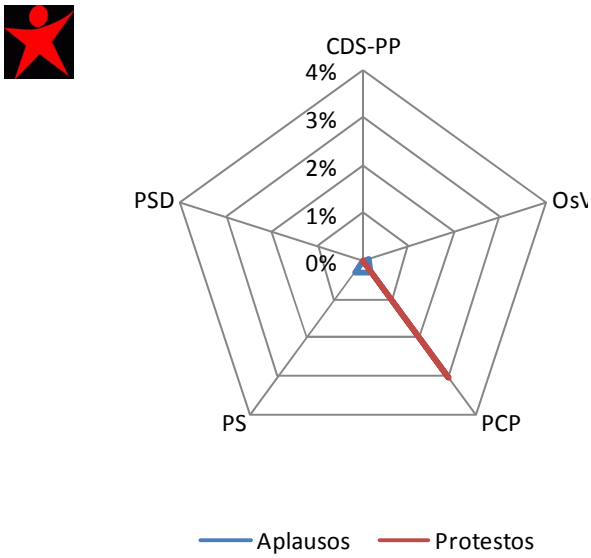
XVI Governo Constitucional



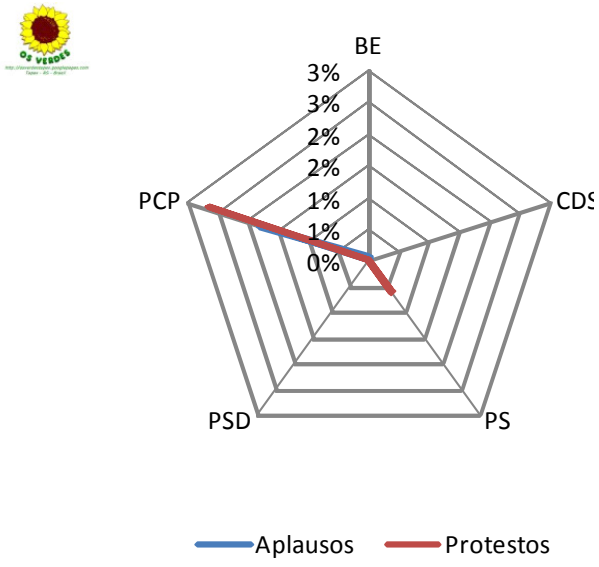
XV Governo Constitucional



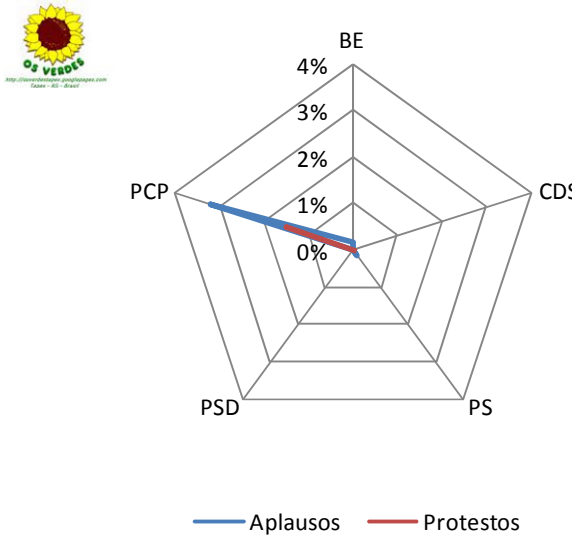
XVI Governo Constitucional



XV Governo Constitucional



XVI Governo Constitucional



4.5.3 Risos

XV Governo Constitucional

		PSD	PS	CDS-PP	PCP	BE	Os Verdes
Total GP		39%	37%	34%	22%	6%	2%
GP isolado/ total das reacções		11%	29%	6%	12%	2%	0%
Indicador de Isolamento		28%	79%	18%	53%	38%	14%
União	**Esquerda**	0%	99%	0%	100%	100%	100%
	Direita	100%	1%	100%	0%	0%	0%

XVI Governo Constitucional

		PSD	PS	CDS-PP	PCP	BE	Os Verdes
Total GP		40%	28%	37%	23%	13%	2%
GP/ total das reacções		8%	22%	5%	11%	5%	0%
Indicador de Isolamento		20%	79%	14%	48%	39%	0%
União	**Esquerda**	0%	100%	0%	100%	100%	100%
	Direita	100%	0%	100%	0%	0%	0%

Na categoria "risos", PSD e CDS-PP revelam o mesmo comportamento coeso manifestado nas categorias anteriores, com elevadas percentagens na totalidade das reacções de cada um dos GP – 39% e 40% para o PSD; 34% e 37% para o CDS-PP, no XV e no XVI Governos Constitucionais, respectivamente. Tal como nos casos já vistos, estas percentagens elevadas correspondem a um grande número de manifestações conjuntas, já que os risos destes GP apresentam valores mais baixos quando vistos isoladamente. Da mesma forma, o PSD manifesta maior grau de autonomia do que o CDS-PP, com índices de isolamento superiores: 28% e 20% para o PSD; 18% e 14% para o CDS-PP. Note-se, não obstante, que o CDS-PP tem mais protagonismo nesta categoria do que nas anteriormente detalhadas (exceptuando a categoria protestos no XV GC).

Já o Partido Socialista revela, nesta categoria, um comportamento similar ao da categoria "protestos". De facto, apresenta um total de risos do GP elevado nos dois GC, embora superior no governo liderado por Durão Barroso, com 37%, e com apenas 28% no governo de Santana Lopes. Apesar do índice de

isolamento ser superior a 50%, o PS apresenta uma maior incidência de reacções conjuntas com outros partidos, quase sempre de esquerda.

O PCP sobe a sua participação nesta categoria (com o dobro da percentagem apresentada na categoria aplausos, por exemplo, com valores superiores aos 20% nos dois GC - 22% e 23%, respectivamente). Já o indicador de isolamento é mais baixo, aproximando-se dos 50% nos dois GC, revelando que este partido tem um comportamento mais "grupal" (com 100% de união com os partidos de esquerda), nos momentos mais bem dispostos dos debates parlamentares.

O BE destaca-se de Os Verdes nesta categoria, sobretudo no governo liderado por Santana Lopes, no qual a percentagem total de risos deste partido mais que duplica a apresentada no governo de Durão Barroso: 13% durante o XVI GC e 6% no XV GC. Nos dois governos o indicador de isolamento deste partido situa-se abaixo dos 50%, mostrando que também o BE dá mais gargalhadas em conjunto com outros GP, nomeadamente os de esquerda. A participação de Os Verdes é, também aqui, pouco relevante e maioritariamente associada à esquerda.

4.5.4 Vozes

XV Governo Constitucional

		PSD	PS	CDS-PP	PCP	BE	Os Verdes
Total GP		38%	25%	29%	20%	5%	2%
GP/ total das reacções		23%	23%	14%	17%	4%	0%
Indicador de Isolamento		62%	93%	50%	83%	69%	27%
União	**Esq**	2%	84%	1%	98%	100%	100%
	Dta	98%	16%	99%	2%	0%	0%

XVI Governo Constitucional

		PSD	PS	CDS-PP	PCP	BE	Os Verdes
Total GP		36%	26%	27%	20%	5%	3%
GP/ total das reacções		24%	25%	15%	16%	4%	1%
Indicador de Isolamento		66%	97%	55%	81%	73%	18%
União	**Esq**	1%	86%	0%	100%	100%	100%
	Dta	99%	14%	100%	0%	0%	0%

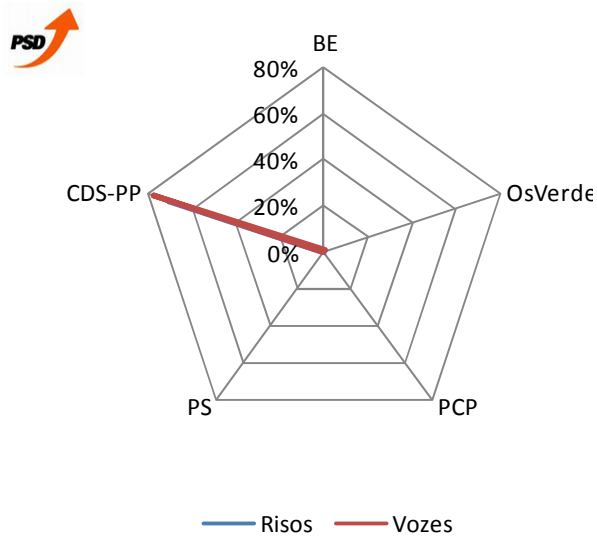
Vejamos, por último, o comportamento dos GP na categoria "vozes". Destaca-se a presença da coligação de incidência parlamentar, com 38% e 36% de valores totais para o PSD e 29% e 27% de valores totais para o CDS-PP. No entanto, quer o PSD, quer o CDS-PP reagem mais vezes sozinhos, apresentando os maiores indicadores de isolamento nos dois governos constitucionais: 62% no XV e 66% no XVI, para o PSD; 50% no XV e 55% no XVI para o CDS-PP. Ainda relativamente a estes dois partidos, cabe salientar que nesta categoria ambos apresentam momentos de coincidência com partidos de esquerda.

O comportamento do PS é, nesta categoria, consistente com o revelado nas restantes: forte presença total (26% no XV GC e 25% no XVI GC), com elevado cariz individualista (96% de indicador de isolamento no XV GC e 99% no XVI GC) e, nos momento de associação a outros GP, com coesão prevalentemente à esquerda.

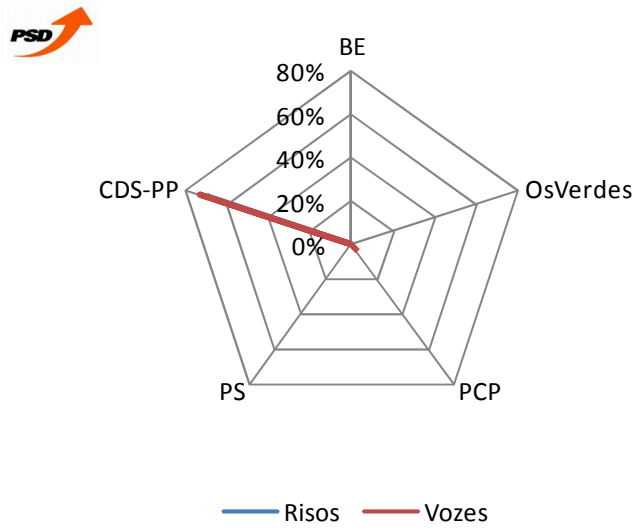
Também relativamente ao PCP, não há grandes desvios comportamentais a assinalar: é um GP com uma boa participação para a pequena representatividade (11% e 10% no XV e XVI GC, respectivamente). Nesta categoria apresenta um índice de isolamento ligeiramente ao das outras categorias em análise (83% e 79% nos governos de Durão Barroso e de Santana Lopes, respectivamente) e união à esquerda.

O BE e Os Verdes têm uma participação pouco significativa, sendo mais uma vez evidente o maior individualismo das reacções do BE quando comparado com Os Verdes, o que mais uma vez se compreende no âmbito da coligação parlamentar existente entre PCP e Os Verdes (CDU).

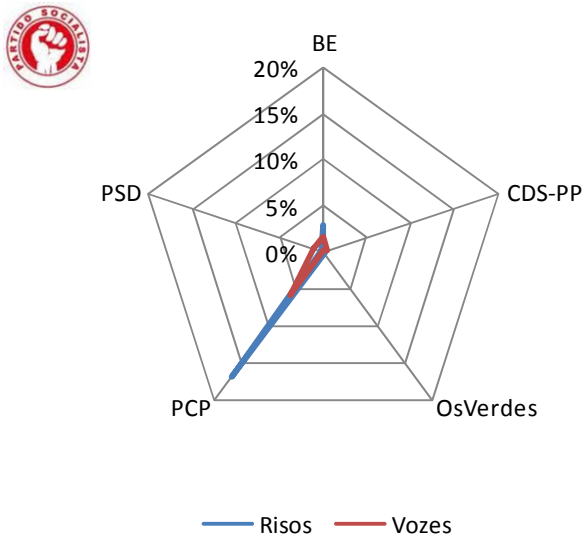
XV Governo Constitucional



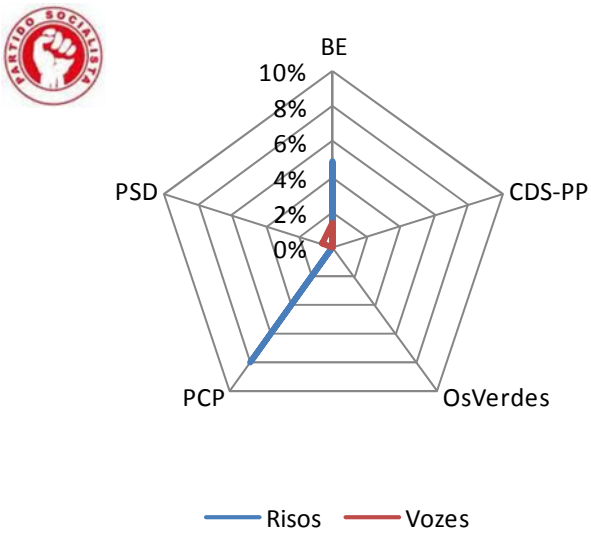
XVI Governo Constitucional



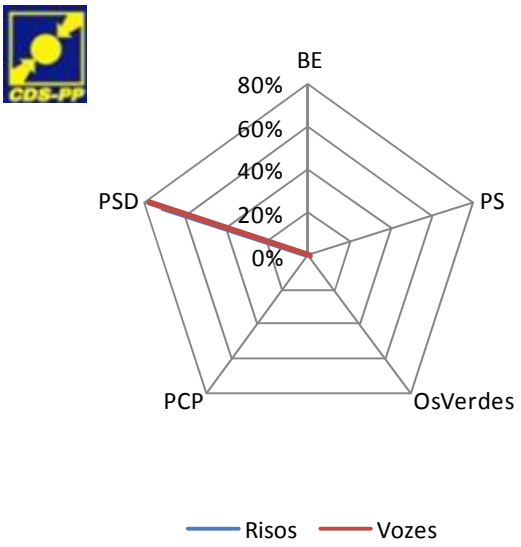
XV Governo Constitucional



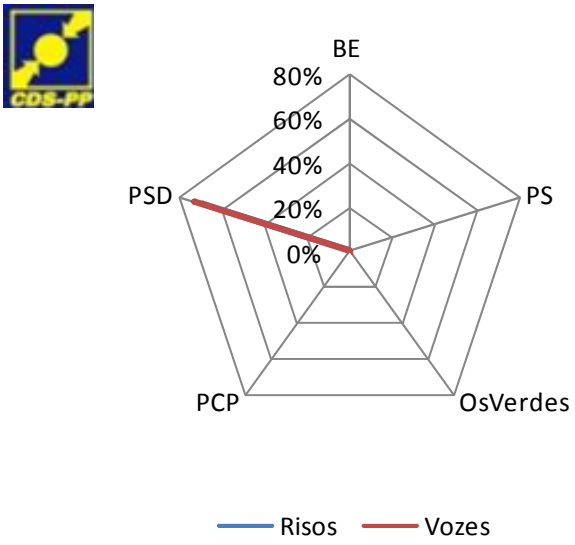
XVI Governo Constitucional



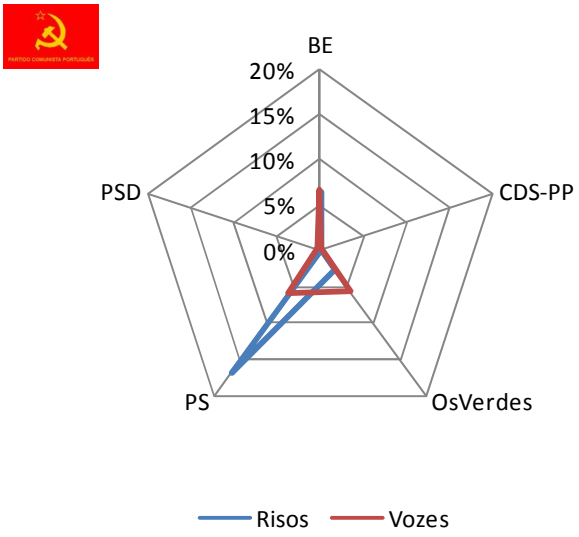
XV Governo Constitucional



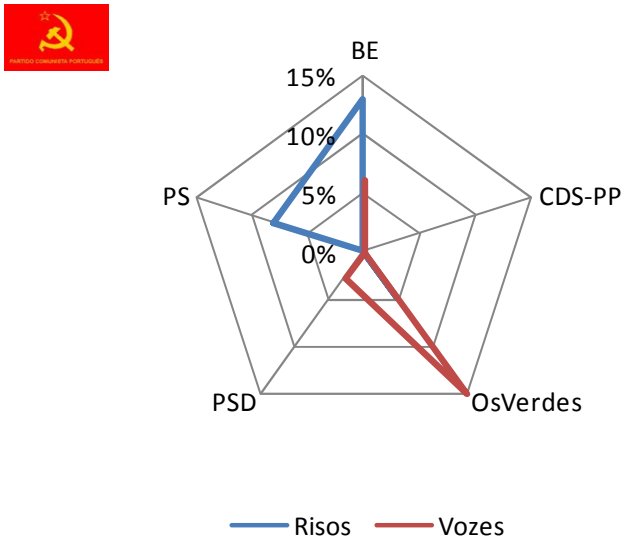
XVI Governo Constitucional



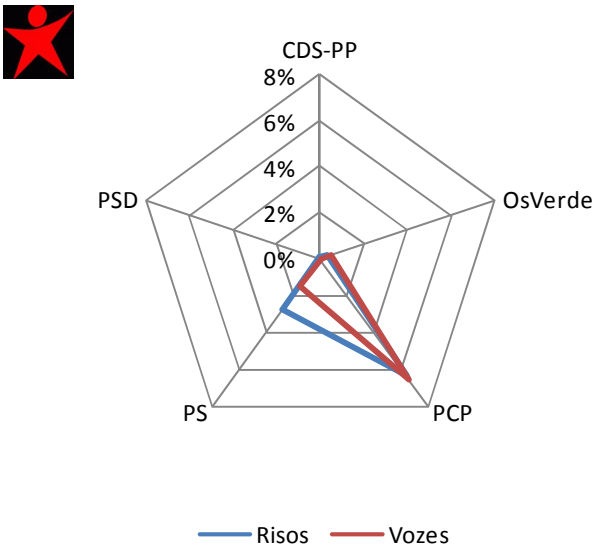
XV Governo Constitucional



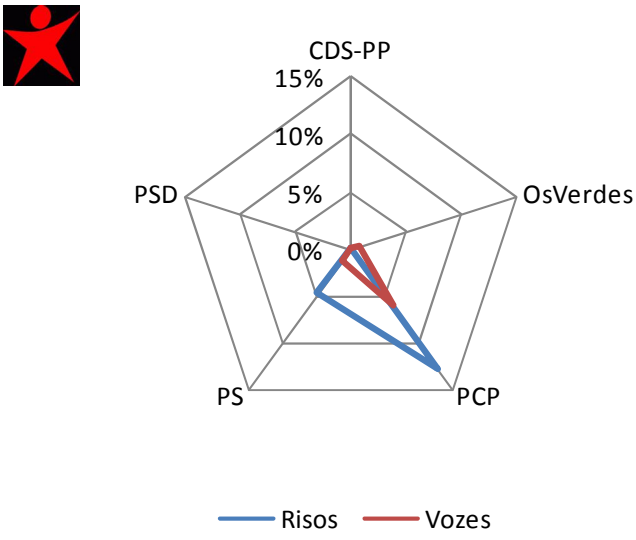
XVI Governo Constitucional



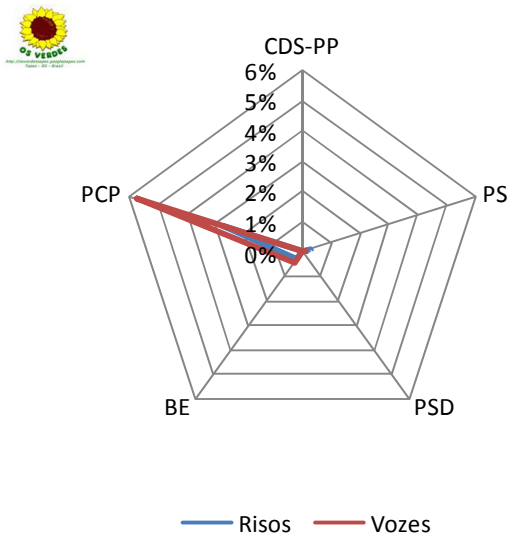
XV Governo Constitucional



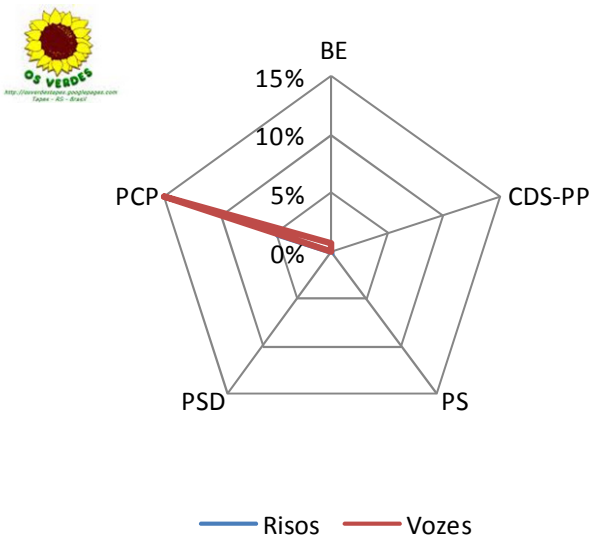
XVI Governo Constitucional



XV Governo Constitucional



XVI Governo Constitucional



À laia de conclusão, com base na análise de resultados previamente apresentada, vejamos de que forma este projecto nos ajudou a dar resposta às questões de investigação apresentadas no ponto 3.4.

(1) Como é que os Grupos Parlamentares se unem nas emoções manifestadas?
Verificou-se com clareza a uniformidade das reacções revelada pelos GP do PSD e do CDS-PP, o que faz, efectivamente, todo o sentido, no contexto da IX Legislatura, em que o acordo entre o PPD/PSD e o CDS-PP liderou o XV e o XVI Governos Constitucionais. Note-se que este fenómeno também é visível, a uma escala menor, na coligação entre PCP e Os Verdes.

(2) Que relação existe entre a coesão na manifestação de emoções de dois GP e as suas orientações políticas (por exemplo, os partidos da esquerda e da direita aplaudem sempre em conjunto)?
Efectivamente, na legislatura em análise, constatámos que os partidos de esquerda tendem a unir-se com os de esquerda, e os de direita com os de direita, verificando-se a inexistência quase absoluta de concordância entre os extremos (BE & CDS-PP; BE & PSD; Os Verdes & CDS-PP; Os Verdes & PSD; PCP & CDS-PP; PCP & PSD). Não obstante, seria interessante confrontar esta legislatura com uma legislatura governada pelo PS, por exemplo, pois atrevemo-nos a considerar que esta união entre esquerda e direita não seria tão linear.

(3) Como é que as emoções transmitidas se articulam com o poder?
As emoções transmitidas articulam-se com a ordem instituída, quer seja na união dos dois partidos que representam a coligação com incidência parlamentar, quer seja na consonância da oposição nas reacções conjuntas.

(4) Concretamente em relação à legislatura em análise (a IX), que diferenças existem entre o governo liderado por Durão Barroso (XV) e o dirigido por Santana Lopes (XVI)?
Muito provavelmente por se tratar de uma comparação entre dois governos constituídos exactamente pelas mesmas forças políticas no poder, não se verificam diferenças de comportamento assinaláveis dos GP no XV e no XVI Governos Constitucionais.

(5) Qual o grau de isolamento das reacções dos GP quando reagem emotivamente?
Com esta questão de investigação procurámos aferir o "carisma" ou a "atitude" dos GP. É muito interessante verificar que os dois partidos que assumiam o poder apresentam baixos valores neste indicador em quase todas as categorias. Ou seja, à coligação política PSD / CDS-PP corresponde uma elevada coesão de emoções e de atitudes. Por outro lado, é de relevar que ao principal partido da oposição, o PS, equivale grande individualismo e carácter na manifestação de emoções, sendo que este indicador também apresenta valores interessantes em partidos de menor dimensão, como é o

caso do PCP ou mesmo do BE. Nos Verdes, em relação ao PCP, assiste-se a um fenómeno idêntico ao do CDS-PP com o PSD, pois o facto de Os Verdes integrarem uma coligação com o PCP leva-os a reagir quase sempre em conjunto com este partido.

(6) Há relação entre o número de deputados representados e a capacidade de demonstrar emoções?

Tal como já tivemos oportunidade de referir, o programa utilizado não nos permite responder com segurança a esta questão, pois não permite fazer a correlação entre as reacções dos GP e os deputados efectivamente presentes em cada sessão. Não obstante, pode-se constatar que há uma relação entre a maior representatividade parlamentar e a capacidade de mostrar emoções (vejam-se os casos do PS e do PSD, com elevada representatividade e grande capacidade de demonstrar emoções), muito embora um partido com baixa representação parlamentar, como o PCP, apresente sempre valores isolados superiores aos do CDS-PP, com representação parlamentar ligeiramente superior na legislatura e governos em questão.

5 Capítulo 5

5.1 Conclusão

Este trabalho de projecto, desenvolvido no âmbito do Mestrado em Estatística e Gestão de Informação, permitiu-nos desenvolver capacidades de investigação e trabalho numa área emergente e apelativa para diversos sectores da actividade económica, em contínua inovação e com oportunidades interessantes para o futuro.

Por um lado, tornou possível a utilização de um programa de processamento automático de dados não estruturados, tendo este aspecto sido particularmente estimulante, uma vez que o SAS nos deu a hipótese de utilizar, "em primeira mão", um *software* totalmente inexplorado pela delegação portuguesa da empresa.

Não obstante, não podemos escamotar o facto de o modelo aqui desenvolvido apresentar algumas limitações na análise, determinadas pelas contingências do *software* utilizado:

(1) Impossibilidade de relacionar as reacções dos deputados com as intervenções que as precederam

Uma das maiores debilidades da análise é o facto de, com este programa de categorização textual, não ser possível estabelecer uma relação entre as reacções dos deputados e as intervenções que as precederam. De facto, os resultados seriam mais ricos se pudéssemos avaliar em reacção a quem aplaudem ou protestam os partidos, ou de que situações se riem ou vozeiam. Pensamos que esta restrição poderá ser ultrapassada com a utilização de um *software* de *text mining*, que, como vimos, não foi possível utilizar no contexto do presente trabalho.

(2) Impossibilidade de contabilizar todas as ocorrências de uma dada expressão

Outra condicionante da nossa análise prende-se com os resultados propriamente ditos. Os números indicados pelo programa não correspondem exactamente às ocorrências de uma dada expressão num documento *html*, mas ao número de documentos em que estas aparecem pelo menos uma vez. Dada a dimensão reduzida dos documentos *html* utilizados como dados de *input* (que, como já referimos, correspondem a uma visualização da página da internet), esta restrição não inviabiliza os resultados.

Tentámos ultrapassar esta debilidade recorrendo ao *Teragram Concept Extractor*, que nos permitiria obter a contagem de todas as ocorrências de uma dada expressão num documento de *input*. No entanto, apesar de esta ferramenta ter sido explorada (o intuito seria cruzar resultados, após análise dos mesmos documentos de *input,* com as duas ferramentas do Teragram TK240, i.e.,

o *Teragram Concept Extractor* e o *Teragram Categorizer*), o *software* não suportou o processamento da totalidade dos ficheiros *html* da IX Legislatura e não produziu resultados.

(3) Inexistência de listagens dos documentos classificados em cada categoria

A terceira limitação que nos é imposta pelo *software* é o facto de não ser possível identificar de forma intuitiva e fácil os documentos classificados dentro de cada subcategoria. Com efeito, embora os documentos atribuídos em cada subcategoria apareçam listados na janela *testing*, podendo ser individualmente seleccionados para verificação do seu contexto, não é possível descarregar para um ficheiro à parte uma listagem dos documentos classificados em cada categoria.

Em suma, o programa é bastante eficaz na categorização automática dos documentos e de fácil utilização na criação de categorias. No entanto, quando se procura efectuar um estudo que ultrapasse a categorização textual, o programa não apresenta recursos suficientes para uma análise mais rica e profícua dos dados.

Sendo a categorização textual uma área em actualização constante, uma das lacunas que também tentámos, modestamente, ajudar a colmatar, foi, precisamente, a escassa bibliografia existente em português sobre o assunto. No entanto, o tempo disponível para realizar este trabalho de projecto não nos permitiu ir além de um breve enquadramento teórico (o mais consistente possível no período em causa), com clarificação de alguns conceitos e processos implicados numa investigação em processamento automático de informação textual. É um modesto contributo para um trabalho que fica por fazer – a elaboração de um "estado da arte" sobre categorização textual, processamento semi-automático e automático de dados não estruturados e *text mining*, pois estas são áreas que merecem uma investigação mais aprofundada em português europeu.

A falta de credibilidade da política, o desinteresse crescente da opinião pública sobre estes assuntos e um certo desconhecimento do que se passa na AR levou-nos a escolher como objecto de análise os debates parlamentares, e a analisar, nestes, os elementos habitualmente não observados de forma sistemática – as emoções e as reacções dos GP em AR. Do nosso ponto de vista, os resultados obtidos revelaram ser extremamente interessantes, porque transparecem exactamente a orientação política vigente na AR. Lamentamos apenas não termos tido a oportunidade de aprofundar o nosso estudo de forma longitudinal por falta de acesso aos dados, o que poderá ser facilmente resolvido.

Na sequência deste trabalho de projecto, abrem-se assim linhas de investigação para o futuro. Se tivermos como ponto de partida o protótipo criado (com todas as contingências a que este foi sujeito no curtíssimo prazo disponível para a sua realização), haverá pelo menos dois caminhos a seguir.

Por um lado, pode-se aplicar este protótipo a todas as legislaturas, desde 1974. Seria possível verificar, dessa forma, que alterações se verificam nas reacções dos GP relativamente ao partido que se encontra no poder. Por exemplo, seria pertinente verificar se as reacções em bloco do PSD e CDS-PP se verificam mesmo quando a coligação destes dois partidos não assume a liderança do Governo, e se, num contexto recente, em que o PS governava com maioria absoluta, existiria uma igual coesão entre este partido e o PCP. Este estudo longitudinal poderia ser feito com o *Teragram TK 240*.

Outra hipótese - que dependeria, no entanto, da utilização de um *software* de *text mining* eficaz para o português europeu - seria analisar os DAR (sem dúvida carregados de informação de interesse público por analisar e explorar), elaborando uma pesquisa exploratória, com vista a avaliar quais foram os temas mais frequentemente debatidos na AR no período decorrente entre 1976 e 2005. Poder-se-ia, numa segunda fase, e com recurso à metodologia de análise de *clusters*, procurar relações entre os debates parlamentares do ponto de vista dos temas abordados, desde a primeira legislatura (1976-1980) até à actualidade.

Em suma, acreditamos que este é um projecto inovador, não só na utilização do software Teragram TK240; mas também pela abordagem de linhas do saber actuais, do ponto de vista da revisão da literatura; e pela exploração de dados pouco estudados, com uma abordagem inédita. Esperamos que seja merecedor de interesse e cremos que não pecará, certamente, por falta de originalidade.

6 Referências bibliográficas

Advanced Approaches in Analyzing Unstructured Data. New York, Cambridge University Press.

ANDROUTSOPOULOS, I., J. KOUTSIAS & K. V. CHANDRINOS (2000). An experimental comparison of naive Bayesian and keyword based anti-spam filtering with personal e-mail messages. In *Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval*, Athens, Greece, 2000: 160-167.

APTE, C., F. DAMERAU & S. WEISS (1994). Automated Learning of Decision Rules for Text Categorization. *ACM Transactions on Information Systems*, 12, 3: 233-251.

APTE, C., F. DAMERAU & S. WEISS (1994). Towards language independent automated learning of text categorization models. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, USA, 1994.

APTE, C., F. DAMERAU & S. WEISS (1998). Text mining with decision rules and decision trees. *Proceedings of the Conference on Automated Learning and Discovery*. Pittsburg, 1998.

BAÇÃO, F. L. (2007). *Data Mining.* Lisbon, Portugal, ISEGI – UNL, 2007.

BAEZA-YATES, R. & B. RIBEIRO-NETO (1999). *Modern Information Retrieval.* New York, ACM Press.

BAKER, L. DOUGLAS & A. K. MCCALLUM (1998). Distributional clustering of words for text categorization. In *Proceedings of the 21th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval,* Melbourne, Australia: 96-103.

BELKIN, N. J. & W. B. CROFT (1992). Information Filtering and Information Retrieval: two sides of the same coin? *Commun. ACM 35*, 12, 29-38.

BORKO, H. & M. BERNICK (1963). Automatic Document Classification. *J. Assoc. Comput. Mach. 10,* 2, 151-161.

CAVNAR, W. B. & J. M. TRENKLE (1994). N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas: 161-175.

CHEN, H. (2001). *Knowledge Management Systems. A Text Mining perspective.* Arizona, Knowledge Computing Corporation, 2001.

COHEN, A. M. & W. R. HERSH (2005). A survey of current work in biomedical text mining. *BRIEFINGS IN BIOINFORMATICS* **6**(1): 57-61.

COHEN, WILLIAM W. (1995). Text Categorization and Relational Learning. *The Twelfth Internacional Conference on Machine Learning*. Morgan Kaufmann.

COHEN, WILLIAM W. & YORAM SINGER (1996). Context-sensitive learning methods for text categorization. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval:* 307-315.

CORREIA, A. M. R. (2007). *Materiais de apoio para as UC Metodologias de Investigação do Mestrado em Estatística e Gestão da Informação Lisboa*, ISEGI - UNL.

CRESWELL, J. W. (2003). *Research Design. Qualitative, Quantitative and Mixed Methods Approaches.* California, Sage Publications.

DELEN, D. & M. D. CROSSLAND (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems With Applications* **34**: 1707-1720.

DRUCKER, H. , V. VAPNIK & D. WU (1999). Automatic Text Categorization and its applications to text retrieval. *IEEE Trans. Neural Netw., 10, 5*: 1048-1054.

EPPLER, M. J. & J. MENGIS (2004). The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines. *The Information Society **20**(5): 325 - 344.*

FAYYAD, U. & R. UTHURUSAMY (2002). Evolving Data Mining into Solutions for Insights. *Communications of the ACM* **5**(8): 28-31.

FELDMAN, R. & I. DAGAN (1995). Knowledge discovery in textual databases (KDT). *Knowledge Discovery and Data Mining*.

FELDMAN, R. & J. SANGER (2007). *The Text Mining Handbook*.

FIELD, B. (1975). Towards automatic indexing: automatic assignment of controlled-language indexing and classification from free indexing. *J. C. Document.* 31, 4, 246-265.

FORSYTH, R. S. (1999). New directions in text categorization. In *Causal Models and Intelligent Data Management,* A. Gammerman, Heidelberg, Germany, Springer, 151-185.

FUHR N., S. HARTMANNA, G. LUSTIG, M. SCHWANTNER & K. TZERAS (1991). Air/x – a rule-based multistage indexing systems for large subject fields. *Proceedings of RIAO'91*.

GABRILOVICH, E. & S. MARKOVITCH (2004). Text Categorization with Many Redundant Features: Using Aggressive Feature Selection to Make SVMs Competitive with C4.5. *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004.

GRAY, W. A. & A. J. HARLEY (1971). Computer Assited Indexing. *Inform. Storage Retrieval 7,* 4: 167- 174.

HAYES, P. J., P. M. ANDERSEN, I. B. NIRENBURG & L. M. SCHMANDT (1990). Tcs: a shell for content-based text categorization. In *Proceedings of CAIA-90, 6th IEEE Conference on Artificial Intelligence Applications*, Santa Barbara, California: 320-326.

HEAPS, H. (1973). A Theory of relevance for automated text classification. *Inform. Control 22,* 3: 268-278.

HEARST, M. (1999). *Untangling Text Data Mining***:** 3-10.

HUSEMAN, R. C. AND J. P. GOODMAN (1999). *Leading with Knowledge. The Nature of Competition in the 21st Century*, Sage Publications.

JOACHIMS, THORSTEN (1998). Text Categorization with Support Vector Machines: Learning with many relevant features. *European Conference on Machine Learning (ECML)*.

KESSLER, B., G. NUNBERG & H. SCHÜTZE (1997). Automatic detection of text genre. In *Proceedings of ACL-97, 35th Annual Meeting of the Association of Computational Linguistics*, Madrid: 32-38.

KLOPTCHENKO, A. (2003). *Text Mining Based on the Prototype Matching Method*. Åbo, Faculty of Economics and Social Sciences, Åbo Akademi University.

KOLLER, D. & M. SAHAMI (1997). Hierarchically classifying documents using very few words. *Fourteenth International Conference on Machine Learning (ICML)*: 170-178.

LAM W. & C.Y. HO. (1998). Using a generalized instance set for automatic text categorization. *Proceedings of the 21th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval*: 81-89.

LARKEY, L. S. (1999). A patent search and classification system. In *Proceedings of DL-99, 4th ACM Conference on Digital Libraries,* Berkeley: 179-187.

LEWIS, D. D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, Copenhagen, Denmark: 37-50.

LEWIS, D. D. & M. RINGUETTE (1994). Comparison of two learning algorithms for text categorization. *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*.

LEWIS, D. D., ROBERT E. SCHAPIRE, JAMES P. CALLAN & RON PAPKA (1996). Training Algorithms for linear text classifiers. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*: 298-306.

LEWIS, D. D., Y. YANG, T. ROSE & FAN LI (2004). RCV1: A New Benchmark Collection for text categorization research, *Journal of Machine Learning Research* 5: 361-397.

MARON, M. (1961). Automatic indexing: an experimental inquiry, *J. Assoc. Comput. Mach. 8,* 3, 404-417.

MASAND, B., G. LINOFF & D. WALTZ (1992). Classifying news stories using memory based reasoning. *15th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval*: 59-64.

MCCALLUM, A. & K. NIGAM (1998). A comparison of event models for naive bayes text classification. *AAAI-98 Workshop on Learning for Text Categorization.*

MCCALLUM, A., R. ROSENFELD, T. MITCHELL & A. NG (1998). Improving Text Classification by Shrinkage in a Hierarchy of Classes. In *Proceedings of the Fifteenth International Conference on Machine Learning,* San Francisco: 359-367

MCKNIGHT, W. (2005). Text Data Mining in Business Intelligence. *DM Review*: 21-22.

MERCKL, D. (1998). Text classification with self-organization maps: Some lessons learned. *Neuro-computing* 21, 1/3, 61-77.

MILLER, T. W. (2005). *Data and text mining: a business applications approach*. New Jersey, Pearson/Prentice Hall.

MOULINIER, I., G. RASKINIS & J. GANASCIA (1996). Text categorization. A symbolic approach. *Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval.*

MOULINIER, I. (1997). Is learning bias an issue on the text categorization problem?. *Technical Report, LAFORIA-LIP6, Université Paris VI*.

MYERS, K., M. KEARNS, S. SINGH & M. A. WALKER (2000). A boosting approach to topic spotting on subdialogues. In *Proceedings of ICML-00, 17th International Conference on Machine Learning*, Stanford, 2000: 655-662.

NG., H.T., W.B. GOH AND K.L. LOW (1997). Feature Selection, perceptron learning, and a usability case study for text categorization. *20th Ann Int ACM SIGIR Conference on Research and Development of Information Retrieval*: 67-73.

PASSARIN, D. (2005). *Text Mining no Aperfeiçoamento de Consultas e Definição de Contextos de uma Central de Notícias Baseada em RSS*. Palmas, Centro Universitário Luterano de Palmas.

SABLE, C. L. & V. HATZIVASSILOGLOU (2000). Text based approaches for non-topical image categorization. *Internat. J. Dig. Libr. 3,* 3: 261-275.

SANTOS, M. I. G. D. (2004). *Uma aplicação da* Competitive Intelligence *em contexto organizacional. Identificação das Necessidades de Informação de um Parque de Ciência e Tecnologia. O Caso do Madan Parque,* Universidade Nova de Lisboa.

SCHAPIRE, R. E. & Y. SINGER (2000). BoosTexter: a boosting-based system for text categorization. *Mach. Learn. 39,* 2/3: 135-168.

SEBASTIANI, FABRIZIO (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, vol. 34, No. 1: 1-47.

TZERAS K., & S. HARTMAN (1993). Automatic Indexing based on bayesian inference networks. *Proc 16th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval*: 22-34.

WIENER E., J.O. PEDERSEN & A.S. WEIGEND (1995). A neural network approach to topic spotting. *Proceeedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval*.

WENG, S.-S. AND Y.-J. LIN (2003). A study on searching for similar documents based on multiple concepts and distribution of concepts. *Expert Systems With Applications* **25**(3): 355-368.

WENG, S.-S. AND C.-K. LIU (2004). Using text classification and multiple concepts to answer e-mails. *Expert Systems With Applications* **26**(4): 529-543.

YANG, H.-C. AND C.-H. LEE (2005). A text mining approach for automatic construction of hypertexts. *Expert Systems With Applications* **29**(4): 723-734.

YANG, Y. (1994). Expert network: Effective and Efficient Learning from human decisions in text categorization and retrieval. *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval,* Dublin, Ireland**:** 13-22.

YANG, Y. & C.G. CHUTE (1994). An example based mapping method for text categorization and retrieval. *ACM Transaction on Information Systems*: 252-277.

YANG, Y. & J. P. PEDERSON (1997). A Comparative Study on Feature Selection in Text Categorization**.** *Proceedings of Fourteenth International Conferences on Machine Learning*: 412-420.

YANG, Y. (1999). An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval*, The Netherlands: 69-90.

YANG, Y. & XIN LIU (1999). A re-examination of text categorization methods. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval,* Berkeley: 42 - 49

YIN, R. K. (2003). *Case Study Research. Design and Methods*. California, Sage Publications.

ZANASI, A., ED. (2005). *Text Mining and its Applications To Intelligence, CRM and Knowledge Management*. Advances in Management Information. Boston, WIT Press.

ZHANG, Y. AND J. R. JIAO (2007). An associative classification-based recommendation system for personalization in B2C e-commerce applications. *Expert Systems With Applications* **33**(2): 357-367.

7 ANEXOS

7.1 Mapa do Sítio da Assembleia da República

Mapa do Sítio da Assembleia da República

Retirado de http://www.parlamento.pt/Paginas/MapaSite.aspx

Página Inicial
Presidente
Acessibilidade
Administrador
Contactos
Correio
Correio AHP
Correio BIB
Correio CICRP
Correio DAC
Correio DAPAT
Correio DAPLEN
Correio DE
Correio DGF
Correio DILP
Correio DP
Correio DRAA
Correio DRHA
Correio DRI
Correio DSAF
Correio DSATS
Correio DSDIC
Correio GARIP
Correio Parlamento Jovens
Ficha Técnica
Glossário
Iniciativas Discussão Pública
Mapa do Sítio
Orgãos exteriores com representação da AR
Outras Ligações
Parlamentos do Mundo
Proximos Agendamentos
Reunião da Troika do V Fórum Parlamentar Iberoamericano
Trabalhos do Dia
Últimas Iniciativas Entradas
Últimos Textos Aprovados

Actividade Parlamentar e Processo Legislativo
Actividades Parlamentares
Diplomas Aprovados
Iniciativas Legislativas
Perguntas ao Governo e Requerimentos
Petições
Relatórios/Estatísticas
Reuniões Plenárias
Arquivo e Documentação
Arquivo Audiovisual
Arquivo de Destaques
Biblioteca do Parlamento
Livros on-line
Comissões Parlamentares
10ª Saúde
11ª Trabalho, Segurança Social e Administração Pública
12ª Ética, Sociedade e Cultura
1ª Assuntos Constitucionais, Direitos, Liberdades e Garantias
2ª Negócios Estrangeiros e Comunidades Portuguesas
3ª Defesa Nacional
4ª Assuntos Europeus
5ª Orçamento e Finanças
6ª Assuntos Económicos, Inovação e Desenvolvimento Regional
7ª Poder Local, Ambiente e Ordenamento do Território
8ª Educação e Ciência
9ª Obras Públicas, Transportes e Comunicações
Acompanhamento das Questões Energéticas
Acompanhamento e Avaliação da Política Nacional de Defesa da Floresta contra Incêndios
Inquérito sobre a Situação que Levou à Nacionalização do BPN e sobre a Supervisão Bancária Inerente
Portal das Comissões Parlamentares
Deputados e Grupos Parlamentares
Blogs
Estatuto dos Deputados
Comissão Permanente
Conferência de Líderes
Deputados
Grupos Parlamentares
Mesa da Assembleia
Páginas Pessoais
Presenças e Faltas dos Deputados às Reuniões Plenárias
Presidentes dos Grupos Parlamentares
Resultados Eleitorais
Diário da Assembleia da República
DAR I Série
DAR II Série

Separatas
Dossiers Temáticos
Fiscalização Política
Apreciação de Decretos-Lei
Comissões de Inquérito
Conta Geral do Estado
Inquéritos Parlamentares
Interpelações
Moções
Perguntas ao Governo
Petições
Programa do Governo
Relatórios de Entidades Externas
Relatórios de Segurança Interna
Requerimentos
Gestão do Parlamento
Balanço Social
Conselho de Administração
Contratação Pública
Orçamento e Conta de Gerência
Recrutamento de Pessoal
Secretário-Geral
Serviços da Assembleia da República
Intervenções e Debates
Debates Parlamentares
Intervenções em Plenário
Legislação
Constituição da República Portuguesa
Direito de Petição
Estatuto do Direito de Oposição
Estatuto dos Deputados
Lei da Iniciativa Legislativa dos Cidadãos
Lei das Precedências do Protocolo do Estado Português
Lei de Acompanhamento, Apreciação e Pronúncia pela Assembleia da República no Âmbito do Processo de Construção da União Europeia
Lei de Organização e Funcionamento dos Serviços da Assembleia da República
Lei do Financiamento dos Partidos Políticos e das Campanhas Eleitorais
Lei dos Partidos Políticos
Lei Eleitoral da Assembleia da República
Lei Orgânica do Regime do Referendo
Regime Jurídico de Incompatibilidades e Impedimentos
Regime Jurídico dos Inquéritos Parlamentares
Regimento da Assembleia da República
Livraria Parlamentar
Orçamento do Estado e Contas Públicas
Conta Geral do Estado

Grandes Opções do Plano
Orçamento do Estado
Programa de Estabilidade e Crescimento

Parlamento

Apontamentos Históricos
Competência
Estatuto e Eleição
Organização e Funcionamento
Processo Legislativo Comum

Relações Internacionais

Actividade do Presidente
Boletim
Cooperação Interparlamentar
Delegações Permanentes
Deslocações
Grupos Parlamentares de Amizade
O Parlamento e a União Europeia
Visitas Oficiais

Revisões Constitucionais

Revisão Constitucional de 2005
Revisões Constitucionais anteriores

7.2 Evolução do sítio da AR

O sítio da AR (http://www.parlamento.pt/Paginas/default.aspx) foi alvo de diversas alterações e melhoramentos durante o nosso trabalho. No âmbito do presente projecto, interessavam-nos, em particular, as secções a partir das quais fosse possível extrair informação, directa ou indirectamente, sobre o objecto da nossa análise – os debates parlamentares. De modo a desenvolver um projecto que acrescentasse valor ao serviço já disponibilizado, foram analisadas com maior atenção as secções "Intervenções e Debates" (que se subdivide em Intervenções em Plenário e Debates Parlamentares) e "Deputados e Grupos Parlamentares" (que tem as subsecções: 1) Deputados, 2) Grupos Parlamentares, 3) Mesa da Assembleia, 4) Conferência de Líderes, 5) Presidentes dos Grupos Parlamentares, 6) Páginas Pessoais, 7) Blogs, 8) Resultados Eleitorais, 9) Estatutos dos Deputados, 10) Presenças e Faltas dos Deputados às Reuniões Plenárias).

É, em seguida, descrito o conteúdo do sítio relativamente a estes pontos em particular:

7.2.1 Intervenções e debates

De acordo com a informação disponibilizada pelo sítio, na secção "Intervenções e Debates", estão disponíveis as intervenções dos Deputados de cada GP e do Governo desde a VI Legislatura (Outubro de 1991), feitas no âmbito do processo legislativo e da actividade parlamentar, entre as quais se incluem a discussão de iniciativas legislativas (projectos de revisão constitucional, projectos e propostas de lei, projectos e propostas de resolução e de referendo e projectos de deliberação), petições dos cidadãos, declarações políticas, perguntas ao Governo e outras intervenções produzidas no decurso dos debates ocorridos em plenário. Os resultados das pesquisas realizadas neste sítio interligam-se com a informação contida na base de dados de Debates Parlamentares, que contém os textos integrais de todas as intervenções feitas em plenário desde a Assembleia Constituinte de 1821 até à actualidade.

7.2.2 Intervenções em Plenário

No campo “Intervenções em Plenário”, a busca da intervenção pode ser feita por legislatura, sessão legislativa, assunto, data de intervenção, GP e orador (tal como é visível nas ilustrações 2 a 4).

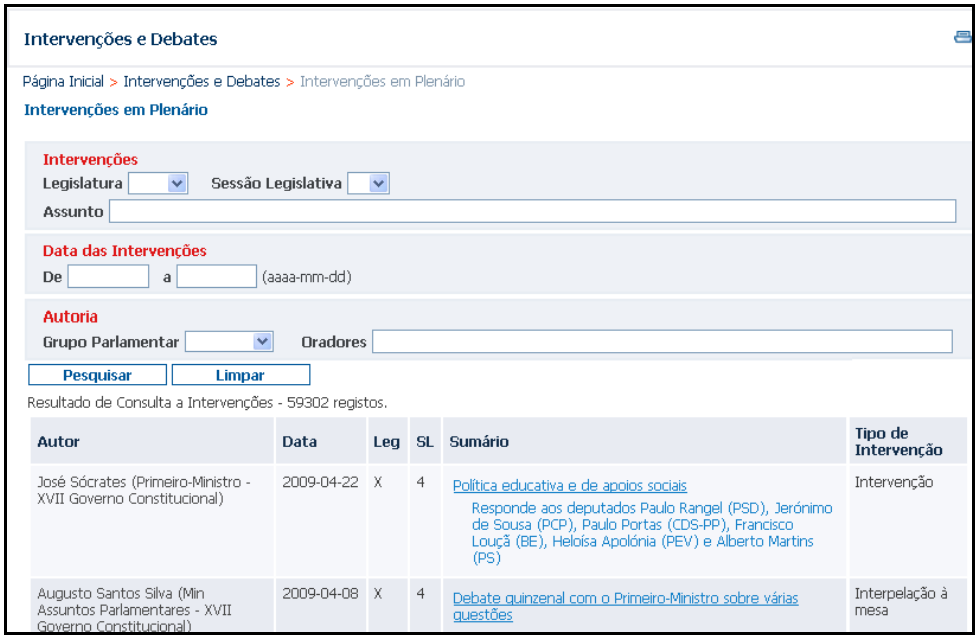


ILUSTRAÇÃO 5 - PÁGINA DA AR, SECÇÃO INTERVENÇÕES E DEBATES, INTERVENÇÕES EM PLENÁRIO

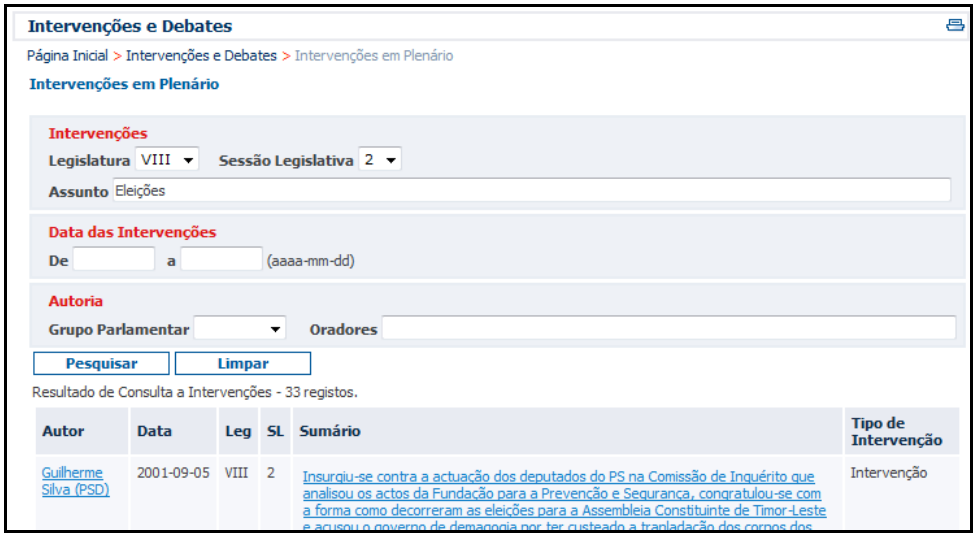


ILUSTRAÇÃO 6 - PÁGINA DA AR, SECÇÃO INTERVENÇÕES E DEBATES, INTERVENÇÕES EM PLENÁRIO, ONDE É VISÍVEL O TIPO DE PESQUISA QUE SE PODE REALIZAR: POR LEGISLATURA, SESSÃO LEGISLATIVA, ASSUNTO, DATA DE INTERVENÇÃO, GP E ORADOR.

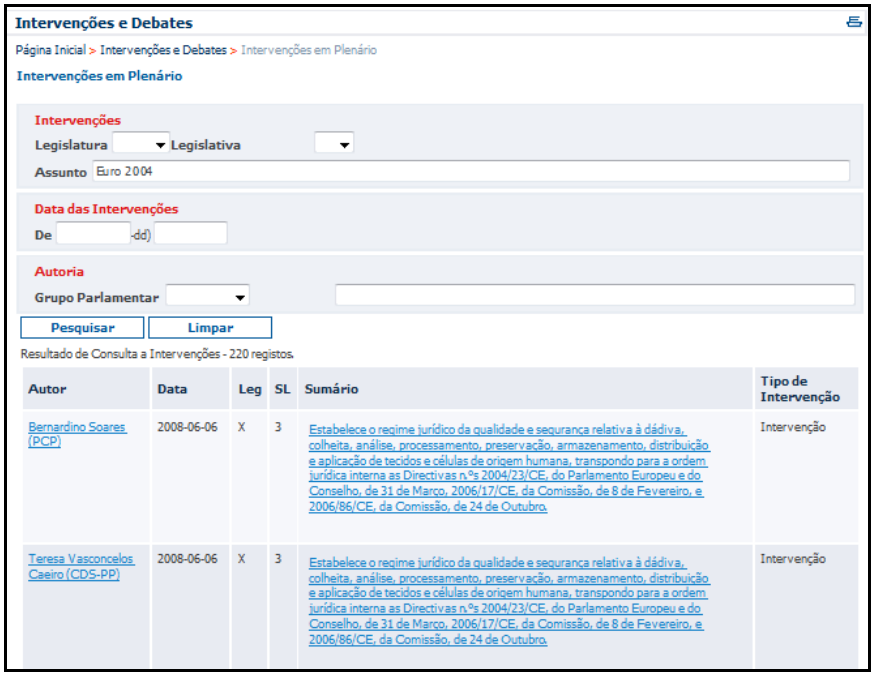


ILUSTRAÇÃO 7 - PÁGINA DA AR, SECÇÃO "INTERVENÇÕES E DEBATES, INTERVENÇÕES EM PLENÁRIO", ONDE SÃO VISÍVEIS ALGUNS RESULTADOS DA PESQUISA POR "EURO 2004".

7.2.3 Debates Parlamentares

Esta secção encontra-se organizada em quatro separadores, de acordo com a organização cronológica/política dos órgãos governativos nacionais, da Monarquia Constitucional à actualidade: 1) 3.ª República; 2) Estado Novo (1935-1974); 3) 1.ª República (1910-1926); 4) Monarquia Constitucional (1821-1910).

Em cada um destes separadores é possível consultar as publicações relativas ao período respectivo. A "Monarquia Constitucional" cobre a actividade das "Cortes Geraes e Extraordinárias da Nação Portugueza" (1821-1822), da "Câmara dos Senhores Deputados" (1822-1910), da "Câmara dos Pares do Reino" (1826-1838), das "Cortes Geraes, Extraordinárias e Constituintes da Nação Portugueza" (1837-1838), da "Câmara dos Senadores" (1838-1842) e da Câmara dos Pares do Reino (1842-1910). A "1.ª República" inclui os trabalhos parlamentares da "Assembleia Nacional Constituinte" (1911), da "Câmara dos Deputados" (1911-1926), do Senado da República (1911-1926) e do Congresso da República (1911-1926). O "Estado Novo" possui os textos referentes aos Diários das Sessões da "Assembleia Nacional" (1935-1974) e da "Câmara Corporativa" (1935-1974). No período da 3.ª República, a base de dados divide-se da seguinte forma:

A I Série contém os textos integrais de todas as intervenções parlamentares feitas em plenário, quer na "Assembleia Constituinte" (1975-1976), quer na "Assembleia da República" (1976-);

A II Série é composta por cinco sub-séries:

II Série-A - onde são publicados os decretos, resoluções e deliberações do Plenário, os textos dos projectos de revisão constitucional, projectos e propostas de lei; projectos e propostas de resolução e de referendo, projectos de deliberação, pareceres e outros textos aprovados em Comissão;

II Série-B - onde são publicados os textos dos votos, interpelações, inquéritos parlamentares, as perguntas formuladas por escrito ao Governo e os requerimentos referidos nas alíneas d) e e) do artigo 156.º da Constituição, bem como as respectivas respostas, e os textos e relatórios das petições que devam ser publicados nos termos da lei e aqueles a que a comissão parlamentar competente entenda dar publicidade;

II Série-C – que contém os relatórios da actividade das comissões parlamentares, bem como das delegações da AR e as actas das comissões parlamentares e das audições parlamentares, quando deliberada a sua publicação;

II Série-D – onde são publicadas as intervenções dos deputados em instâncias internacionais, quando em representação da AR, desde que constem integralmente dos respectivos registos, bem como das delegações da Assembleia e os documentos relativos à constituição e composição dos grupos parlamentares de amizade;

Série-E – que inclui os despachos do Presidente da Assembleia e dos Vice-Presidentes, o orçamento e as contas da AR, e os relatórios da actividade da Assembleia e da Auditoria Jurídica, as deliberações, recomendações, pareceres e relatórios dos órgãos independentes que funcionam junto da AR, como a Comissão Nacional de Eleições (CNE), a Comissão de Acesso aos Documentos Administrativos (CADA) ou a Entidade Reguladora para a Comunicação Social (ERCS), documentos relativos ao pessoal da AR e outros documentos que, nos termos da lei ou do Regimento, devam ser publicados, bem como os que o Presidente da AR entenda mandar publicar.

Na base de dados Debates Parlamentares estão ainda disponíveis a I e II Série RC que abrangem os textos dos debates relativos às sucessivas revisões constitucionais (1982, 1989, 1992, 1997, 2001 e 2004 e 2005) realizadas em plenário (I Série RC) e nas comissões eventuais para a revisão constitucional (II Série RC).

7.2.4 Deputados e Grupos Parlamentares

A secção "Deputados e Grupos Parlamentares" encontra-se subdividida em várias subcategorias:

7.2.4.1 Deputados

Aqui podem consultar-se os nomes dos deputados presentes na AR tendo em conta a legislatura, o GP, o círculo eleitoral, a situação[54] e a data (ver, de seguida, ilustração 5).

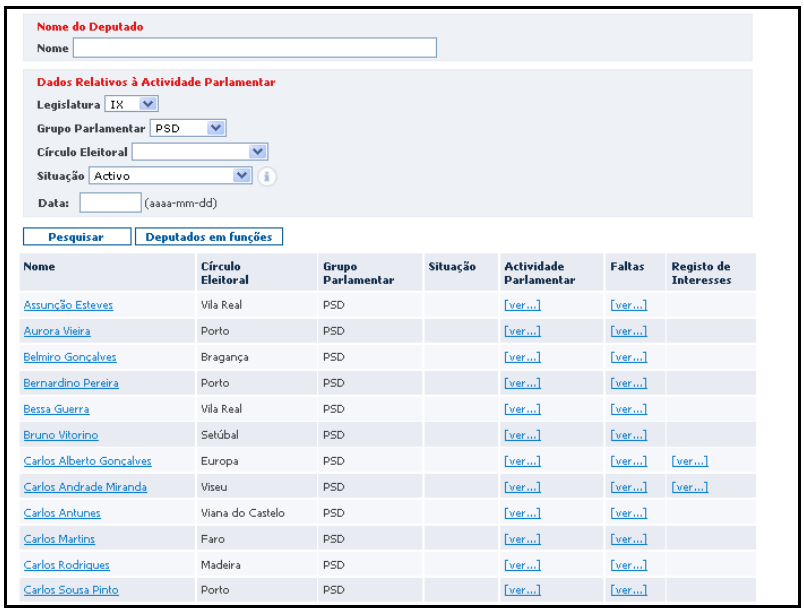
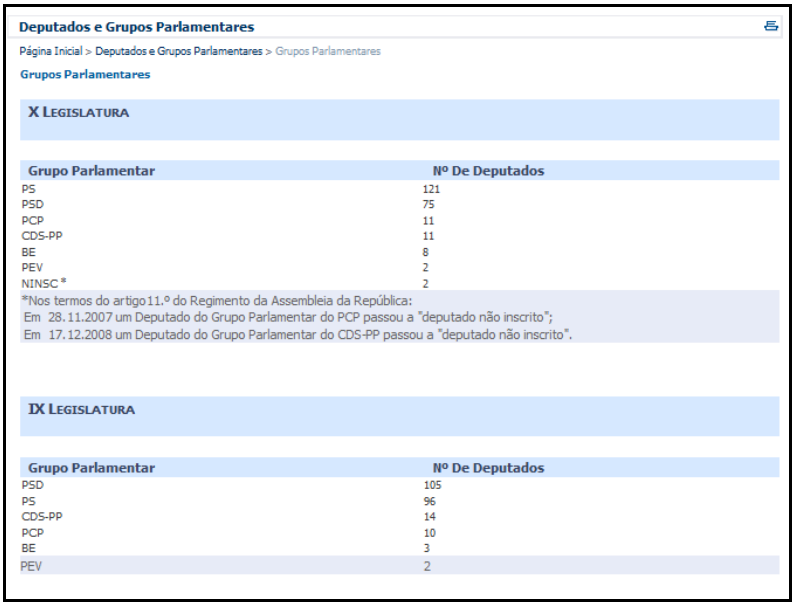


ILUSTRAÇÃO 8 - PÁGINA DA AR, ONDE SE ILUSTRA A PESQUISA DE DEPUTADOS TENDO EM CONTA A LEGISLATURA, O GP E A SITUAÇÃO.

7.2.4.2 Grupos Parlamentares

Neste campo não é possível efectuar pesquisas, sendo fornecida ao utilizador informação sobre a constituição partidária dos grupos parlamentares nas quatro legislaturas mais recentes (ver ilustração 6).

[54] Activo; efectivo; efectivo definitivo; efectivo temporário; impedido; inactivo; renunciou; suplente; suspenso (efectivo def); suspenso (eleito); suspenso (não eleito).



Deputados e Grupos Parlamentares

Página Inicial > Deputados e Grupos Parlamentares > Grupos Parlamentares

Grupos Parlamentares

X LEGISLATURA

Grupo Parlamentar	Nº De Deputados
PS	121
PSD	75
PCP	11
CDS-PP	11
BE	8
PEV	2
NINSC *	2

*Nos termos do artigo 11.º do Regimento da Assembleia da República:
Em 28.11.2007 um Deputado do Grupo Parlamentar do PCP passou a "deputado não inscrito";
Em 17.12.2008 um Deputado do Grupo Parlamentar do CDS-PP passou a "deputado não inscrito".

IX LEGISLATURA

Grupo Parlamentar	Nº De Deputados
PSD	105
PS	96
CDS-PP	14
PCP	10
BE	3
PEV	2

ILUSTRAÇÃO 9 - PÁGINA DA AR, SECÇÃO GRUPOS PARLAMENTARES.

7.2.5 Mesa da Assembleia

Permite estudar a constituição da Mesa da Assembleia, seleccionando a legislatura e/ou datas pretendidas (ver ilustração 7).

Deputados e Grupos Parlamentares

Página Inicial > Deputados e Grupos Parlamentares > Mesa da Assembleia

Mesa da Assembleia

Legislatura: IX Data: (aaaa-mm-dd)

Pesquisar

Nome	Cargo	Grupo Parlamentar
António Filipe	Vice-Presidente de 2002-04-05 a 2005-03-09	PCP
António Galamba	Secretário de 2002-04-05 a 2005-03-09	PS
Duarte Pacheco	Secretário de 2002-04-05 a 2005-03-09	PSD
Fernando Santos Pereira	Vice-Secretário de 2002-04-05 a 2005-03-09	PSD
Henrique Campos Cunha	Secretário de 2004-01-22 a 2005-03-09	CDS-PP
Manuel Alegre	Vice-Presidente de 2002-04-05 a 2005-03-09	PS
Manuel Oliveira	Vice-Secretário de 2002-04-05 a 2005-03-09	PSD
Maria Leonor Beleza	Vice-Presidente de 2002-04-05 a 2005-03-09	PSD
Miguel Coelho	Vice-Secretário de 2002-04-05 a 2005-03-09	PS
Mota Amaral	Presidente de 2002-04-05 a 2005-03-09	PSD
Narana Coissoró	Vice-Presidente de 2002-04-05 a 2005-03-09	CDS-PP
Rodeia Machado	Secretário de 2002-04-05 a 2005-03-09	PCP
Rosa Maria Albernaz	Vice-Secretário de 2002-04-05 a 2005-03-09	PS

ILUSTRAÇÃO 10 - PÁGINA DA AR ONDE SE ILUSTRA A PESQUISA SOBRE A MESA DA ASSEMBLEIA, TENDO EM CONTA A LEGISLATURA SELECCIONADA.

7.2.6 Conferência de Líderes

Possibilita a consulta de informação sobre os membros que integram a conferência de líderes, seleccionando a legislatura e/ou datas pretendidas (ver ilustração 8).

ILUSTRAÇÃO 11 - PÁGINA DA AR ONDE SE DEMONSTRA A PESQUISA DE INFORMAÇÃO SOBRE A CONFERÊNCIA DE LÍDERES, TENDO EM CONTA A LEGISLATURA SELECCIONADA.

7.2.6.1 Presidentes dos Grupos Parlamentares

Não é possível efectuar pesquisas, sendo apresentada uma listagem dos presidentes dos grupos parlamentares da legislatura vigente (no momento da realização deste trabalho, encontrava-se disponível a listagem dos presidentes dos grupos parlamentares da X Legislatura - ver ilustração 9).

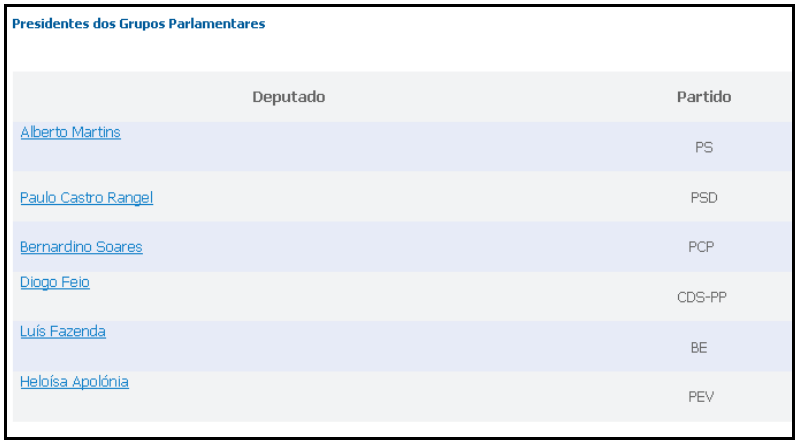


ILUSTRAÇÃO 12 - PÁGINA DA AR ONDE SE VISUALIZAM OS NOMES DOS PRESIDENTES DOS GRUPOS PARLAMENTARES NO MOMENTO DA X LEGISLATURA.

7.2.7 Comissão Permanente

Consulta da lista de membros que integram a comissão permanente, seleccionando a legislatura e/ou datas pretendidas. No momento de realização do nosso trabalho, encontrava-se indisponível informação sobre legislaturas anteriores à VI (ver ilustração 10).

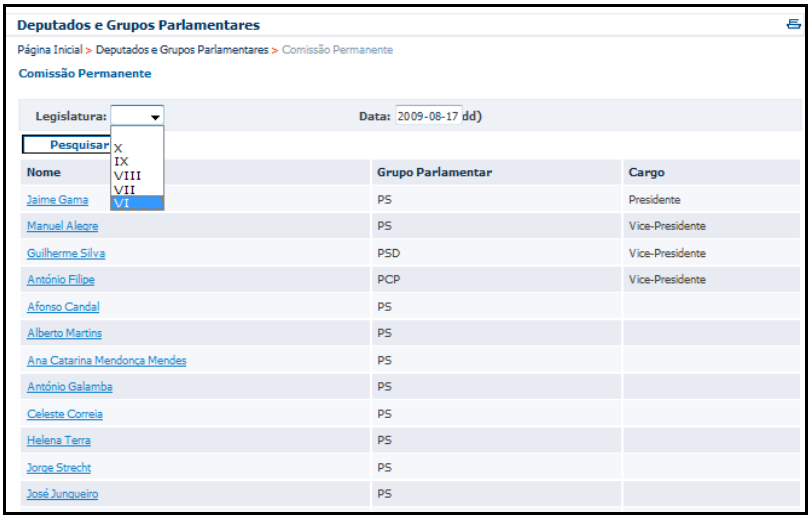


ILUSTRAÇÃO 13 - PÁGINA DA AR ONDE SE ILUSTRA A PESQUISA DE INFORMAÇÃO SOBRE COMISSÃO PERMANENTE, TENDO EM CONTA A LEGISLATURA SELECCIONADA.

7.2.8 Páginas Pessoais

Neste separador são apresentados *links* para as páginas pessoais dos deputados dos vários GP (ver ilustração 11).

ILUSTRAÇÃO 14 - PÁGINA DA AR, SECÇÃO DEBATES PARLAMENTARES, PÁGINAS PESSOAIS, ONDE É POSSÍVEL VISUALIZAR OS LINKS ASSOCIADOS AOS MEMBROS DOS DIFERENTES PARTIDOS POLÍTICOS.

7.2.9 Blogs

É aberta uma nova página (http://blogs.parlamento.pt/indice/), onde se pode consultar um blogue com participações dos diferentes deputados.

7.2.10 Resultados Eleitorais

Apresentação dos resultados eleitorais desde a I legislatura (ver ilustração 12).

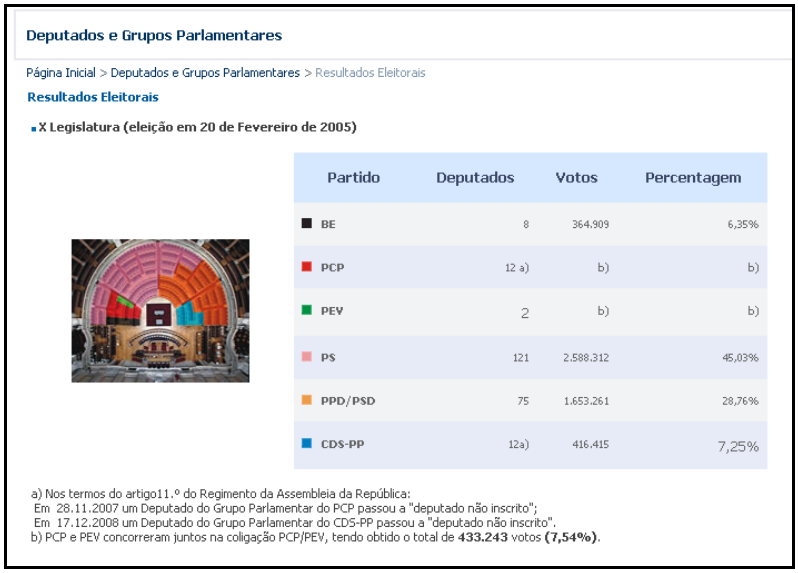


ILUSTRAÇÃO 15 - PÁGINA DA AR, SECÇÃO DEBATES PARLAMENTARES, RESULTADOS ELEITORAIS, ONDE SE VÊEM OS RESULTADOS ELEITORAIS DA X LEGISLATURA.

7.2.11 Estatuto dos Deputados

Somos redireccionados para uma página com legislação sobre o estatuto dos deputados (ver ilustração 13).

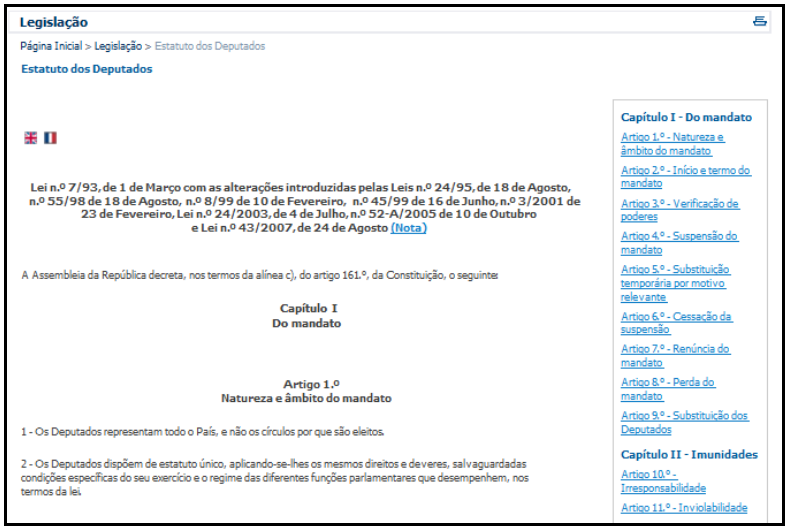
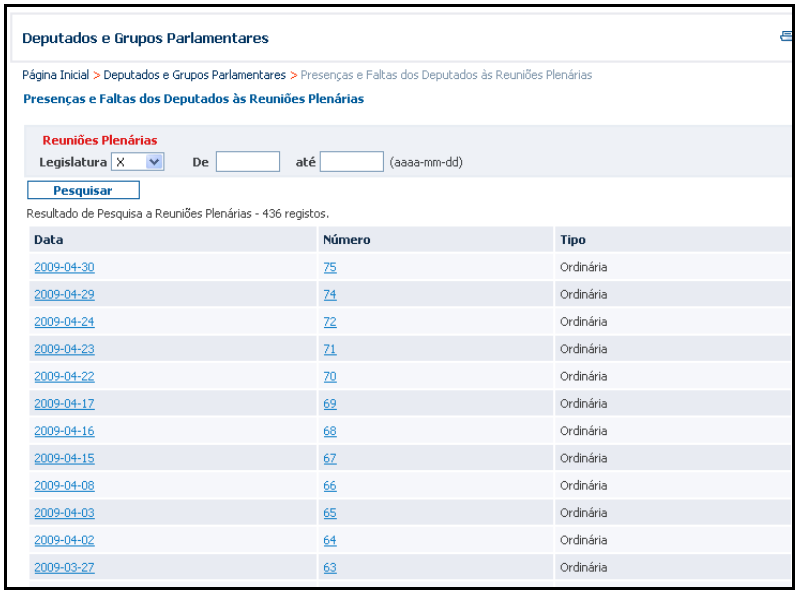


ILUSTRAÇÃO 16 - PÁGINA DA AR, SECÇÃO DEBATES PARLAMENTARES, ESTATUTO DOS DEPUTADOS.

7.2.12 Presenças e faltas dos deputados às reuniões plenárias

Permite consultar, por sessão plenária, quais os deputados presentes e os faltosos (ver ilustrações 14 e 15).



Deputados e Grupos Parlamentares

Página Inicial > Deputados e Grupos Parlamentares > Presenças e Faltas dos Deputados às Reuniões Plenárias

Presenças e Faltas dos Deputados às Reuniões Plenárias

Reuniões Plenárias

Legislatura X De até (aaaa-mm-dd)

Pesquisar

Resultado de Pesquisa a Reuniões Plenárias - 436 registos.

Data	Número	Tipo
2009-04-30	75	Ordinária
2009-04-29	74	Ordinária
2009-04-24	72	Ordinária
2009-04-23	71	Ordinária
2009-04-22	70	Ordinária
2009-04-17	69	Ordinária
2009-04-16	68	Ordinária
2009-04-15	67	Ordinária
2009-04-08	66	Ordinária
2009-04-03	65	Ordinária
2009-04-02	64	Ordinária
2009-03-27	63	Ordinária

ILUSTRAÇÃO 17 - PÁGINA DA AR, SECÇÃO DEBATES PARLAMENTARES, PRESENÇAS E FALTAS DOS DEPUTADOS ÀS REUNIÕES PLENÁRIAS, ONDE É VISÍVEL O TIPO DE PESQUISA POR SESSÃO PLENÁRIA.

Deputados e Grupos Parlamentares

Página Inicial > Deputados e Grupos Parlamentares > Reunião Plenária

Reunião Plenária

voltar à pesquisa

Reunião Plenária Ordinária de 2008-02-21.

Deputado	Grupo Parlamentar	Presença/Falta	Motivo
Abel Baptista	CDS-PP	Presença (P)	
Adão Silva	PSD	Presença (P)	
Afonso Candal	PS	Presença (P)	
Agostinho Branquinho	PSD	Presença (P)	
Agostinho Gonçalves	PS	Presença (P)	
Agostinho Lopes	PCP	Presença (P)	
Alberto Antunes	PS	Presença (P)	
Alberto Arons de Carvalho	PS	Presença (P)	
Alberto Martins	PS	Presença (P)	
Alcídia Lopes	PS	Presença (P)	
Aldemira Pinho	PS	Presença (P)	
Ana Catarina Mendonça Mendes	PS	Presença (P)	
Ana Couto	PS	Presença (P)	
Ana Drago	BE	Presença (P)	

ILUSTRAÇÃO 18 - PÁGINA DA AR, SECÇÃO DEBATES PARLAMENTARES, PRESENÇAS E FALTAS DOS DEPUTADOS ÀS REUNIÕES PLENÁRIAS, ONDE É VISÍVEL UMA PESQUISA FEITA À PRESENÇA E FALTA DE DEPUTADOS PARA A SESSÃO DE DIA 21-02-2008

7.3 Imagens dos resultados obtidos com o Teragram TK240, durante a realização dos testes

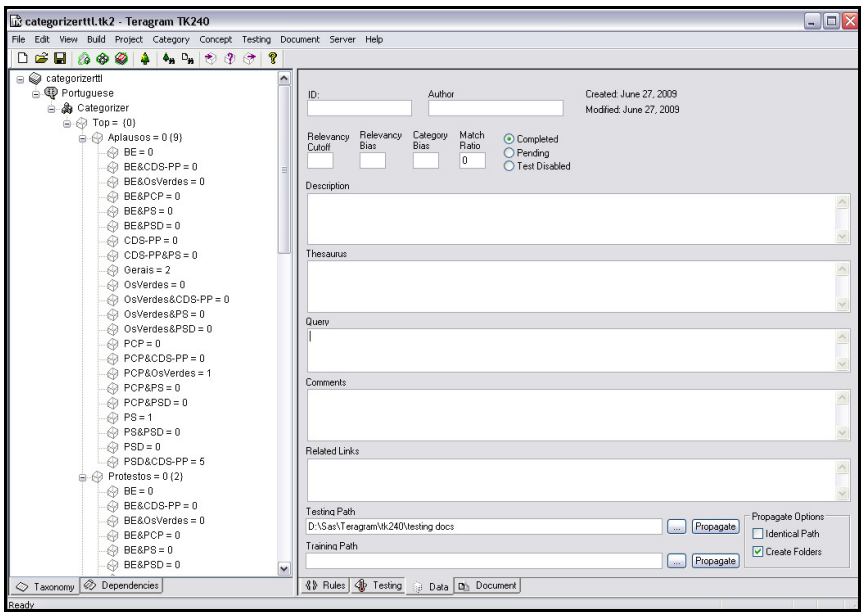


ILUSTRAÇÃO 19 - RESULTADOS DA CATEGORIA "APLAUSOS" (DOCUMENTOS DE TESTE).

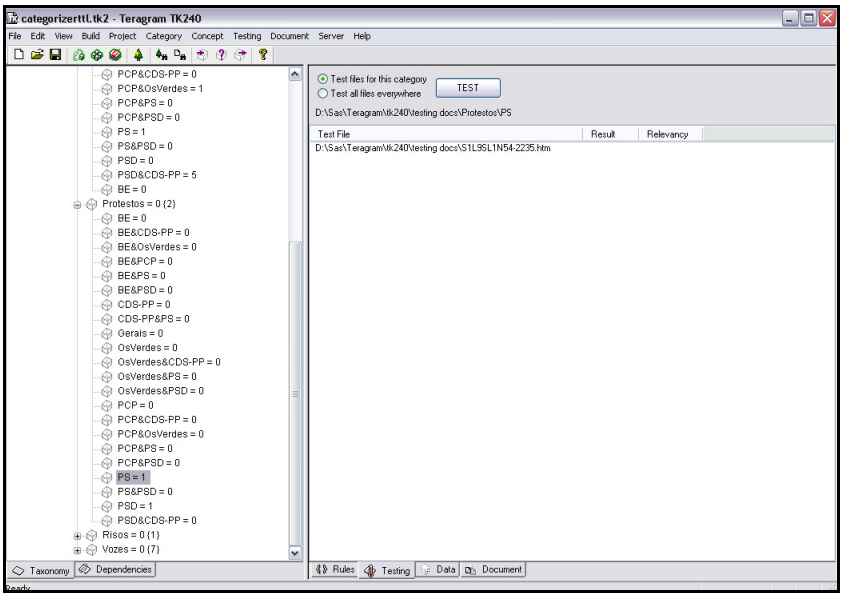


ILUSTRAÇÃO 20 - RESULTADOS DA CATEGORIA "PROTESTOS" (DOCUMENTOS DE TESTE).

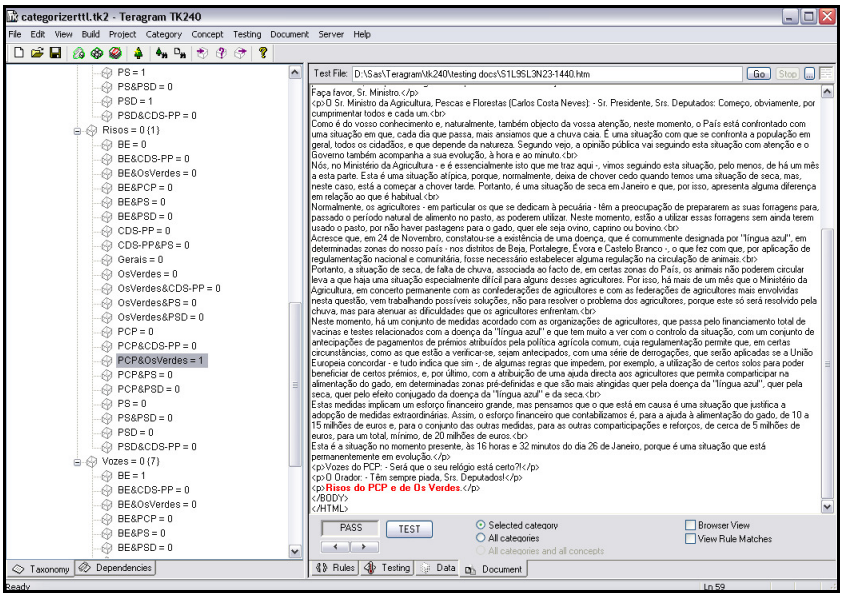


ILUSTRAÇÃO 21 - RESULTADOS DA CATEGORIA "RISOS" (DOCUMENTOS DE TESTE).

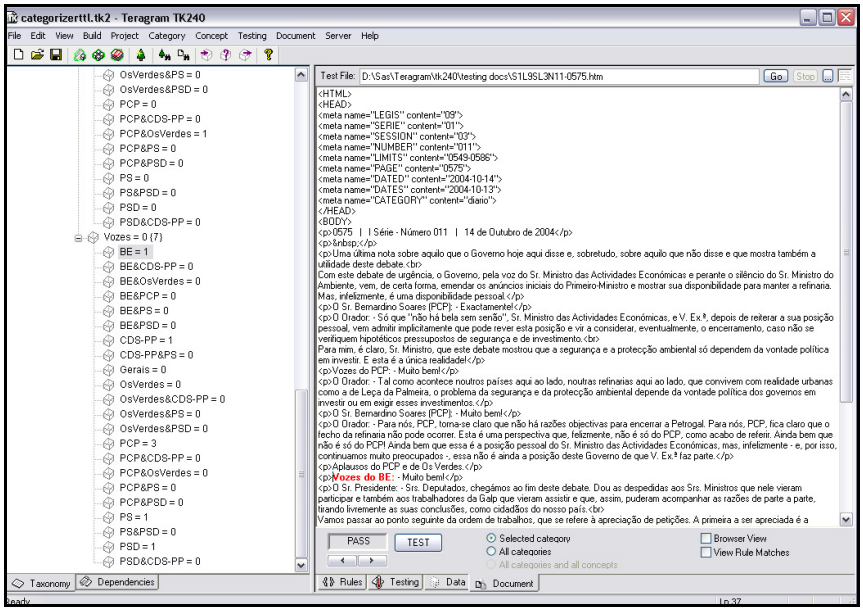


ILUSTRAÇÃO 22 - RESULTADOS DA CATEGORIA "VOZES" (DOCUMENTOS DE TESTE).

7.4 Imagens dos resultados obtidos com o Teragram TK240, durante o processamento dos ficheiros

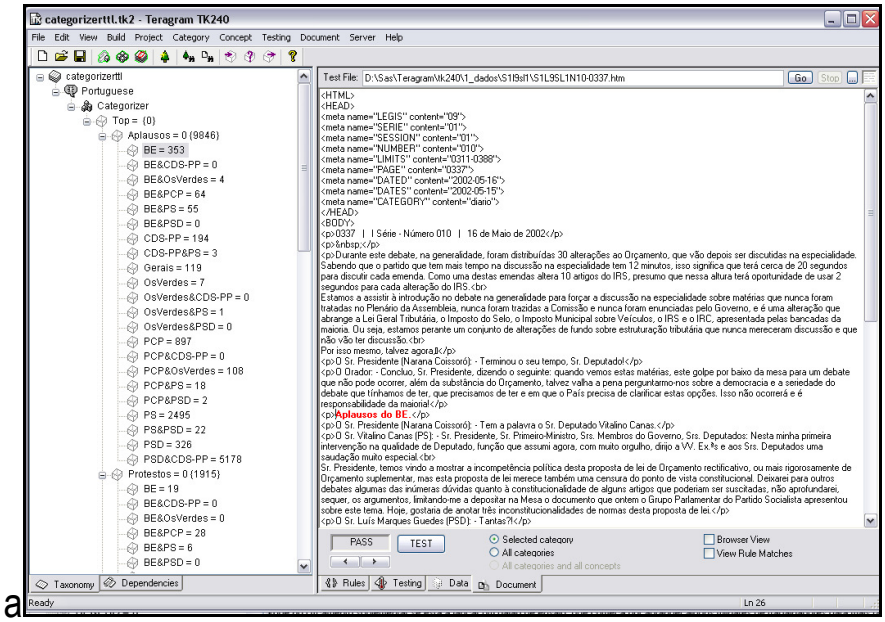


ILUSTRAÇÃO 23 - RESULTADOS OBTIDOS NA CATEGORIA "APLAUSOS".

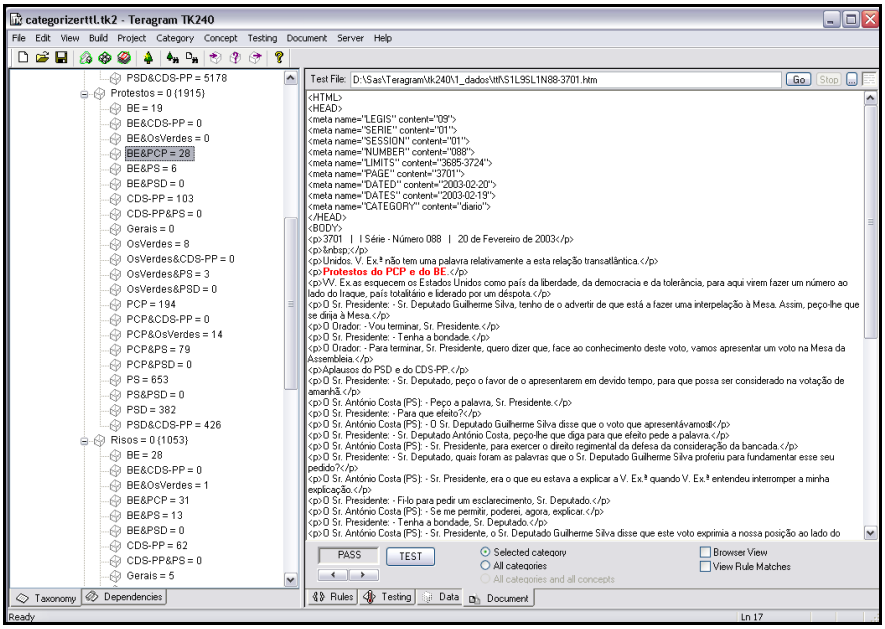


ILUSTRAÇÃO 24 - RESULTADOS OBTIDOS NA CATEGORIA "PROTESTOS".

I

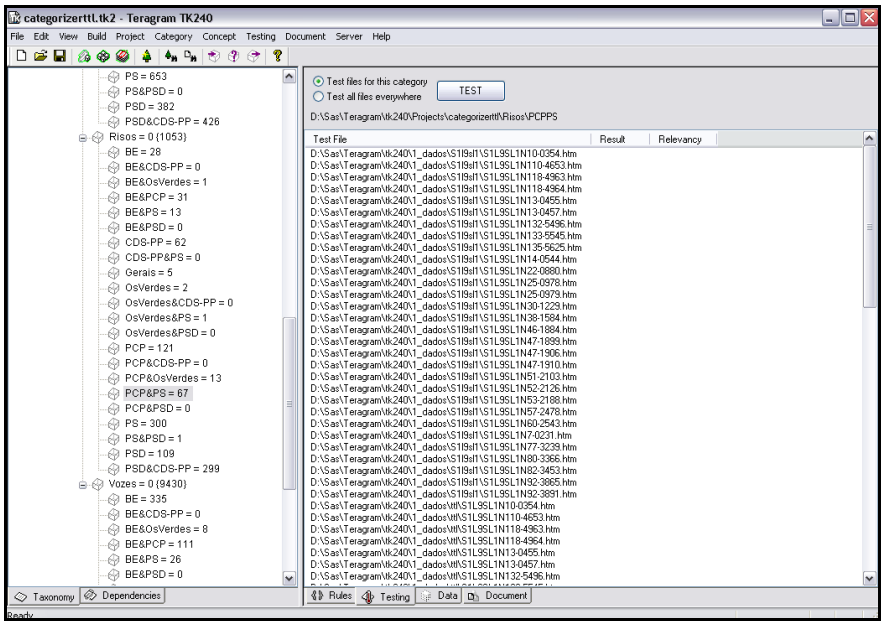


ILUSTRAÇÃO 25 - RESULTADOS OBTIDOS NA CATEGORIA "RISOS"

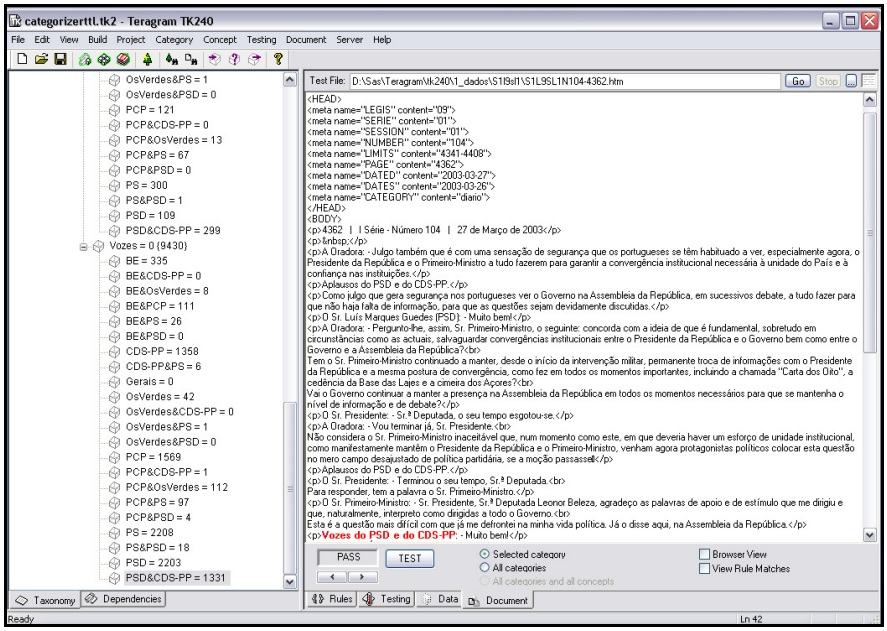


ILUSTRAÇÃO 26 - RESULTADOS OBTIDOS NA CATEGORIA "VOZES"

7.5 Valores Absolutos relativamente aos dados processados

Aplausos XV Governo Constitucional

		PSD	PS	CDS-PP	PCP	BE	Os Verdes
Total do GP		4859	2306	4714	971	434	96
GP Isolado		298	2213	172	804	318	7
GP com outros Partidos	**Total**	4561	93	4542	167	116	89
	Esquerda	22	69	3	166	116	89
	Direita	4539	24	4539	1	0	0

Aplausos XVI Governo Constitucional

		PSD	PS	CDS-PP	PCP	BE	Os Verdes
Total do GP		658	286	650	117	40	23
GP Isolado		28	282	22	93	35	0
GP com outros Partidos	**Total**	630	4	628	24	5	23
	Esquerda	2	3	0	23	5	23
	Direita	628	1	628	1	0	0

Protestos XV Governo Constitucional

		PSD	PS	CDS-PP	PCP	BE	Os Verdes
Total do GP		698	677	462	290	49	23
GP Isolado		334	595	98	178	17	7
GP com outros Partidos	**Total**	364	82	364	112	32	16
	Esquerda	0	82	0	112	32	16
	Direita	364	0	364	0	0	0

Protestos XVI Governo Constitucional

		PSD	PS	CDS-PP	PCP	BE	Os Verdes
Total do GP		106	63	63	24	4	2
GP Isolado		48	58	5	16	2	1
GP com outros Partidos	**Total**	58	5	58	8	2	1
	Esquerda	0	5	0	8	2	1
	Direita	58	0	58	0	0	0

Risos XV Governo Constitucional

		PSD	PS	CDS-PP	PCP	BE	Os Verdes
Total do GP		354	343	310	200	55	14
GP Isolado		98	270	55	106	21	2
GP com outros Partidos	**Total**	256	73	255	94	34	12
	Esquerda	1	72	0	94	34	12
	Direita	255	1	255	0	0	0

Risos XVI Governo Constitucional

		PSD	PS	CDS-PP	PCP	BE	Os Verdes
Total do GP		54	38	50	31	18	3
GP Isolado		11	30	7	15	7	0
GP com outros Partidos	**Total**	43	8	43	16	11	3
	Esquerda	0	8	0	16	11	3
	Direita	43	0	43	0	0	0

Vozes XV Governo Constitucional

		PSD	PS	CDS-PP	PCP	BE	Os Verdes
Total do GP		3237	2128	2457	1714	435	135
GP Isolado		1993	1987	1227	1424	302	37
GP com outros Partidos	**Total**	1244	141	1230	290	133	98
	Esquerda	21	118	7	285	109	98
	Direita	1223	23	1223	5	24	0

Vozes XVI Governo Constitucional

		PSD	PS	CDS-PP	PCP	BE	Os Verdes
Total do GP		319	228	239	180	45	28
GP Isolado		210	221	131	145	33	5
GP com outros Partidos	**Total**	109	7	108	35	12	23
	Esquerda	1	6	0	35	12	23
	Direita	108	1	108	0	0	0

7.6 Exemplo de Utilização do Software Teragram TK 240

Como criar um projecto:

Exemplo: criar um novo projecto ("Categorizerttl") e arquivá-lo no caminho seleccionado: D:\SAS\Teragram\tk240\Projects

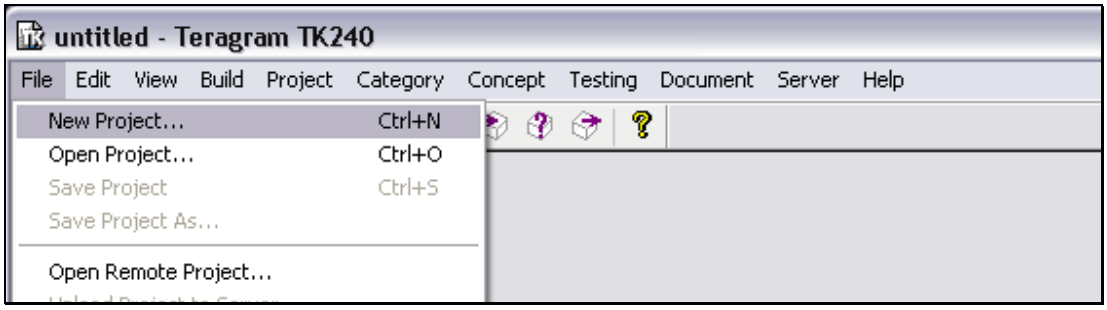


ILUSTRAÇÃO 27 - CRIAÇÃO DE UM NOVO PROJECTO.

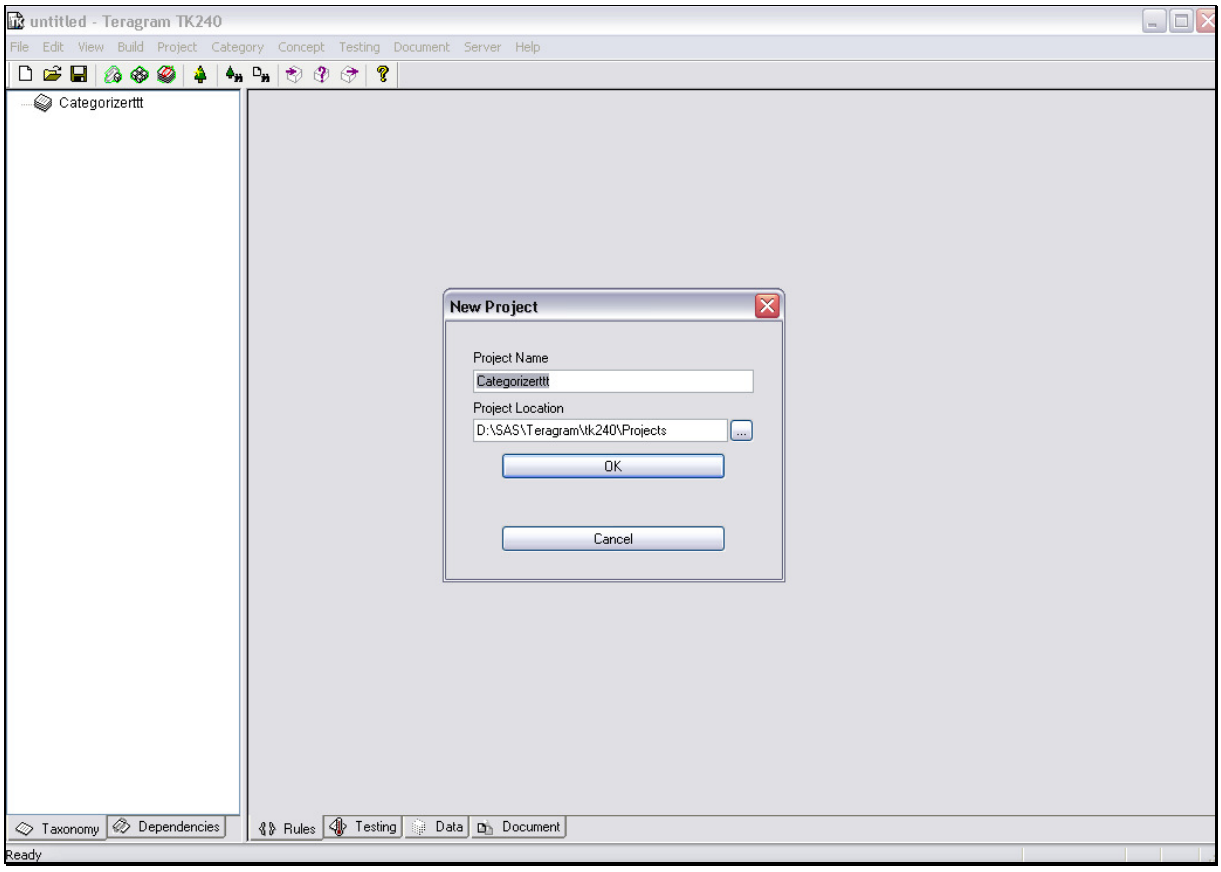


ILUSTRAÇÃO 28 - NOMEAR O PROJECTO E SELECCIONAR O CAMINHO ONDE ESTE FICARÁ GUARDADO

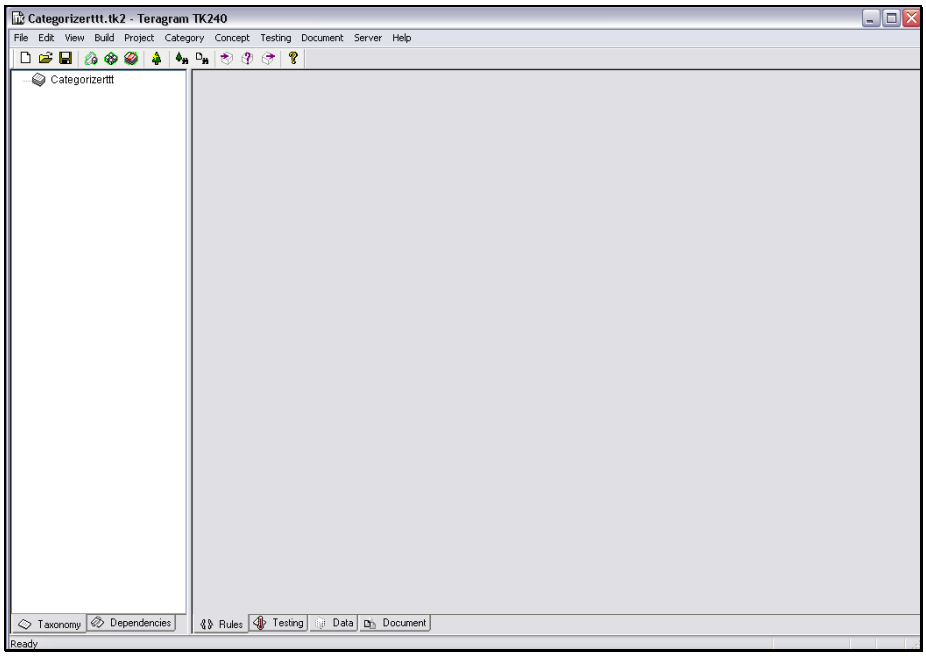


ILUSTRAÇÃO 29 - NOVO PROJECTO CRIADO E IDENTIFICADO, CORRESPONDENTE AO NÓ MAIS ALTO DA HIERARQUIA

Seleccionar a Língua em que se vai realizar o projecto:

O *Teragram* detecta automaticamente se a língua seleccionada necessita de codificação UTF-8[55]. Nos casos em que esta codificação não é necessária, é utilizado o Latin-1 como código de caracteres. Tal é o caso do português, como se pode ver na ilustração abaixo. Optámos por não utilizar a codificação UTF-8, uma vez que para tal seria necessário que o computador tivesse esta codificação de caracteres disponível.

[55] UTF-8 é um tipo de codificação de dimensão variável para Unicode. Permite representar todos os caracteres em standard Unicode, sendo compatível com ASCII. Por estes motivos, tem vindo a ser adoptada como a codificação preferencial para e-mails, páginas de internet e outros documentos em que os caracteres são armazenados.

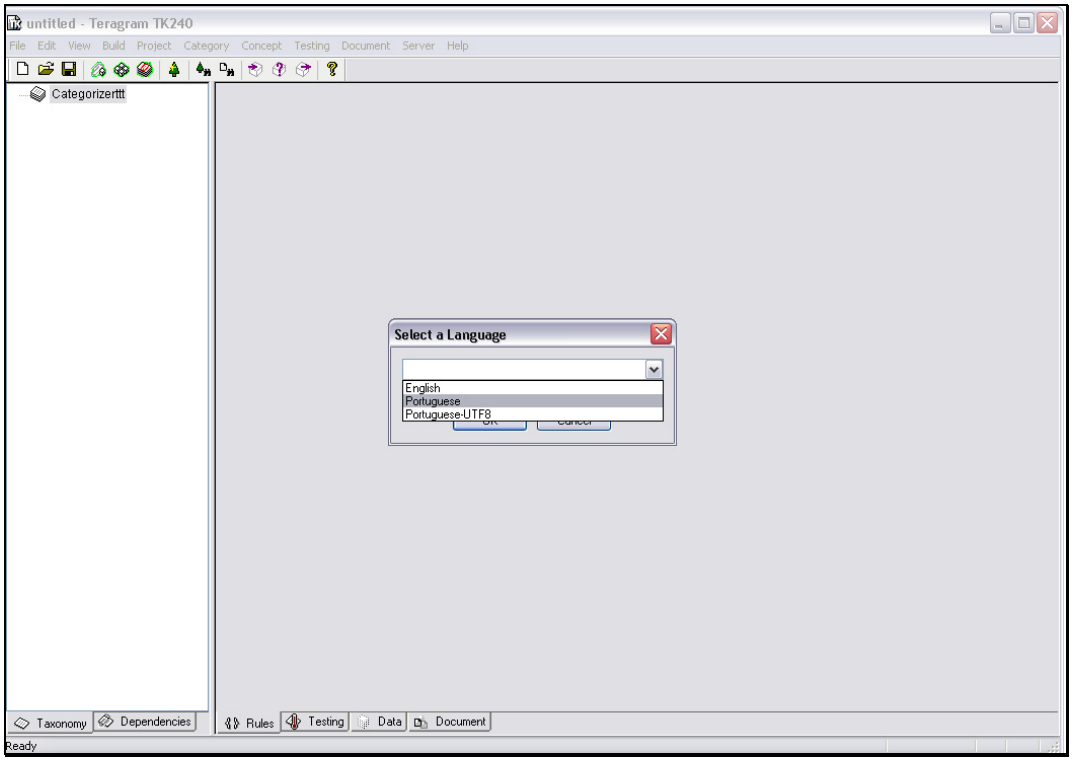


ILUSTRAÇÃO 30 - SELECÇÃO DA LÍNGUA EM QUE SE VAI REALIZAR O PROJECTO (NO CASO FOI SELECCIONADO O PORTUGUÊS).

ILUSTRAÇÃO 31 - PORMENOR DA SELECÇÃO DA LÍNGUA EM QUE SE VAI REALIZAR O PROJECTO (PORTUGUÊS).

Criar o Categorizer:

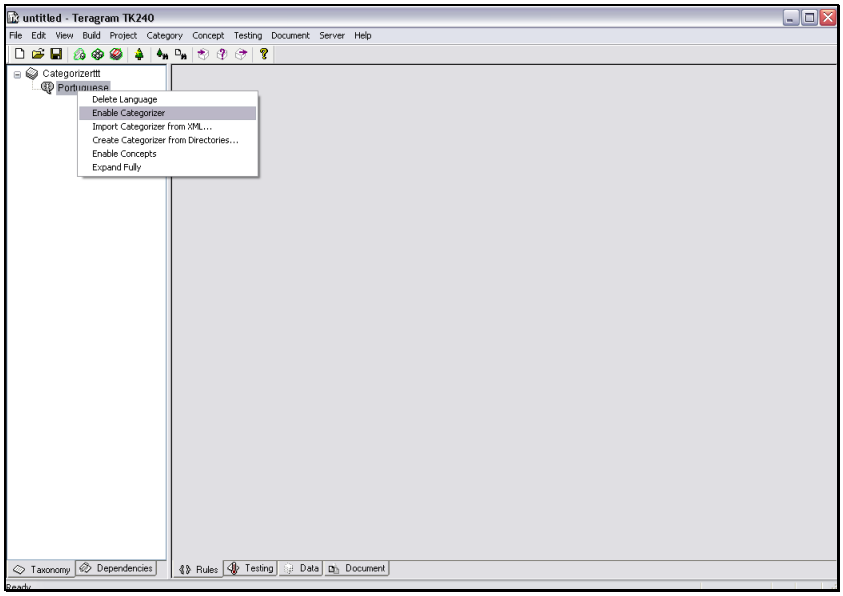


ILUSTRAÇÃO 32 - CRIAÇÃO DO CATEGORIZER, COM A SELECÇÃO DA OPÇÃO "ENABLE CATEGORIZER".

Criar uma categoria "pai":

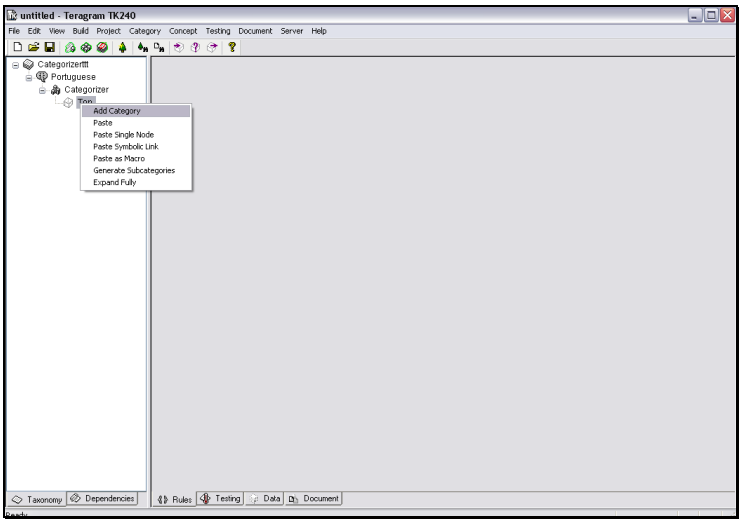


ILUSTRAÇÃO 33 - ADICIONAR UMA CATEGORIA "PAI" NA CONSTRUÇÃO DA TAXONOMIA

Criar uma categoria "filho":

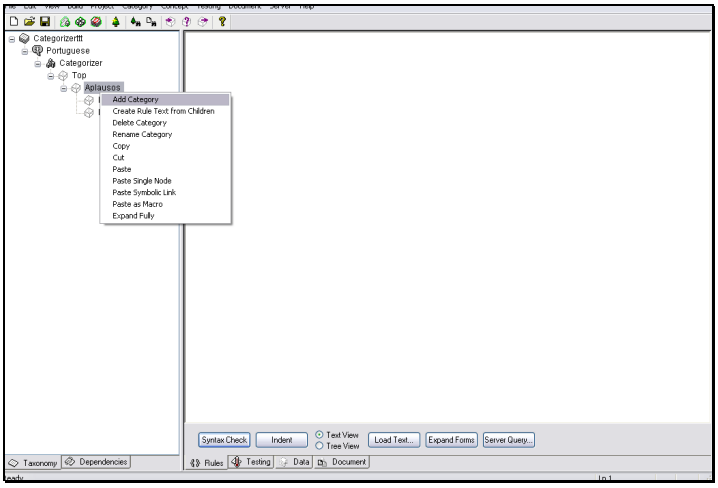


ILUSTRAÇÃO 34 - ADICIONAR UMA CATEGORIA "FILHO" NA CONSTRUÇÃO DA TAXONOMIA.

Criar as regras (janela "rules"):

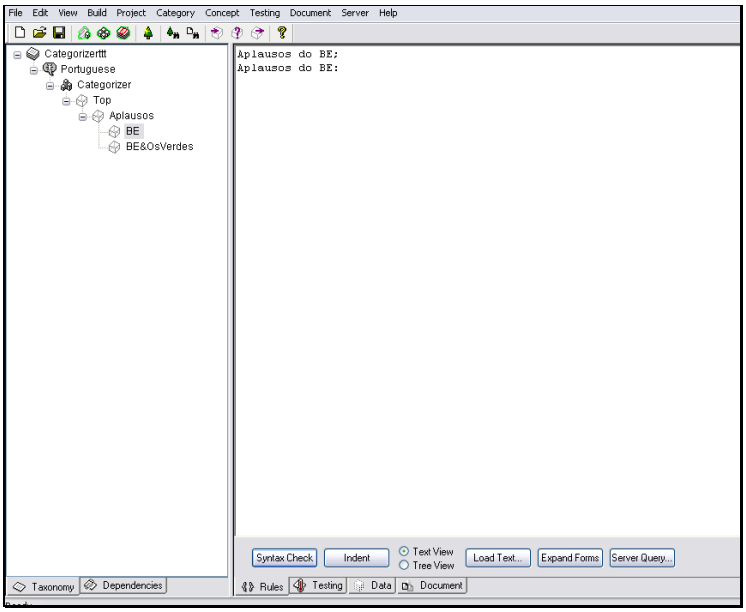


ILUSTRAÇÃO 35 - CRIAÇÃO DAS REGRAS LINGUÍSTICAS DENTRO DE UMA DADA CATEGORIA (NESTE CASO, CRIAÇÃO DAS REGRAS LINGUÍSTICAS PARA A CATEGORIA BE).

Criar a estrutura de pastas:

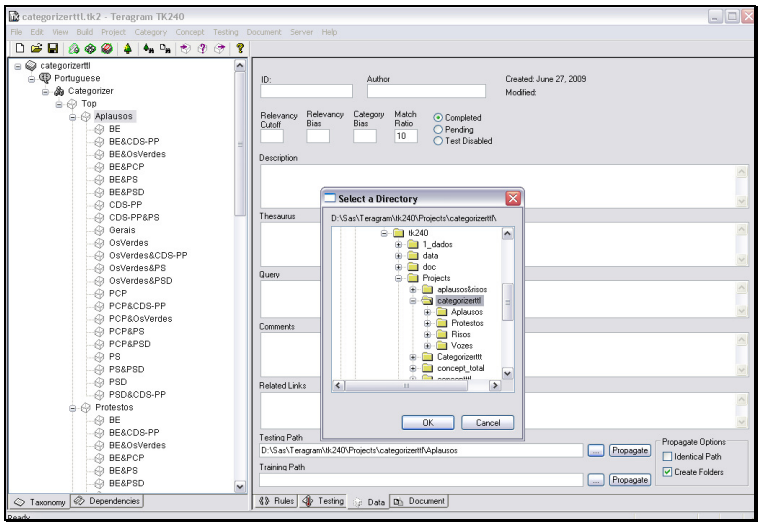


ILUSTRAÇÃO 36 - SELECCIONAR O CAMINHO, NO DISCO, ONDE SERÁ CRIADA AUTOMATICAMENTE UMA ESTRUTURA DE PASTAS IDÊNTICA À TAXONOMIA

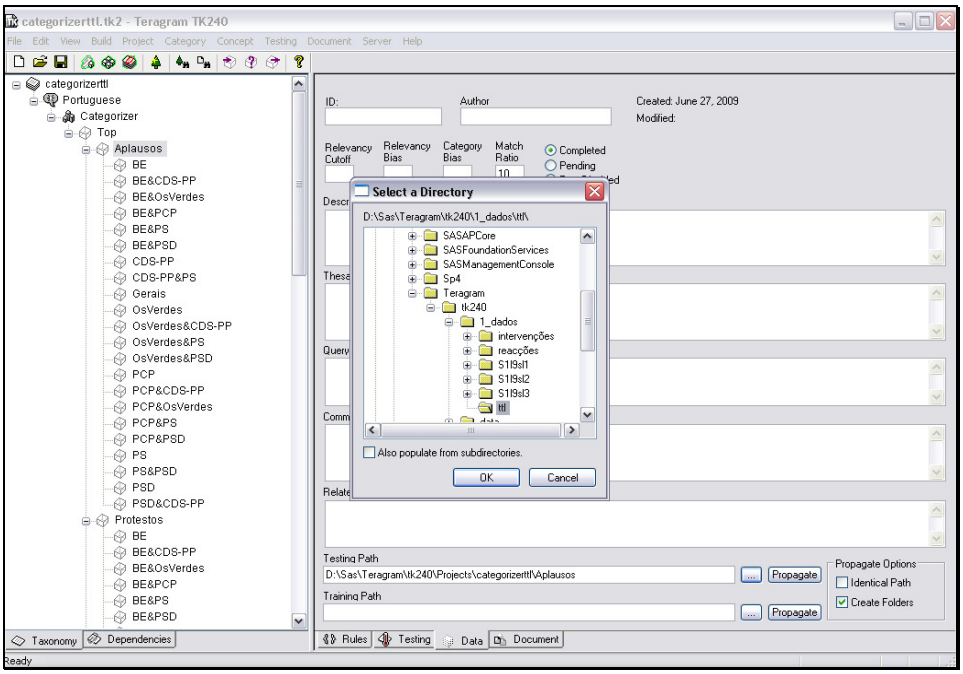


ILUSTRAÇÃO 37 - SELECÇÃO DO CAMINHO ONDE ESTÃO OS DOCUMENTOS DE INPUT (JANELA DATA).

Introduzir os documentos de *input* (popular a estrutura de pastas criada):

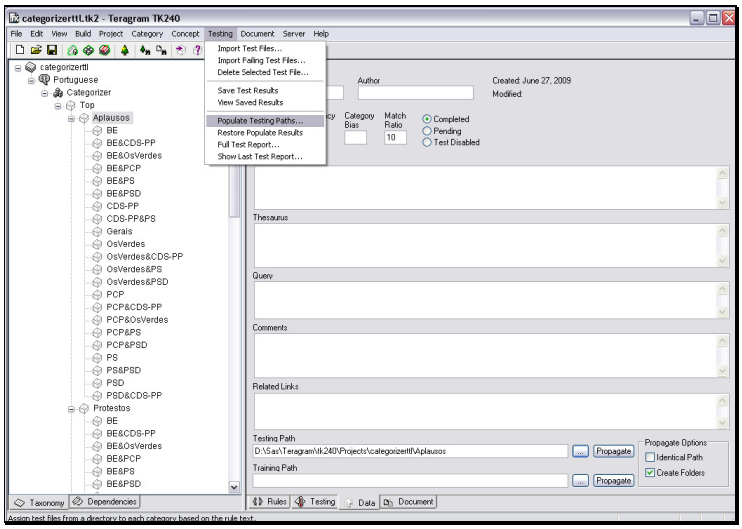


ILUSTRAÇÃO 38 - SELECÇÃO DA OPÇÃO POPULATE TESTING PATHS, DANDO-SE ASSIM INDICAÇÃO AO PROGRAMA PARA ORGANIZAR OS DOCUMENTOS DE INPUT NAS RESPECTIVAS CATEGORIAS.

Consultar os resultados do teste:

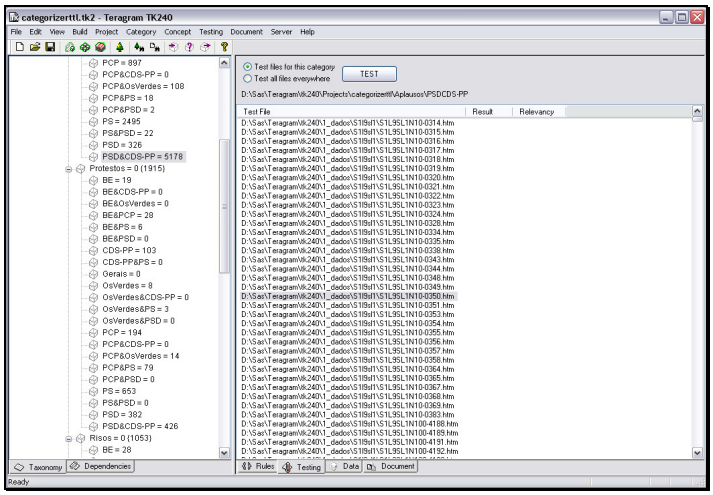


ILUSTRAÇÃO 39 - CONSULTAR A LISTAGEM DE DOCUMENTOS CATEGORIZADOS NUMA DADA CATEGORIA (NESTE CASO, APLAUSOS PSD&CDS-PP)

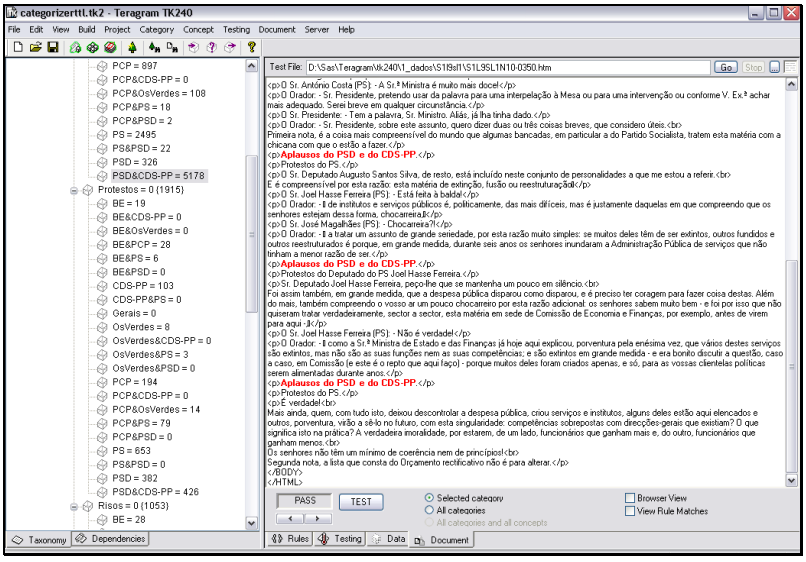


ILUSTRAÇÃO 40 - SELECÇÃO DE UM DOCUMENTO EM CONCRETO, ONDE ESTÃO ASSINALADAS A VERMELHO AS OCORRÊNCIAS DAS REGRAS LINGUÍSTICAS UTILIZADAS.

Pedir um Relatório dos resultados:

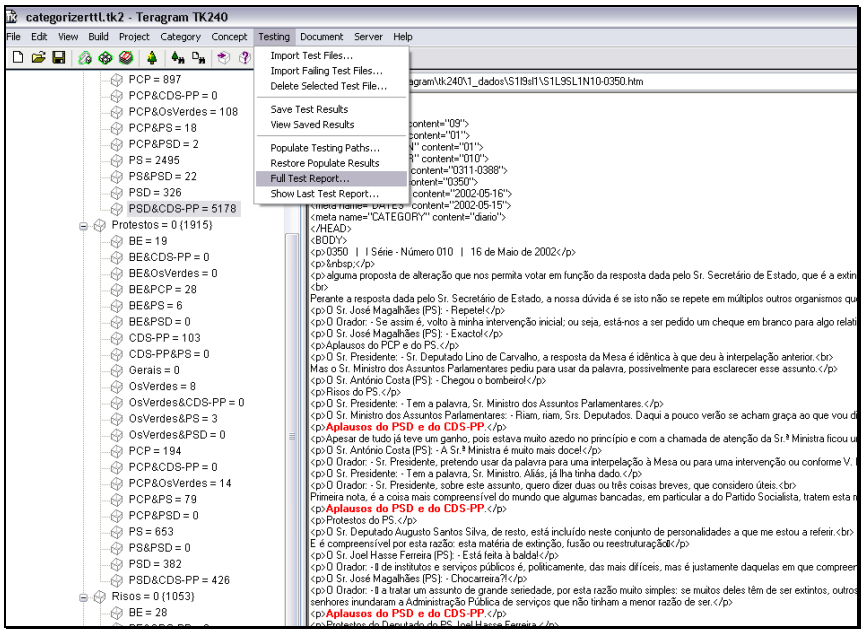


ILUSTRAÇÃO 41 - ESCOLHA DA OPÇÃO "FULL TEST REPORT" NA JANELA TESTING

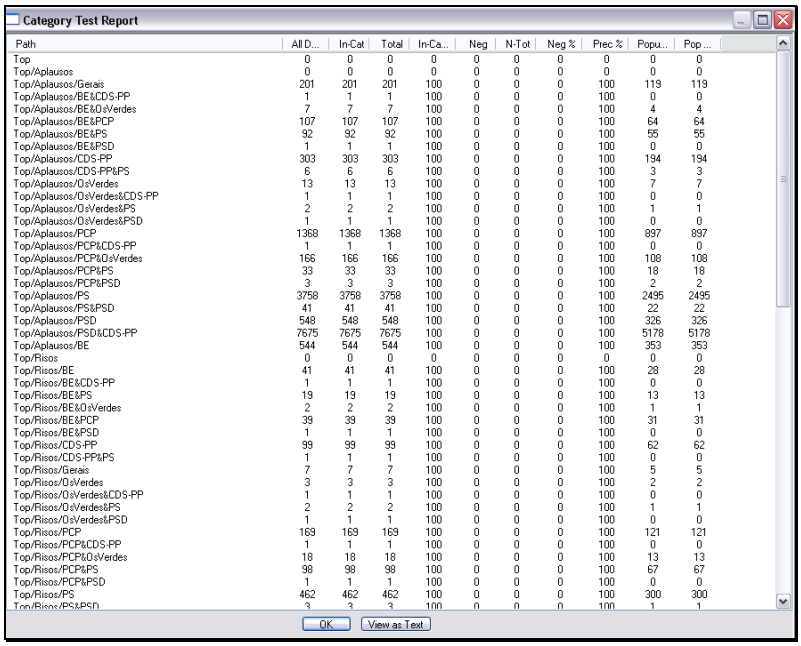


ILUSTRAÇÃO 42 - RELATÓRIO DOS RESULTADOS FORNECIDO PELO PROGRAMA TERAGRAM TK240

7.7 XV e XVI Governos Constitucionais

7.7.1 XV Governo Constitucional

6 Abril 2002 a 17 Julho 2004

Primeiro Ministro: José Manuel Durão Barroso		
Ministro	Estado	Manuela Ferreira Leite
Ministro	Finanças	Manuela Ferreira Leite
Ministro	Defesa Nacional	Paulo Portas
Ministro	Estado	Paulo Portas
Ministro	Negócios Estrangeiros e das Comunidades Portuguesas	Teresa Gouveia/ Martins da Cruz
Ministro	Admnistração Interna	António Figueiredo Lopes
Ministro	Justiça	Celeste Cardona
Ministro	Presidência	Nuno Morais Sarmento
Ministro	Assuntos Parlamentares	Luís Marques Mendes
Ministro	Adjunto do Primeiro Ministro	José Luís Arnaut
Ministro	Economia	Carlos Tavares
Ministro	Agricultura, Desenvolvimento Rural e Pescas	Armando Sevinate Pinto
Ministro	Educação	David Justino
Ministro	Ciência e Ensino Superior	Graça Carvalho
Ministro	Cultura	Pedro Roseta
Ministro	Saúde	Luís Filipe Pereira
Ministro	Segurança Social e Trabalho	António Bagão Félix
Ministro	Obras Públicas, Transportes e Habitação	António Carmona Rodrigues
Ministro	Cidades	Valente de Oliveira

7.7.2 XVI Governo Constitucional

17 Julho 2004 a 12-03-05

Primeiro-Ministro: Pedro Santana Lopes		
Ministro	Estado e Actividades Económicas	Álvaro Barreto
Ministro	Estado, Defesa Nacional e Assuntos do Mar	Paulo Portas
Ministro	Estado e Presidência	Nuno Morais Sarmento
Ministro	Finanças e Administração Pública	António Bagão Félix
Ministro	Negócios Estrangeiros e Comunidades Portuguesas	António Monteiro
Secretário de Estado	Assuntos Europeus	Mário David
Ministro	Administração Interna	Daniel Sanches
Ministro	Justiça	José Pedro Aguiar-Branco
Ministro	Cidades, Administração Local, Habitação e Desenvolvimento Regional	José Luís Arnaut
Ministro	Agricultura, Pescas e Florestas	Carlos da Costa Neves
Ministro	Educação	Maria do Carmo Seabra
Ministro	Saúde	Luís Filipe Pereira
Ministro	Ciência, Inovação e Ensino Superior	Maria da Graça Carvalho
Ministro	Cultura	Maria João Bustorff
Ministro	Segurança Social, Família e Criança	Fernando Negrão
Ministro	Obras Públicas, Transportes e Comunicações	António Mexia
Ministro	Ambiente e Ordenamento do Território	Luís Nobre Guedes
Ministro	Turismo	Telmo Correia
Adjunto PM	Adjunto do PM (17-07-04 a 24-11-04)	Henrique Chaves
Adjunto PM	Adjunto do PM (24-11-04 a 12-03-05)	Rui Gomes da Silva
Ministro	Juventude, Desporto e Reabilitação (24-11-04 a 02-12-04)	Henrique Chaves
Ministro	Assuntos Parlamentares (24-11-04 a 12-03-05)	Rui Gomes da Silva