FACULDADE DE CIÊNCIAS SOCIAIS E HUMANAS
UNIVERSIDADE NOVA DE LISBOA

Luís António Pissarra de Matos Agonia Pereira

Tese de Doutoramento em Ciências da Educação

Especialidade de Tecnologias, Redes e Multimédia na Educação e Formação

Contribuições para a Formulação de uma Metodologia de Ensino e Avaliação baseada na Análise Estatística de Textos em Português

Julho, 2013

Tese apresentada para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Ciências da Educação, Especialidade de Tecnologias, Rede e Multimédia na Educação e Formação, realizada sob a orientação científica dos

Professor Doutor João Nogueira

Professor Doutor Valter Martins Vairinhos

Aos meus pais

AGRADECIMENTOS

Ao meu Orientador, Professor Doutor João Nogueira, Professor desta Faculdade, pela orientação prestada e empenho com que supervisionou o desenvolvimento deste trabalho.

Ao meu Coorientador, Professor Doutor Valter Martins Vairinhos, expresso o meu profundo reconhecimento por toda a sua dedicação, empenho e motivação. Os seus conselhos e encorajamentos foram uma constante: a sua orientação foi fundamental para a concretização deste trabalho.

Ao Ministério da Educação, ter aceite e autorizado o pedido de consulta e acesso aos diferentes arquivos das Escolas Secundárias Nacionais, às Provas de Exames Nacionais realizadas por estudantes do Ensino Secundário, com a respetiva garantia do anonimato de todos os envolvidos.

Aos Senhores Diretores José António Sousa, Pedro Folgado e Manuela Dias, respetivamente das Escolas Secundárias D. Dinis, Lisboa; Damião de Gois, Alenquer; e Portela, Sacavém, por terem permitido a consulta e fotocópia das Provas dos Exames Nacionais efetuados pelos seus estudantes, em arquivo, respeitando o anonimato de todos os envolvidos, tanto dos Professores corretores como dos estudantes.

À Professora Doutora Florinda Matos pela cedência dos textos de resposta a questões de gestão - que se revelaram extremamente úteis.

Aos meus colegas e amigos o interesse com que sempre me acompanharam, nomeadamente às irmãs Hermínia Ribeiro, pelo primeiro contato estabelecido com a Escola Secundária da Portela, Sacavém, e Isabel Ribeiro, pela inestimável colaboração que prestou na correção de Provas de Exame Nacionais de Português assim como no primeiro contacto estabelecido com a Escola Secundária Damião de Gois, Alenquer.

À Rute Antunes pela paciência e por todo o trabalho que teve na transcrição dos textos.

Finalmente, à minha família a satisfação com que acompanharam o meu entusiasmo pelo trabalho desta Tese.

CONTRIBUIÇÕES PARA A FORMULAÇÃO DE UMA METODOLOGIA DE ENSINO E AVALIAÇÃO BASEADA NA ANÁLISE ESTATÍSTICA DE TEXTOS EM PORTUGUÊS

por

LUÍS ANTÓNIO PISSARRA DE MATOS AGONIA PEREIRA

RESUMO

PALAVRAS-CHAVE: Língua Portuguesa; Avaliação automática; Análise da Semântica Latente; Biplots.

Esta investigação procurou identificar e avaliar problemas associados ao desenvolvimento de instrumentos de ensino / aprendizagem, a serem utilizados por professores e estudantes, com base na análise estatística de textos, com especial relevo dado à monitorização, quer pelos professores quer pelos estudantes, dos processos de aquisição de conhecimentos e de conteúdos pedagógicos nos quais o português ocupa papel central.

Apesar de existirem importantes projetos de investigação ligados à língua portuguesa, por exemplo a investigação ligada à atividade do Instituto de Linguística Teórica e Computacional, ILTEC e do NLX-Group (*Natural Language and Speech Group* Da Faculdade de Ciências, Universidade de Lisboa) constata-se que na área do objeto de estudo desta investigação, poucos trabalhos têm sido publicados.

No início desta investigação demos particular atenção aos trabalhos do grupo de Landauer e sua equipa, na Universidade do Colorado, EUA, que, tendo por base a metodologia da Análise da Semântica Latente (ASL), visam a avaliação automática de conhecimentos a partir de textos escritos em resposta a questões abertas. Apesar da ASL ser uma metodologia interessante, constatámos que a sua base computacional (a decomposição em valores e vetores singulares das matrizes de frequências de palavras em textos) coincide com a dos biplots de Gabriel e Galindo, pelo que usámos os biplots como instrumentos de visualização ao construir um sistema protótipo que, de uma forma simples e flexível, permite, em contexto de sala de aula, e tendo por base um processo de avaliação contínua, a avaliação dos estudantes por parte do Professor.

Como principais resultados desta investigação, citam-se o desenvolvimento de um protótipo experimental que incorpora um sistema de mineração de textos de resposta a questões abertas em apoio da atividade dos professores na respetiva atividade letiva, o treino e teste de sistemas automáticos de classificação de textos (SAAT) assentes no conceito de ASL, dotados de uma interface gráfica assente nos biplots de Galindo e ainda os resultados das análises dos dados reais recolhidos (textos de manuais usados no ensino e textos de respostas a questões abertas incluídas em exames nacionais e em testes de avaliação contínua num politécnico) – realizadas pelo protótipo – que permitiram concluir, experimentalmente, pela razoabilidade de criar instrumentos de software com as características desejadas, embora separando os problemas de classificação automática com questões incluídas em exames sumativos finais das questões ligadas à formação contínua.

ABSTRACT

KEYWORDS: Portuguese Language; automatic assessment; Latent Semantic Analysis; Biplots.

This research aims to identify and evaluate problems associated with the development of teaching and learning instruments, to be used by teachers and students, based on statistical analysis of texts, with special emphasis on monitoring, both by teachers and students, the acquisition of knowledge and pedagogical content, in which the Portuguese occupies the central role.

Although there are important research projects related to the Portuguese language, for example the research linked to the activity of the Institute of Theoretical and Computational Linguistics, ILTEC, and NXL-Group (Natural Language and Speech Group four Faculdade de Ciências, Lisbon University, we observed that on the object of study of this research, few works have been published.

At the beginning of this research we have given particular attention to research work of Landauer and its associates at the University of Colorado, USA that based on the methodology of Latent Semantic Analysis (LSA), the automatic assessment of knowledge. Despite LSA being an interesting methodology, we found that LSA and the concept of biplot from Gabriel and Galindo are based on the same computational techniques: singular value decomposition of frequency matrices. That is the reason why we used biplots as visualization instrument in the construction of a system that, in a simple and flexible way, allows teachers, in the context of the classroom and of a continuous assessment process, to evaluate students.

As the main results of this investigation, we mention the development of an experimental prototype system that incorporates text mining of texts resulting from response to open questions, in support to teaching activities, the training and test of systems for automatic text classification based on the concept of LSA, equipped with a graphical user interface based on Galindo biplots and, also, the results of the analysis of real data collected (text of books used in teaching and text answers to open questions included in national examinations and texts resulting from continuous assessment in a polytechnic) - performed by the prototype - which allowed us to conclude, experimentally, for the reasonableness of creating software instruments with the intended characteristics, though separating sharply the problems of automatic classification of high stakes tests from issues related to continuous assessment.

ÍNDICE

LISTA DE ABREVIATURAS

AA	Aprendizagem Auto-regulada
ACCESS; Microsoft Access	Sistema de Gestão de Base de Dados da Microsoft, inserido no pacote Microsoft Office Professional
AES	Automatic Essay Scoring
ALGLIB	Numerical Analysis library (*http://www.alglib.net/*)
APA	American Psychological Association
ASL	Análise da Semântica Latente
BDT	Base de Dados Textual
BETSY	Bayesian Essay Teste Scoring System
CRI	Constructed Response Item
ECD	Evidence Centered Design
EDM	Educational Data Mining
ES	Espaço(s) Semântico(s)
ETS	Educational Testing Service
EXCEL; Microsoft Excel	Programa de Folha de Cálculo da Microsoft, incluído no pacote Microsoft Office
GMAC	Graduate Management Admission Council
GMAT	Graduate Management Admission Test
GRE	Graduate Record Examination
IEA	Intelligente Essay Assessor

ILTEC	Instituto de Linguística Teórica e Computacional
IRT	Item Response Theory
LAD	Language Acquisition Device
LSA	Latent Semantic Analysis
MDE	Mineração de Dados Educacionais
MDS	Multi Dimensional Scaling
NLX-Group	Natural Language and Speech Group
NLP	Natural Language Processing
OCR	Optical Character Recognition
PAET	Programa de Análise Estatística de Textos
PEG	Project Essay Granding
SAAT	Sistema Automático de Avaliação de Textos
SDL	Software Development Lohninger
SPSS	Statistical Package for the Social Sciences
SVD	Singular Value Decomposition
TCT	Teoria Clássica dos Testes
TOEFL	Test Of English as a Foreign Language
TRI	Teoria da Resposta ao Item
WAD	Word Acquisition Device

Introdução

Neste trabalho investiga-se a possibilidade da utilização das técnicas de análise estatística de textos (*text-mining*) em tarefas relacionadas com a avaliação, a monitorização do processo de aquisição de conhecimentos e no apoio ao trabalho do professor e do estudante.

O tema que mais nos motiva é, de um modo geral, a problemática da avaliação e o da construção de instrumentos de apoio ao professor nas suas tarefas de avaliação. De um modo mais específico, a motivação imediata desta tese teve a ver com a problemática da avaliação de respostas a questões abertas – sob a forma de textos elaborados pelos estudantes e as questões de *feedback* a estudantes em sistema de auto ensino. Trata-se de analisar e, eventualmente, construir instrumentos capazes de lidar com grandes quantidades de texto envolvidas nos manuais de estudo e nas respostas dos estudantes a questões que exigem como resposta destes a elaboração de textos.

É um facto sobejamente conhecido por todos os professores que, de um modo geral, estes são deixados a si próprios – isto é, não têm, em geral, qualquer apoio metodológico ou outro – quando têm de elaborar testes de avaliação de conhecimentos e competências nos diversos momentos do ciclo letivo anual e, depois de aplicados esses testes, avaliar as respostas e publicar os resultados dessa avaliação. É o que sucede com particular incidência, naquelas disciplinas em que o texto escrito em linguagem natural tem importância preponderante e em que a correção e classificação das respostas a questões abertas consomem um volume enorme de trabalho sem que isso garanta a homogeneidade de critério e ausência de ambiguidade na avaliação dessas respostas.

Numa altura em que quase todos os envolvidos no processo de ensino – aprendizagem têm acesso a processamento eletrónico de textos, em que as plataformas de ensino, nomeadamente em *software Open Source* como o *Moodle*, proliferam a todos os níveis e em que vão estando disponíveis poderosas metodologias de tratamento estatístico de textos, justifica-se investigar a possibilidade de identificar, avaliar ou mesmo construir alguns instrumentos que apoiem os professores nas tarefas de avaliação de respostas a questões abertas e na preparação de testes em que os textos tenham papel predominante.

Esses mesmos instrumentos estatísticos podem ter um papel fundamental no processo de monitorização pelo Professor da aquisição de conhecimentos pelos estudantes como dos próprios estudantes envolvidos em processos autónomos de aprendizagem, possibilitando o fornecimento, em tempo oportuno, de *feedback* baseado em informação objetiva que os motive e permita o autocontrolo e a auto-motivação para aprender mais e melhor.

Nesta perspetiva, procura-se neste trabalho avaliar – do ponto de vista do seu potencial para apoio do trabalho dos professores e estudantes – certas técnicas e metodologias gerais de análise estatística de textos – como a Classificação Automática, o *Multi Dimensional Scaling* (MDS) e técnicas mais especializadas como a *Latent Semantic Analisis* (LSA). Esta avaliação assenta na elaboração de protótipos que permitam a sua utilização experimental.

Que instrumentos podem ser desenvolvidos e com que metodologias para que, sem envolver mais do que os recursos habituais da parte dos intervenientes (informação textual, processador de texto, folha de cálculo, entre outros), estes possam utilizá-los no seu dia-a-dia?

A abordagem básica é a de procurar dotar os intervenientes (professores e estudantes) de instrumentos que gerem informação útil ao processo de ensino – aprendizagem, mais do que de programas de avaliação automática de conhecimentos, sendo os objetivos gerais desta tese os seguintes:

1. Identificar e avaliar técnicas de análise estatística de textos que possam ser usados na construção de instrumentos de apoio a professores e estudantes como instrumentos de monitorização dos processos de aquisição de conhecimentos e avaliação;
2. Avaliar a possibilidade de usar a técnica de LSA em tarefas de apoio à avaliação de conhecimentos com base em textos de resposta a questões abertas;
3. Elaboração de um protótipo que permita experimentar técnicas de análise estatística de textos na perspetiva do apoio ao professor e a estudantes identificando problemas a ter em conta em futuros desenvolvimentos.

As questões para as quais procurámos respostas ao longo desta investigação são as seguintes:

Questão I – Qual o estado da investigação relativo à avaliação automática de conhecimentos com base na análise de textos de resposta a questões abertas?

Questão II – Quais os problemas que se põem quando se procura construir instrumentos baseados em análise estatística de textos e outras técnicas de mineração de textos que permitam apoiar os professores no seu esforço para reduzir os subjetivismos na análise dessas respostas dos estudantes, permitindo, sem aumento brutal da sobrecarga de trabalho, generalizar o uso de questões de resposta aberta?

Questão III – Que impacto podem as técnicas de mineração de textos ter no modo de abordar o processo ensino-aprendizagem e qual a possibilidade, a curto prazo, de construir sistemas baratos de apoio a professores e estudantes usando essas tecnologias?

Este trabalho é uma tentativa de dar resposta a cada uma destas questões. Especificamente:

No **Capítulo I** procura-se responder à Questão I. Neste capítulo visa-se fazer o ponto de situação relativa à investigação da questão da avaliação automática de conhecimentos com base na análise de textos de resposta a questões abertas. Surpreendentemente, constata-se que em Portugal não se deteta na literatura ou nos organismos oficiais iniciativas nesta direção, quando a investigação parece apontar para o facto de, neste momento, ser difícil avaliar as diferenças entre os comportamentos estatísticos das classificações atribuídos pelos sistemas automáticos mais importantes e as atribuídas por professores. Sendo essa questão considerada resolvida, nos Estados Unidos, o próximo passo parece ser o de realizar importantes avaliações sumativas a nível nacional usando sistemas que têm vindo a ser aperfeiçoados ao longo dos últimos 20 anos, sob a liderança de organizações como o *Educational Testing Service* (ETS).

Nos **Capítulos I** e **II** procura-se dar resposta à Questão II.

Especificamente, no **Capítulo II** procura-se estabelecer a metodologia a usar na construção de um modelo que procure responder a esta questão, dando-se especial atenção

às características de um sistema para realizar avaliações – sobretudo formativas – que possa apoiar a avaliação de textos produzidos por estudantes em resposta a questões abertas, usando principalmente a técnica da Análise da Semântica Latente (ASL) – designação de agora em diante usada para designar a LSA. Tendo-se reconhecido que subjacente a esta metodologia está a mesma técnica computacional em que se baseiam representações gráficas de dados multivariados – a *Singular Value Decomposition* (SVD) – a estrutura do modelo a desenvolver explora este facto e prevê, explicitamente, a produção de gráficos (biplots, árvores de classificação) que possam ser usados na comparação dos textos produzidos pelos diversos estudantes de uma turma com melhoria dos critérios de homogeneidade e redução de subjetividade do Professor.

É ainda neste capítulo – central de toda a tese – que se apresenta o procedimento usado na estruturação de uma amostra que permite validar com dados reais as principais metodologias.

Como condição fundamental para a possível utilização prática das técnicas e instrumentos como os que neste trabalho se preconizam, assume-se que os estudantes usam, na produção dos respetivos textos, meios informáticos (computadores e processadores de texto). A tecnologia do reconhecimento ótico de caracteres, OCR, não atingiu ainda o grau de perfeição que permita encarar a utilização de textos manuscritos dos estudantes neste tipo de processamentos.

Infelizmente residiu aqui uma das principais dificuldades deste projeto: o facto de o Ministério da Educação não guardar de um modo centralizado as provas classificadas e o facto de todos os textos usados serem textos manuscritos que tiveram previamente de ser digitalizados sem qualquer apoio económico para a realização dessa tarefa.

Ainda no contexto da resposta à Questão II, no **Capítulo III** descreve-se a estrutura de uma base de dados a usar para registar os textos das respostas dos estudantes a questões de resposta aberta, descrevendo-se os instrumentos usados na recolha e o programa usado para o efeito. São ainda apresentados os resultados das análises realizadas e a respetiva interpretação. Uma das conclusões a que se chegou foi a de que, embora existam correlações significativas entre as classificações atribuídas pelos professores e as atribuídas pelo sistema implementado, essas correlações não atingem os níveis relatados em estudos publicados, por exemplo, pelo ETS. Atribui-se isto - para lá do facto de os dados usados não serem suficientemente volumosos para esse efeito – ao não ter sido possível dotar o

protótipo construído de todos os componentes necessários à cobertura de questões linguísticas.

Nos **Capítulos II** e **III** procura-se responder à Questão III, tendo-se concluído, tanto através do estudo da literatura como através da experimentação possibilitada pelo protótipo de sistema por nós construído, que as técnicas de mineração de textos (*data mining*), em particular, e de exploração de dados, em geral (através do conceito *Educational Data Mining* (EDM)) vão ter, como suspeitávamos, um impacto enorme na completa revisão, em curso, do atual paradigma do processo ensino – aprendizagem. Em particular, dado o uso combinado de sensores para aspetos físicos (biometria) com a exploração estatística sistemática dos dados da interação estudantes-sistemas, combinados com técnicas psicométricas (Teoria da Resposta ao Item (TRI), em particular), estão criadas as condições que vão permitir, a breve prazo, criar sistemas baseados no conceito designado por *Evidence Centered Design* (ECD).

O protótipo por nós criado (e o próprio processo de criação) para apoiar as investigações descritas – cujo manual do utilizador constitui o Anexo A desta tese – permitiu-nos uma perceção prática do tipo de problemas a resolver em futuras investigações.

Capítulo I: Enquadramento Teórico / Estado de Arte.

1.1. Introdução.

Neste capítulo procura-se sintetizar o significado e evolução histórica de alguns dos conceitos fundamentais relacionados com a nossa própria investigação – nomeadamente a avaliação de conhecimentos no âmbito da atividade letiva do professor e do estudante, procurando-se caracterizar o contexto das atuais potencialidades criadas pela revolução em curso da ciência dos dados e das disciplinas com ela relacionadas: *data mining*, *text mining*, *analytics*, de agora em diante designadas em português por mineração de dados, mineração de textos e analítica.

Sendo um dos objetivos deste trabalho investigar a possibilidade de criar instrumentos informáticos que facilitem o trabalho dos professores e dos estudantes, principais atores envolvidos no processo de ensino-aprendizagem, procura-se neste capítulo caraterizar algumas dessas possibilidades enquadrando-as nos contextos históricos e teóricos adequados.

Face à disponibilidade objetiva de instrumentos informáticos (*hardware* e *software*) a atividade de avaliação, a que dedicamos a maior parte do presente capítulo, está a sofrer profundas transformações alimentadas por novos instrumentos teóricos e metodológicos.

Estas transformações implicam a reconceptualização de alguns problemas de ensino-aprendizagem – face a novas possibilidades de análise de dados como as fornecidas pela mineração de textos e de novos instrumentos metodológicos como sejam sistemas de mineração de dados e aprendizagem automática (*learning machines*).

A título de exemplo, a Teoria da Resposta ao Item (TRI) modelo teórico em que assenta a elaboração de testes com caraterísticas psicométricas controladas (fiabilidade, validade e grau de dificuldade) é atualmente objeto de intensa atividade de investigação desencadeada pelos desafios resultantes das possibilidades criados por essas novas tecnologias (Rauch & Marting, 2010).

O conceito de EDM (Mineração de Dados Educacionais) é a manifestação no domínio das Ciências da Educação de uma nova realidade resultante da disponibilidade de grandes massas de dados em todos os domínios (em particular na atividade da Educação),

que desafiam as capacidades de armazenamento e análise dos computadores disponíveis e dos métodos e metodologias existentes.

Até à década de 70 do século passado, a informação era restrita a certas pessoas, em determinados lugares e num formato específico. A questão era então a escassez de informação e a dificuldade do respetivo acesso. Hoje a informação está acessível a qualquer pessoa que tenha um computador com acesso à Internet.

Face ao colossal volume de informação disponível (Manyika, Chui, Brown, Bughin, Dobbs, Roxburgh, & Byers, 2011) torna-se absolutamente essencial usar tecnologia que consiga obter informação correta, isto é, contextualizada, no espaço de tempo certo. Ranadivé e Maney (2011) conseguem descrever perfeitamente esta ideia no seu livro "*The Two-Second Advantage*", mostrando que uma vantagem de apenas dois segundos pode fazer toda a diferença no mundo atual em que vivemos, graças aos avanços da ciência nos campos da pesquisa neurológica e das Tecnologias da Informação.

Uma das palavras-chave do atual estado do mundo – *big data*, ver por exemplo Executive Office of the President (2010) – tem a ver com a superabundância de dados existentes e com a tendência exponencial para o seu aumento: organizações (nomeadamente as ligadas ao ensino), empresas, famílias e indivíduos acumulam, em suportes magnéticos, tanto os dados das observações científicas como os dados do seu trabalho normal e das suas vidas, num movimento que tende a refletir nos computadores o estado do mundo e das sociedades em cada instante (Manyika et al., 2011). Este facto tem como consequência que não só praticamente todas as atividades humanas estejam dependentes de dados existentes em suporte magnético como, do ponto de vista da investigação científica, uma parte cada vez maior dessa investigação passa a ser realizada sobre a representação do mundo já existente nos computadores e em contínua atualização ao ritmo da vida. Veja-se o exemplo do *Google Earth*[1].

Como consequência, têm assumido importância crescente as atividades ligadas à exploração científica desses dados em benefício de quase todas as disciplinas, desde as Ciências Humanas, à Física e à Biologia.

Em particular, nas Ciências da Educação, é de origem recente o conceito de Mineração de Dados Educacionais - Exploração de Grandes Massas de Dados em Ciências

[1] *Google Earth* realiza a cartografia do planeta, juntando imagens obtidas de várias fonte sobre um globo 3D.

da Educação (Romero, Ventura, Pechenizkiy, & Baker, 2011). Trata-se de utilizar também, no domínio das Ciências da Educação, conceitos que têm feito o seu caminho, com resultados assinaláveis, noutras áreas no sentido de desenvolver técnicas que permitam explorar a estrutura específica dos dados existentes sobre suportes magnéticos nos diversos domínios e vertentes das Ciências da Educação.

Quais as consequências desse facto nas técnicas de ensino, no processo de ensino-aprendizagem, no funcionamento das escolas, na estruturação dos programas e seus conteúdos, nas relações entre professores e estudantes, nos métodos de avaliação?

Não se trata apenas de superabundância de dados, mas também do facto do acesso aos mesmos ser altamente facilitado através da Internet – no sentido de que está, supostamente, disponível para todos e imediatamente acessível. Professores e estudantes têm acesso, em princípio, à mesma informação, ao mesmo tempo, no mesmo local e nas mesmas condições. Quais as consequências deste facto para as relações professor-estudante no processo de aprendizagem e para a problemática da avaliação de conhecimentos?

Quando se diz "acessível" não estamos apenas a dar conta da possibilidade de consultar, estudar, ler esses dados, mas da possibilidade de os processar usando software sob o controlo do utilizador.

Entre outras consequências, o que é ensinado e o modo como é ensinado pode ficar sob um escrutínio potencialmente muito mais rigoroso e intenso por parte da sociedade – instituições, pais, estudantes, escola, tribunais; o professor deixa de ser a fonte principal do conhecimento e os seus atos e decisões avaliados objetivamente.

Os intervenientes no processo de ensino-aprendizagem continuam a ser os mesmos: sociedade, escola, professor, estudante; contudo a realidade descrita tem profundas consequências no desempenho dos respetivos papéis e seu conteúdo.

1.2. Importância, Formas de Avaliação e Monitorização de Conhecimentos.

A importância da avaliação e certificação de conhecimentos e competências bem como os seus objetivos e condições de realização garante que a respetiva validade e fiabilidade constituíam já uma preocupação importante da China de há 3000 anos, como se pode ver em Wainer, H., 1990. Este autor citando a obra do historiador chinês Ssu-yu Têng

(1943), afirma que já em 1115 a. C. existia na China um sistema de exames a que se submetia os funcionários públicos antes de serem graduados.

Na obra de Têng (1943) pode, inclusivamente, ler-se que a primeira referência na Europa ao Exame Imperial deve-se ao missionário português Gaspar da Cruz (*Tratado em que se contam muito por extenso as Cousas da China com suas particularidades e assim do Reino de Ormuz*, 1569). Este Exame Imperial veio influenciar decisivamente o sistema de avaliação dos funcionários públicos ingleses.

Se é verdade que os estudantes têm, com a tecnologia atual, potencialmente, acesso às mesmas fontes, conteúdos e instrumentos de análise que os professores, tal não significa que a aquisição de conhecimentos e competências por parte dos estudantes seja, por esse facto, um dado adquirido.

Pelo contrário, a necessidade de avaliação de conhecimentos e certificação de competências têm uma importância crescente nas sociedades complexas em que as questões de qualidade – impostas pela legislação e pela necessidade de fiabilidade e segurança têm importância cada vez maior.

Por definição, o ensino é um processo que tenta promover o desenvolvimento dos estudantes, sendo a aprendizagem dos estudantes o objetivo dos professores e das escolas. A sociedade tende a atribuir a tarefa de avaliar o desempenho, o desenvolvimento escolar e o potencial dos estudantes em grande parte às escolas e aos professores, responsabilizando-os pela aprendizagem daqueles. As escolas acabam por ser sistemas humanos influenciados quer pelas pessoas que nelas trabalham, de forma interdependente, como pela comunidade e sociedade onde estão integradas.

A sociedade não pode correr o risco de atribuir responsabilidades a incompetentes nem de confiar a pessoas impreparadas o desempenho de funções técnicas e sociais importantes.

Diversos estudos concluíram que o comportamento dos professores tem uma importante influência sobre a vontade que os estudantes demonstram em cooperar e persistir nas tarefas de aprendizagem (Dolezal, Welsh, Pressley, & Vincent, 2003; Pintrich & Schunk, 2002). O papel do professor na orientação e no sucesso dos estudantes é muito importante. O modo como o estudante é motivado na sala de aula para que atinja o comportamento e a aprendizagem desejada são fatores que não podem ser descuidados. Ou seja, a atribuição de boas notas, elogios e privilégios nas aulas são exemplos de incentivos

e recompensas que os professores podem utilizar de modo a criarem, nos estudantes, hábitos de estudo e fomentarem comportamentos desejados. Da mesma forma, as más notas, os castigos e a perda de privilégios são usados como forma de desencorajar ações e comportamentos indesejados. O professor aplica a teoria do reforço, com recurso à utilização de reforços positivos e negativos. O professor deverá ser competente e eficaz na forma como identifica os comportamentos desejados e os reforços que são adequados e utiliza estes reforços de forma a fortalecer e encorajar os comportamentos desejados.

Do mesmo modo, o Professor, no decurso do ano letivo necessita de controlar o processo de aquisição de conhecimentos e competências dos seus estudantes no sentido de realizar as correções necessárias e fornecer aos estudantes elementos de *feedback* que permitem a autocorreção ou funcionem como elementos de motivação.

É consensualmente aceite que os índices da motivação nos estudantes são alteráveis, sendo um dos papeis mais importantes que o professor tem no contexto do ensino-aprendizagem, devendo estar atento àqueles procurando potenciá-los.

Efetivamente, as escolhas dos professores, em relação às estratégias de motivação adotadas influenciam a manutenção do envolvimento dos estudantes na aprendizagem. Há bastante tempo atrás, John Dewey (1916), observou que as crianças aprendiam quando participavam em contextos sociais. Também Jerome Bruner (1996) e Vygotsky (1978, 1994) defenderam que as pessoas estabelecem significados a partir das relações e da participação em cada cultura. De facto, a vida numa sala de aula pode, de alguma forma, espelhar a vida fora da escola. Assim, também os grupos e as comunidades de aprendizagem são representativas das necessidades individuais e de grupo tornando-se um importante aspeto do ensino, devendo os professores atender e corresponder às múltiplas caraterísticas dos estudantes para a sua criação e estabelecimento, uma vez que se sabe que estas comunidades também podem limitar a iniciativa de um indivíduo e / ou promover normas que limitem a criatividade e a aprendizagem.

Cabe ao professor atribuir tarefas aos estudantes que sejam para eles interessantes, de modo a que estes continuem nas suas tarefas de aprendizagem, havendo, contudo, estudantes mais persistentes que outros. Também é aceite que as aprendizagens são mais significativas em ambientes que se caraterizam pelo respeito mútuo.

Os próprios estudantes envolvidos em processos de ensino-aprendizagem, com ou sem professores, necessitam com frequência de aferir o respetivo progresso antes de

avançarem para temas mais complexos e dependentes de conhecimento anterior. Efetivamente, a utilização da Mineração de Dados Educacionais (EDM) tem contribuído muito para o desenvolvimento das Ciências da Educação, nomeadamente na análise das trajetórias dos estudantes à medida que se desenvolvem ao longo de um período de tempo. Os estudos têm incidido sobre períodos de tempo pequenos. Amershi e Conati (2009) estudaram as relações entre comportamentos em Ambientes de Aprendizagem (A. A.) exploratória e os resultados da aprendizagem. Arroyo e Woolf (2005) examinaram as relações entre um conjunto de comportamentos AA e as motivações num sistema inteligente de tutoria. No entanto, existem estudos nos quais os períodos de tempo em questão são maiores, como o de Arnold (2010), que teve a duração de um semestre. Este assunto está intimamente ligado à AA, na qual um estudante procura, por um lado, otimizar a forma de chegar ao seu objetivo e, por outro, otimizar a forma como essa forma é construída. With Hadwin e Winne (Winne & Hadwin, 1998; Winne, 2001, 2011) propuseram um modelo de 4 fases sobre A. A.

Se é certo que a Avaliação sempre tem levantado, e continua a levantar, muitas reservas por parte de muitos educadores – ver, por exemplo, Miller, Linn, e Gronlund (2009) – torna-se evidente do exame da bibliografia consultada, mesmo a mais recente, que a avaliação não só não pode ser dispensada como tende a tornar-se cada vez mais presente e decisiva.

No contexto atrás descrito, o papel do Professor não fica enfraquecido, sendo ainda mais importante e difícil – uma vez que não se trata agora, apenas, de transmitir conhecimentos cabendo-lhe ensinar algo muito mais importante: o significado das coisas e da informação a que todos têm acesso, as atitudes perante o conhecimento; a semântica, mais do que o conhecimento formal. Cabe-lhe também um papel ainda mais fundamental e difícil na avaliação dos conhecimentos adquiridos e da certificação de competências num contexto em que os processos de avaliação são objeto de sistemática contestação por parte de grupos sociais importantes.

A avaliação dos estudantes consome grande parte do tempo dos professores. Shaefer (1991) refere que os professores passavam até 10% do seu tempo com assuntos relacionados com a classificação e a avaliação, e Stiggins (2004) refere que este tempo pode representar cerca de um terço.

Assim, é neste quadro, que cabe ao Professor realizar a "medição". Ou seja, fazer a recolha e síntese da informação sobre os seus estudantes e as suas salas de aula. É um processo contínuo. A recolha desta informação pode ser efetuada de maneira formal, através de testes, relatórios, apresentações, entre outros; e de forma informal, por observação, conversação, etc. Pode fazer parte deste processo informações sobre a sala de aula assim como a forma de ensinar do professor.

Outro dos papéis importantes que o Professor tem a seu cargo é a "avaliação". Realizar juízos, dar classificações – atribuição de um valor à informação obtida – e decidir sobre o mérito. As avaliações separam-se em avaliações formativas e sumativas. As formativas são recolhidas antes ou durante a formação destinando-se a informar os professores sobre os conhecimentos e as competências prévias dos estudantes. Estas informações permitem aos professores fazer juízos sobre como agrupar os estudantes, como planificar unidades e aulas assim como estratégias educativas. As sumativas constituem a intenção da utilização da informação sobre os estudantes ou currículos após a realização de uma série de atividades educativas. Têm como objetivo resumir o desempenho de um estudante, ou grupo de estudantes, ou professor mediante um conjunto de objetivos de aprendizagem. Realizam-se juízos sobre os resultados, servindo estes para a determinação das notas e para explicar as mesmas aos estudantes e encarregados de educação.

Tipo de avaliação	Quando é obtida	Tipo de informação obtida	Como é utilizada a informação
Formativa	Antes ou durante a formação	Informação sobre os conhecimentos prévios dos estudantes e / ou dos processos formativos	Ajudar na tomada de decisões do professor
Sumativa	Após a formação	Informação sobre o desempenho do estudante e /ou do professor	Ajudar na elaboração de juízos sobre os resultados obtidos pelos estudantes ou professores

Tabela 1.2.1. Avaliação formativa e sumativa (Arends, 2008).

A informação que os professores e outros profissionais utilizam, para a tomada de decisões deve ser de elevada qualidade (fiável e válida). Os resultados obtidos num teste são fiáveis ou precisos quando são consistentes. No entanto, existem fatores que podem introduzir erros (inconsistência e pouca fiabilidade) nomeadamente as expectativas dos estudantes face às classificações a serem-lhe atribuídas, erros de classificação dos professores e a disposição do estudante no momento do teste. Outra caraterística que um teste deve ter é a de ser justo, ou seja, deve oferecer a todos os estudantes a mesma oportunidade de terem uma boa nota, sem descriminação pela raça, etnia ou género.

1.3. Teoria Clássica da Avaliação e o TRI. O Modelo de RASCH.

Os sistemas atuais de avaliação assentam, essencialmente, em conceitos psicométricos e resultam de uma longa evolução que tem início em 1908 com a escala de Binet e Símon, baseada em 30 problemas ordenados por dificuldade crescente, aplicável a crianças entre os 3 e os 11 anos, escala esta que visava expressar a idade mental ou nível mental.

Na evolução das teorias quantitativas de avaliação assumem especial relevância a Teoria Clássica dos Testes (TCT) e a Teoria da Resposta ao Item (TRI) – Baker (1992), Keeves (1988), Arias (1996), Longford (1995), Muñiz (1996) e Rauch e Harting (2010).

Considerando uma população de P pessoas sobre as quais são observados os valores X_i (i= 1, 2, …, P) de uma variável aleatória – as pontuações obtidas num certo teste, por exemplo – nesta teoria assume-se ainda que, para lá da pontuação observada X_i, existe uma pontuação verdadeira V_i que não pode ser observada (diretamente), donde resulta que

$X_i = V_i + E_i$,

Isto é, o valor observado X_i é o valor verdadeiro V_i mais o erro de observação

$E_i = X_i - V_i$,

assumindo-se que o valor médio do erro é nulo assim como são nulas as correlações dos erros entre si e dos erros com os valores verdadeiros: isto é, não há relação (linear) entre os erros entre si nem entre os erros e os valores verdadeiros.

Este modelo, como mostraram autores como Keeves (1988) e Linden (1997), sofre de consideráveis limitações tanto do ponto de vista da adequabilidade à realidade de alguns

dos seus pressupostos (validade) como do ponto de vista da qualidade dos resultados obtidos. Sobre estas dificuldades pode consultar-se Arias (1996) e Muñiz (1996).

Uma das dificuldades na utilização da TCT reside na não consideração de um aspeto crucial do processo de medição aplicável a fenómenos psicológicos: o da interação entre o respondente e a pergunta, mais especificamente entre o grau de competência do respondente e o grau de dificuldade de um item.

Basicamente, a TRI – que abrange variantes expressas por vários modelos – assume a existência de certas variáveis latentes (não observáveis diretamente) ou traços latentes (competências) que permitem predizer ou explicar o comportamento do estudante ao responder a um item. Uma vez que estas variáveis ou traços latentes, por exemplo, a inteligência do estudante, não são observáveis diretamente, tudo o que se pode fazer é tentar estimar o seu valor através de variáveis observáveis ou manifestas que são função – dependem – desses traços latentes não observáveis (Baker, 1992; Linden, 1997).

Estes traços latentes – admitindo que são variáveis quantitativas – podem ser unidimensionais ou multidimensionais/multivariadas. Isto é, são modeladas por vetores $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ em que k é o número de dimensões da variável latente ou traço latente.

Para cada traço latente - θi –assume-se que existe uma relação crescente (função caraterística) entre o valor desse traço e a probabilidade que o respondente obtenha êxito ao responder ao item. Sobre este tema podem ver-se os citados Arias (1996), Baker (1992), Muñiz (1996) e Linden (1997).

Em Baker (1992) e em Linden (1997) pode ver-se que uma variante específica da TRI é o modelo de RASCH. Se β_n for a aptidão ou competência – valor do traço a ser medido da pessoa número n – e se δ_i for a dificuldade do item (dicotómico) i, então a probabilidade de que esse respondente acerte ao responder a este item é dado pela expressão

$$P\left(X_{n_i}=1\right)=\frac{e^{(\beta_n-\delta_i)}}{1+e^{(\beta_n-\delta_i)}}\,.$$

Em síntese: quanto maior o afastamento entre a competência de um estudante e a dificuldade de um item (aqui expressa por β_n - δ_i), menor é a probabilidade de que um estudante com essa competência acerte na resposta a esse item. E o inverso. Em particular, se $\beta_n = \delta_i$, a probabilidade de êxito é ½; por outras palavras, o resultado é incerto: por vezes os respondentes que respondem a itens ao seu nível de competência acertam na resposta, mas pode suceder, com igual probabilidade, o contrário. Se a competência do

respondente é maior que a dificuldade do item, então a probabilidade de acertar é superior a 0.5. Reciprocamente, se a competência do respondente for inferior à dificuldade do item, a probabilidade de acertar é inferior a 0.5.

Os valores β_n e δ_i não podem ser observados diretamente, pelo que têm de ser estimados a partir de observações resultantes da aplicação de testes. Os valores de β_n aparecem refletidos nas percentagens observadas de acerto em testes com certa dificuldade e os valores de δ_n aparecem refletidos – para estudantes de certa competência fixa – na percentagem de acerto na resposta ao teste.

O modelo de RASCH está em aplicação crescente tanto em testes constituídos por questões (itens) de resposta múltipla como em testes formados por itens de resposta expressa em escalas ordinais, como por exemplo escalas de Likert.

O modelo TRI tem sido um dos fatores explicativos do uso generalizado dos testes de resposta múltipla (binária ou não) dada a simplificação que os dados provenientes deste tipo de itens introduzem nos métodos de estimação e a simplificação da teoria estatística de estimação permitida por este tipo de testes.

A outra razão está na facilidade com que as respostas a testes desse tipo são classificadas e, portanto, no baixo custo da classificação, tanto para a organização como para o professor.

Acresce ainda que a aplicação de itens desse tipo tem permitido a construção de bases de dados com os resultados da respetiva aplicação a grandes conjuntos de respondentes – possibilitando a estimação das chamadas curvas caraterísticas e de outros parâmetros que estão na base da construção automática de testes cujas caraterísticas psicométricas são perspetivadas e otimizadas antes do momento da aplicação (Wainer, 1990).

O aspeto negativo deste estado de coisas tem sido a anterior tendência para não usar ou usar pouco questões de resposta aberta (em que os estudantes elaboram a resposta) mais adequados do que os testes de resposta múltipla na avaliação de certos conhecimentos e competências e latentes.

Recentemente têm-se intensificado – ver, por exemplo, Liu, Lee, e Linn (2011) - os esforços visando a utilização de itens de resposta aberta nos testes de avaliação. Embora desde há muito tempo se tenha reconhecido que os itens de resposta aberta – obrigando a que os respondentes elaborem a resposta na sua língua natural, por exemplo – têm grandes vantagens não só do ponto de vista da aquisição de conhecimentos como pelo que revelam

no momento da avaliação, diversas dificuldades têm impedido a sua utilização mais frequente em testes de avaliação de conhecimentos.

Uma das razões prende-se com o tempo que a respetiva classificação exige dos professores. A outra está ligada à dificuldade em evitar a intervenção de fatores subjetivos e outros enviesamentos nas tarefas de avaliação. Existe, atualmente, uma investigação muito ativa nas questões de adaptação da TRI aos problemas postos pela investigação das respostas a questões abertas (Rauch & Hartig, 2010; Liu, Lee, & Linn, 2011; Oliveri & Davier, 2011; Newman & Vermunt, 2011).

Este estado de coisas tem constituído motivação importante no sentido de explicar tentativas de desenvolvimento de sistemas automáticos de classificação em particular de itens de resposta aberta, tema que é o objeto nos números seguintes.

1.4. Avaliação Automática de Conhecimentos.

1.4.1. Introdução.

A necessidade de desenvolver sistemas automáticos de avaliação de conhecimentos confunde-se, praticamente, com a necessidade de avaliar automaticamente conhecimentos expressos por respostas a questões abertas (itens em que os respondentes respondem elaborando a resposta e não selecionando apenas a resposta como nos testes de itens de resposta múltipla). Em particular, a questões em que a resposta tem a forma de um texto.

Embora existam sistemas automáticos de avaliação de conhecimentos – por exemplo, para matemática – que se baseiam em grafos, símbolos matemáticos ou equações – ver, por exemplo, Williamson, Benmett, Bernstein, Foltz, Landauer, Rusin, Way, e Sweeney, (2010) – a situação de longe mais comum é aquela em que a resposta é um texto: ensaio ou texto mais curto.

Esta tendência tem-se acentuado à medida em que avança a investigação sobre o processamento automático da língua natural, as técnicas de inteligência artificial e sobretudo as técnicas de análise estatística de textos e de mineração de textos.

Surpreendentemente, como se verá em **1.4.3.**, os sistemas disponíveis no mercado têm-se mantido – embora com aperfeiçoamento constante – relativamente estáveis. No que se segue, a sigla SAAT serve para designar sistema Automático de Avaliação de Textos, equivalente ao conceito em inglês *Automatic Essay Scoring* (AES).

1.4.2. Avaliação Baseada em Itens de Resposta Aberta.

Uma das vantagens dos itens de resposta múltipla está – o que vai contra a intuição imediata - na possibilidade de reduzir tanto quanto for necessário ou desejável os inconvenientes da resposta ao acaso. Por exemplo, se um item de resposta múltipla tiver 4 alternativas de resposta, a probabilidade de que um estudante acerte na resposta correta, escolhendo a resposta ao acaso ("adivinhar" a resposta certa), é de ¼= 0.25.

É claro que para *k* alternativas de resposta, essa probabilidade é de $1/k$. Em particular, para *k*= 2, esta probabilidade é enorme (0.5).

Para um teste que tenha *n* itens com *k* alternativas de resposta cada uma, a probabilidade de que um respondente acerte em *a* desses itens respondendo ao acaso (*a*= 0, 1, 2, …, *n*) é dada pela distribuição binomial. Isto é: se *A* designar o número de acertos respondendo ao acaso a um teste de *n* itens cada um com *k*= alternativas, a probabilidade de que *A* tenha o valor *a* é

$$P(A=a)=\binom{n}{a}p^{a}\left(1-p\right)^{n-a}, \text{ em que } p=1/k.$$

Assim, por exemplo, se um teste tem *n*= 20 itens cada um com *k*= 4 alternativas de resposta, a probabilidade de que um respondente acerte em 10 desses itens escolhendo as respostas ao acaso é:

$$P(A=10)=\binom{20}{10}0.25^{10}\times 0.75^{10}=0.0099=0.99\%$$

Em resumo: é extremamente difícil alguém acertar em 10 itens num teste com 20 itens de 4 alternativas, escolhendo as respostas ao acaso.

Admitindo que para "ter positiva" é necessário obter 10 ou mais itens certos, essa probabilidade num teste desse tipo, escolhendo a resposta ao acaso, seria

$$P(A\geq 10)=\sum_{a=10}^{20}\binom{20}{a}0.25^{a}\times 0.75^{10-a}=0.0139=1.39\%.$$

Isto significa que se pode dificultar tanto quanto desejável ou conveniente os efeitos do chamado "*guessing*" (adivinhar a resposta) aumentando o número de itens e/ou o número de alternativas de cada item.

Está aqui uma das grandes vantagens deste tipo de itens. Outras vantagens que explicam a sua enorme popularidade está no facto de ser possível criar bases de dados de itens, aplicados em testes anteriores a grandes populações e sobre as quais é admissível

criar sistemas automáticos de elaboração e calibração de testes. Ver, por exemplo, Wainer (1990) em que é apresentado a estado da arte nesta matéria em 1990.

O uso deste tipo de itens, associado ao TRI (ver atrás) torna possível gerar, por processos puramente automáticos adaptados ao nível de competência do indivíduo (*Computer Assisted Testing*), testes com itens calibrados cujas características psicométricas (validade e fiabilidade) sejam pré-estabelecidas ou estimadas com relativa precisão.

A juntar a estas possibilidades há a considerar as vantagens teóricas e computacionais que ficam relativamente facilitadas quando se usam itens de resposta múltipla (Wainer, 1990; Liu, Lee, & Linn, 2011).

Entre outras vantagens citam-se a facilidade com que os testes são classificados, a ausência de ambiguidade na atribuição da classificação e a consequente economia e facilidade de organização logística.

Apesar de todas estas vantagens têm-se levantado contra este tipo de item as vozes de investigadores na área da educação. Ribeiro (1989) aponta como principais desvantagens deste tipo de item o facto de a respetiva construção consumir muito tempo, ser difícil a identificação de alternativas falsas plausíveis e não servirem para avaliar aptidões de expressão verbal, organização de ideias ou outras em que o respondente organize a resposta (Rauch & Hartig, 2010).

Dum modo geral, este tipo de item não pode ser usado em situações em que o que se pretende avaliar é a capacidade de desempenhar papéis ou a deteção de certos comportamentos ou traços profundos que só se revelam pela realização (*performance*) de certas tarefas.

A resposta natural a estas dificuldades é usar itens em que o respondente elabora a resposta (*Construct Response Item* (CR), na designação anglo-saxónica).

Em particular, usar itens em que o respondente escreve um texto para responder à questão formulada.

Numerosos estudos têm mostrado que a elaboração das respostas a este tipo de questões (Rauch & Hastig, 2010) é em si mesmo instrumento importante de aquisição de conhecimentos. Do ponto de vista da avaliação e classificação, a elaboração da resposta faz intervir e portanto manifestar-se, certos tipos de traços latentes inacessíveis aos itens de resposta múltipla.

Põem, contudo, numerosos problemas, o principal dos quais é o custo da respetiva classificação, dificuldades logísticas, técnicas e problemas de validade e fiabilidade.

Quando é necessário atribuir classificações a um grande número de respondentes em exames nacionais, por exemplo, torna-se necessário recrutar, formar, treinar os professores a quem vai ser atribuída a responsabilidade pelas classificações e depois monitorizar as classificações atribuídas para garantir que a coerência/fiabilidade se mantêm.

Trata-se de um processo muito caro e sujeito a inúmeras dificuldades técnicas, as principais das quais são garantir que há uniformidade de critérios, ausência de subjetivismos e de ambiguidades tanto na atribuição de resultados como na elaboração dos testes (Ribeiro, 1989; Rauch & Hastig, 2010). Em síntese, em termos psicométricos e para lá das questões económicas, põe-se problemas sérios de validade e fiabilidade que tem de ser considerados.

Logo que a tecnologia dos computadores (*software* sobretudo) o possibilitou – por volta de meados da década de 60-70 do século passado – começaram, de modo natural, a aparecer os primeiros sistemas de avaliação e classificação baseados na avaliação automática dos textos de resposta a questões abertas (Shermis & Burstein, 2003).

1.4.3. Sistemas de Avaliação Automática de Textos (SAAT's).

Em Shermis e Burstein (2003) define-se avaliação automática de textos (*Automatic Essay Scoring* (AES)) como a capacidade de avaliar e classificar prosa escrita usando a tecnologia dos computadores.

O primeiro sistema deste tipo surge em 1966 e desde então têm aparecido novos sistemas cujas características principais podem ser apreciadas através das referências a seguir mencionadas.

Em Rudner e Gagne (2001), Yang, Buckendahl, Juszkiewicz, e Bhola (2012), Shermis e Burstein (2003), Dikli (2006), He, Hui, e Quan (2009), Vojac, Kline, Cope, McCarthy, e Kalantzis (2011) e ETS (2013) são apresentadas sucessivas revisões acerca do estado da arte relativa à evolução dos sistemas AES e da investigação em diversos domínios (estatística, linguística, processamento da linguagem natural e *text mining*) que tornou possível o seu desenvolvimento.

Surpreendentemente, a lista de SAAT's, iniciada com o sistema PEG (*Project Essay Granding*) introduzido em 1966 por Ellis Page – Page (1994) – tem-se mantido relativamente estável. É possível constatar através das referências citadas que os sistemas mais salientes em 2008, para lá do já citado PEG, são o *Intelligent Essay Assessor* (IEA)

introduzido em 1997 por Landauer e Foltz (ver à frente), o *E-Rater* desenvolvido no ETS (*Educational Testing Service*) por Jill Burstein, o sistema *Intellimetrics*, desenvolvido em 1998 e o sistema BETSY (*Bayesian Essay Test Scoring System*^{TM}) desenvolvido em 2002.

Consultando Vojac et al., (2011), dez anos depois, vemos que a lista de SAAT's mais usados continuaria a incluir os citados PEG, IEA, *E-Rater*, *Intellimetrics*, bem como novos sistemas muitas vezes desenvolvidos em estreita ligação ou como extensões dos primeiros mas privilegiando agora a interação com o estudante a o respetivo *feedback*, fazendo uso de resultados da investigação na área do processamento automático da linguagem natural, Natural Language Processing (NLP) e da Internet.

Entre os novos sistemas referem-se, citando a referência mencionada, os sistemas *Criterion*, *Write to Learn*, *My Access*, *SA Grades*, entre outros.

Muitos destes sistemas estão disponíveis na Internet e são de livre acesso se bem que a respetiva tecnologia nem sempre seja totalmente acessível (Vojac et al., 2011).

Um dos sistemas que mais tem contribuído para a aceitação cautelosa mas progressiva da utilização destes programas como instrumentos de avaliação de conhecimentos é o sistema PEG descrito em Page (1994). A ideia básica deste primeiro sistema, relevante ainda nos dias de hoje, consiste em tentar medir certas qualidades intrínsecas (não observáveis diretamente) designadas "*trins*"- tais como a dicção, fluência, pontuação e outras. Estas variáveis estão muito correlacionadas com outras – essas sim observáveis diretamente – designadas "*proxies*". Por exemplo, a fluência (um "*trin*") está correlacionada com a "*proxi*" número de palavras do texto e a dicção correlacionada com a "*proxi*" variância do comprimento das palavras (Page, 1994). Se n textos produzidos pelos estudantes forem classificados por um juiz humano que atribui ao texto número i a classificação y_i (i= 1, 2, …, n) então, assumindo que se consideraram no modelo as k "*proxies*" P_1, P_2, …, P_k, a classificação y_i relaciona-se com os "*proxies*" através de um modelo linear expresso por uma regressão múltipla:

$$y_i = a_i + \beta_{i1} P_1 + \beta_{i2} P_2 + \ldots + \beta_{ik} P_k + e_i \ (i= 1, \ldots, n)$$

em que β_{ij} são coeficientes a estimar durante a fase de "treino" (estimação do modelo) usando os resultados de testes previamente classificados por professores, em que os "*proxies*" desempenham o papel de variáveis independentes e as classificações y_i (i= 1, …, n) são variáveis dependentes.

Em geral, *n* varia de 100 a 400, usando-se mais do que um professor para realizar a classificação dos testes da amostra usada na estimação (ou aprendizagem) do modelo, podendo o número de "*trins*" e "*proxies*" ser da ordem das dezenas (Page, 1994).

Como é habitual com os modelos de regressão múltipla, a qualidade global do ajustamento é medida pelo R^2 (coeficiente de determinação). Os valores observados de R^2 foram, desde as primeiras experiências, bastante elevados e encorajantes – acima de 0.75 (Page, 1994; Shermis & Burstein, 2003).

Embora o sistema PEG tenha sido desenvolvido em torno das ideias acima descritas, é difícil não ver nos "*trins*" aquilo a que hoje se designaria por variáveis latentes (traços subjacentes aos textos produzidos ao responder às questões e inacessíveis à observação direta) e nas variáveis designadas "*proxies*" aquilo que se designaria, atualmente, por variáveis manifestas ou indicadores das primeiras.

Isto significa que, implicitamente, subjacente à descrição do sistema PEG - poderia ver-se um modelo de equações estruturais como o sugerido na **figura 1.4.3.1.**

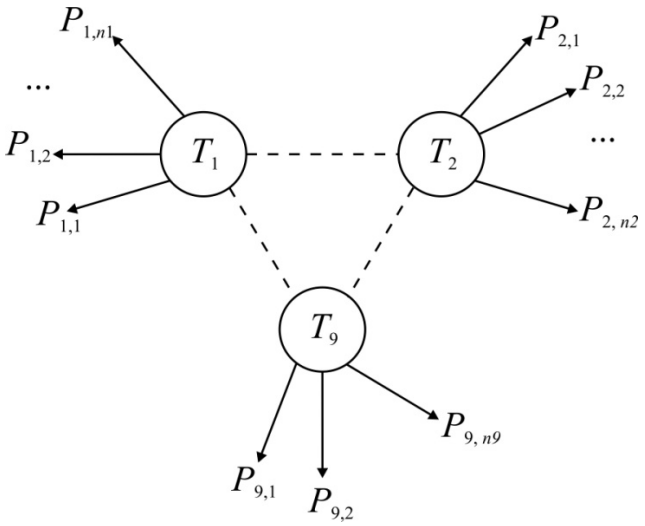


Figura 1.4.3.1. Modelo estrutural sugerido (elaboração própria) pela estrutura do sistema PEG. O modelo estrutural – a tracejado – está indefinido.

Na **figura 1.4.3.1.** (elaboração própria) o modelo observacional (relações lineares entre as variáveis T_1, …, T_9, "*trins*" ou variáveis latentes), $P_{i,j}$ (i= 1, …, 9), (j= 1, …, n_9) são as variáveis indicadoras de cada uma das "*trins*" sugerindo o sentido das setas que os "*proxies*" são manifestações observáveis dos "*trins*". O modelo estrutural (relações entre os "*trins*") está indefinido e por isso representado por tracejados.

O sistema IEA (*Intelligent Essay Assessor*) – Landauer, McNamara, Dennis, e Kintsch (2007) - baseia-se na técnica LSA – *Latent Semantic Analysis* – que assenta na deteção de proximidades entre o significado das palavras usadas num certo conjunto de textos, baseada na técnica computacional da decomposição de matrizes em valores e vetores singulares (SVD). Uma vez que esta técnica é objeto da nossa especial atenção, reservamos para o número **1.4.4.** um desenvolvimento mais detalhado deste tema.

De acordo com Shermis e Burstein (2003) e Attali e Bustein (2004) o *E-rater* é um sistema de classificação automática de textos desenvolvido como instrumento de apoio a usar na avaliação das respostas a questões abertas. As classificações são expressas numa escala 0-6.

O sistema foi desenvolvido no ETS – *Educational Testing Service* americano em cujo *site* podem ser consultados numerosos artigos técnicos, estudos e relatórios relativos à experiência obtida com o sistema e investigação geral sobre este tema. Ver *site* do ETS – *Educational Testing Service*, http://www.ets.org/, consultado em 1/5/2013.

Concetualmente (Shermis & Burstein, 2003) o sistema *E-rater* que é baseado no processamento automático da linguagem natural – procura identificar nos textos caraterísticas (*features*) que expressem qualidades profundas da escrita do respondente. Para isso dispõe de três módulos especializados (sintaxe, discurso, tópicos) cada um dos quais deteta no texto em análise os atributos da respetiva categoria. No final, os valores destes atributos são combinados numa classificação global.

IntelliMetrics® é um produto da firma Vantage, desenvolvido em 1997 e que tem sido usado nos estados Unidos para avaliar respostas a questões abertas contidas em exames nacionais – nomeadamente no GMAT® (*Graduate Management Admission Test*), (Rudner, Garcia, & Welch, 2006; Lam, Dullon, & Chang, 2010). Pode também ser consultado no *site* da firma Vantage (http://www.vantagelearning.com/, consultado em 14 Abril 2013) onde se pode ler uma discrição do respetivo funcionamento. Trata-se do primeiro sistema de avaliação automática de conhecimentos baseado em conceitos de inteligência artificial, conhecendo-se relativamente pouco acerca dos aspetos técnicos específicos do respetivo funcionamento. Em Rudner, Garcia, e Welch (2010) é apresentada a avaliação a que esse sistema foi submetido no contexto de um concurso que culminou com a admissão do *IntelliMetrics*® como instrumento de avaliação automática de conhecimentos do exame GMAT® da associação *Graduate Management Admission*

Council (GMAC). Através dessa avaliação pode verificar-se que esse sistema tem um comportamento praticamente indistinguível do dos classificadores humanos e praticamente a mesma precisão. Com efeito, verificou-se que os coeficientes de correlação média entre as classificações atribuídas pelo sistema e as classificações atribuídas pelos classificadores humanos eram frequentemente acima de 0.98 (Rudner, Garcia, & Welch, 2006).

Uma outra vantagem deste sistema é a de poder realizar a classificação de textos escritos em diversas línguas incluindo, além do Inglês, o Francês, Espanhol, Português, Arábico e Japonês (Dikli, 2006).

O sistema BETSY foi desenvolvido por Laurence M. Rudner - Departamento de Medição, Estatística e Avaliação da Universidade de Maryland.

Em Rudner e Liang (2002) e em Dikli (2006) podem ver-se resumos das abordagens seguidas para desenvolver este sistema, bem como as principais caraterísticas do sistema BETSY, atualmente usado essencialmente em investigação e acessível de modo totalmente grátis na internet (*www.edres.org/betsy*, consultado em 14 Abril 2013).

O interesse deste sistema assenta na metodologia usada no seu desenvolvimento – o teorema de Bayes e redes bayesianas. Dado o texto de resposta a um item, o objetivo é classificar este texto numa das categorias (classes) de um conjunto $\{C_1, C_2, \ldots, C_k\}$ categorias. Por exemplo $\{C_1, C_2, C_3\}$ = {Apropriado, Parcial, Inadequado} ou $\{C_1, C_2, C_3\} = \{A, P, I\}$.

Não havendo qualquer informação *à priori* acerca do comportamento dos estudantes, pode-se assumir que as probabilidade à priori de que o texto em causa pertença a cada uma dessas categorias sejam iguais a $1/3 = P(A) = P(P) = P(I)$ no caso do exemplo ou, no caso de um conjunto de k classes, a $1/k = P(C_1) = P(C_2) = \ldots = P(C_k)$.

Conhecendo os resultados de n testes anteriormente classificados, a distribuição *à priori* de probabilidade poderia ser – por exemplo - $P(C_j)=\frac{\#(C_j)}{n}$ em que $\#(C_j)$ é o número de textos classificados na classe j ($j= 1, 2, \ldots, k$).

Se sobre o texto da resposta em causa foram observados A caraterísticas – por exemplo, as palavras distintas $W_1, W_2, \ldots, W_i, \ldots, W_A$ – podem definir-se as probabilidades $P(W_i \mid C_j)$ ($i= 1, \ldots, A$; $j= 1, \ldots, k$), isto é, a probabilidade de que a palavra W_i ocorra no texto em estudo, assumindo que o texto é da classe j ($i= 1, \ldots k$).

No caso do exemplo anterior com 3 categorias, P (W_i | Não satisfaz), seria a probabilidade de observar a presença da palavra W_i num texto da categoria "Não satisfaz".

Mais uma vez, os valores $P(W_i \mid C_j)$ (j= 1, …, k) podem ser estimados usando os resultados de testes anteriormente classificados.

Sabendo que a palavra W_i ocorreu no texto j já classificado, usando o teorema de Bayes, é possível passar das probabilidades "*à priori*" $P(A)$, $P(P)$, $P(I)$, no caso do exemplo anterior ou, no caso geral – $P(C_j)$, j= 1, …, k – para as probabilidades *à posteriori*:

$$P(C_j \mid W_i) = \frac{P(W_i \mid C_j) \times P(C_j)}{\sum_{e=1}^{k} P(W_i \mid C_e) \times P(C_e)} \quad (j = 1, 2, \ldots, k).$$

Deste modo, para o texto específico em consideração, ter-se-iam k probabilidades *à posteriori* condicionadas à observação (ou não) da palavra i no texto em causa. Isto é, para o texto t, essas probabilidades seriam:

$P(C_1 \mid W_i)$, $P(C_2 \mid W_i)$, …, $P(C_k \mid W_i)$.

A regra de classificação de Bayes consiste em classificar o texto t na classe de maior probabilidade (*à posteriori*).

Por exemplo, usando a escala de três posições antes definidas, $\{C_1, C_2, C_3\} = \{A, P, I\}$, ter-se-ia para o texto em causa t as probabilidades *à priori* $P(A \mid W_i)$, $P(P \mid W_i)$, …, $P(I \mid W_i)$.

Se as probabilidades *à priori* fossem {0.4, 0.4, 0.2} e, realizados os cálculos, as probabilidades *à posteriori* fossem {0.6, 0.3, 0.1} então o texto em análise deveria ser classificado na classe A visto que é ela que tem a maior probabilidade "*à posteriori*".

A simplicidade deste método e a consequente facilidade da sua implementação – bem como a qualidade dos resultados com ele obtidos (Rudner & Liang, 2002; Dikli, 2006) explica o êxito do sistema BETSY que, hoje em dia é essencialmente usado como instrumento de investigação.

1.4.4. Classificação baseada na Análise Semântica Latente.

1.4.4.1. Introdução.

Como se viu atrás, o sistema de avaliação e classificação automática IEA (*Intelligent Essay Assessor*) usa a metodologia LSA designada neste trabalho por ASL.

A ASL assenta computacionalmente na decomposição em valores e vetores singulares (Eckart & Young, 1936; Landauer, McNamara, Dennis, & Kintsch, 2007) de matrizes com frequências de ocorrência de palavras em conjuntos de textos e, do ponto de

vista psicológico, em considerar que essa decomposição é um bom modelo para representar o processo de aquisição do significado de novas palavras por parte das crianças. Ver **Capítulo II** pormenores da nossa própria implementação deste sistema e no **Capítulo III** resultados obtidos com dados reais.

De acordo com Chomsky (1968) todos os seres humanos nascem com um conjunto inato de capacidades que lhes permitem, sem grande esforço, interpretar as estruturas das frases, limitando assim, de modo drástico, o número de significados possíveis de novas palavras que seria necessário considerar para aprender o respetivo significado.

Se bem que esta teoria da existência de algo que se poderia designar por um dispositivo ou órgão inato de aquisição de palavras *Word Acquisition Device* (WAD), (Bloom, 2000; Chomsky, 1968) - ou *Language Acquisition Device* (LAD) tenha em muitas ocasiões sido posta em causa, acaba, recentemente, de receber uma importante confirmação experimental nos trabalhos do grupo *Blue Brain Project* (Hill, Wang, Riachi, Schurmann & Markram, 2012) em que é apresentada evidência física da existência de certas estruturas neuronais inatas que condicionam de modo importante o processo de aquisição da linguagem (e de outros processos) e portanto apoiam implicitamente as teorias de Chomsky.

No que se segue referimo-nos aos trabalhos de Landauer e do seu grupo que levaram à criação da ASL partindo precisamente de pesquisas que visavam criar um modelo para o processo de aquisição de significado de novas palavras pelas crianças.

1.4.4.2. A Análise da Semântica Latente como Modelo de Aquisição do Significado das Palavras pelas Crianças.

No site *http://lsa.colorado.edu/*, consultado em 12 Abril 2012, no Departamento de Psicologia, Instituto de Ciências Cognitivas da Universidade de Colorado podem encontrar-se e obter-se livremente as principais contribuições para a literatura da ASL, nomeadamente Landauer e Dumais (1997) e Landauer, Foltz, e Laham (1998). Em Landauer, McNamara, Dennis, e Kintsch (2007) é apresentado um panorama atualizado destas metodologias.

O referido *site* permite ainda o acesso (utilização "*on-line*") ao *software* IEA da firma Pearson e a realização de experiências de classificação e comparação de textos com esse *software*, bem como acesso aos sites e publicações dos restantes membros do grupo de investigação associadas a Landauer e à metodologia ASL.

Em Landauer e Dumais (1997) são lançados os fundamentos psicológicos e computacionais da ALS a partir da procura de uma "solução" para o chamado "Problema de Platão" – que é afinal o problema de aquisição do significado das palavras por parte das pessoas – as crianças em particular.

Em Landauer e Dumais (1997) e também Bloom (2000) nota-se a aparente extrema facilidade com que as crianças adquirem o significado de novas palavras. Se cerca dos 17 anos um jovem pode já ter adquirido o significado de cerca de 60 000 termos, isso implica uma taxa diária média de 10 termos por dia.

Este ritmo não pode ser justificado pela teoria da aprendizagem clássica que explica a obtenção de novos significados pela associação de palavras a conceitos já aprendidos e com expressão física através da criação de novas ligações neuronais – processo demasiado lento face à evidência experimental da velocidade real de aquisição de novos temas.

Em Landauer e Dumais (1997) procura-se mostrar que a semelhança semântica é o resultado das ocorrências conjuntas (coocorrências) das palavras nos textos e das experiências a que as pessoas são expostas. Esta semelhança semântica (psicológica) resultante das coocorrências citadas, pode ser representada como distância (ou proximidade) num espaço métrico de dimensão (k) adequada.

Isto é, a verosimilhança de coocorrência de duas palavras num mesmo texto escolhido ao acaso é tanto maior quanto mais relacionados estiverem os significados dessas palavras. Esta semelhança, de caráter psicológico, poderia ser captada pela noção de métrica de um espaço de dimensão adequada – em que esta dimensão deveria agora ser estabelecida por métodos computacionais e estatísticos.

Mais especificamente, se considerarmos todas as coocorrências de palavras num conjunto de textos (contando essas coocorrências) e interpretando semelhança psicológica como semelhança de significados, pode pôr-se a questão de obter uma representação dessa informação num espaço métrico de dimensão adequada.

Este modo de formular o problema que se pode encontrar em Landauer e Dumais (1997) conduz naturalmente a um problema de MDS (*Multidimensional Scaling*) que pode ser ilustrado do seguinte modo. Consideremos um texto hipotético contendo apenas as palavras W_1, W_2, W_3, W_4, W_5, W_6.

O "texto" é:

W_1 W_2 W_3 W_4 W_1 W_5 W_6 W_2 W_3 W_2 W_4
W_1 W_2 W_5 W_6 W_3 W_2 W_4 W_5 W_6 W_4.

As 20 coocorrências dessas palavras no "texto" anterior são:

$(W_1\ W_2)$, $(W_2\ W_3)$, $(W_3\ W_4)$, $(W_4\ W_1)$, $(W_1\ W_5)$, $(W_5\ W_6)$, $(W_6\ W_2)$, $(W_2\ W_3)$, $(W_3\ W_2)$, $(W_2\ W_4)$, $(W_4\ W_1)$, $(W_1\ W_2)$, $(W_2\ W_5)$, $(W_4\ W_6)$, $(W_6\ W_5)$, $(W_3\ W_2)$, $(W_2\ W_4)$, $(W_4\ W_5)$, $(W_5\ W_6)$, $(W_6\ W_4)$.

Não considerando a ordem das ocorrências (isto é, considerando que $W_i\ W_j$ é o mesmo que $W_j\ W_i$) obtém-se a seguinte tabela de frequências:

Coocorrências de Palavras	**Frequências Absolutas**	**Frequências Relativas**
$W_1\ W_2$	2	0.10
$W_2\ W_3$	4	0.20
$W_3\ W_4$	1	0.05
$W_1\ W_4$	2	0.10
$W_1\ W_5$	1	0.05
$W_5\ W_6$	3	0.15
$W_3\ W_6$	2	0.10
$W_2\ W_4$	2	0.10
$W_2\ W_5$	1	0.05
$W_4\ W_5$	1	0.05
$W_4\ W_6$	1	0.05
	20	

Tabela 1.4.4.2.1. Coocorrências das "palavras" w_i, w_j num texto.

As frequências relativas observadas podem ser interpretadas como semelhança ou proximidades entre as coocorrências dos significados das palavras. Observe-se que aqui consideram-se coocorrências não as ocorrências simultâneas de duas palavras num texto mas a sua ocorrência de uma seguida da outra.

Representando estas proximidades de ocorrência dos significados das palavras por uma tabela simétrica, obtém-se a **tabela 1.4.4.2.2.** seguinte:

	W_1	W_2	W_3	W_4	W_5	W_6
W_1	1	0.10	X	0.10	0.05	X
W_2	2	1	0.20	0.10	0.05	X
W_3	X	4	X	0.05	X	0.10
W_4	2	2	1	1	0.05	0.05
W_5	1	1	X	1	1	0.15
W_6	X	X	2	1	3	1

Tabela 1.4.4.2.2. Na tabela a parte triangular inferior contém as frequências absolutas e a superior as frequências relativas; o símbolo X significa ausência de informação.

Aplicando a esta tabela um programa de *Multidimensional Scaling*[2] verifica-se, usando, por exemplo, o algoritmo ALSCAL do SPSS© – ver SPSS (2007) – que a dimensão do espaço métrico que melhor representa estas coocorrências com o menor *stress* possível é o espaço de dimensão k= 1, como se vê na **figura 1.4.4.2.1.**

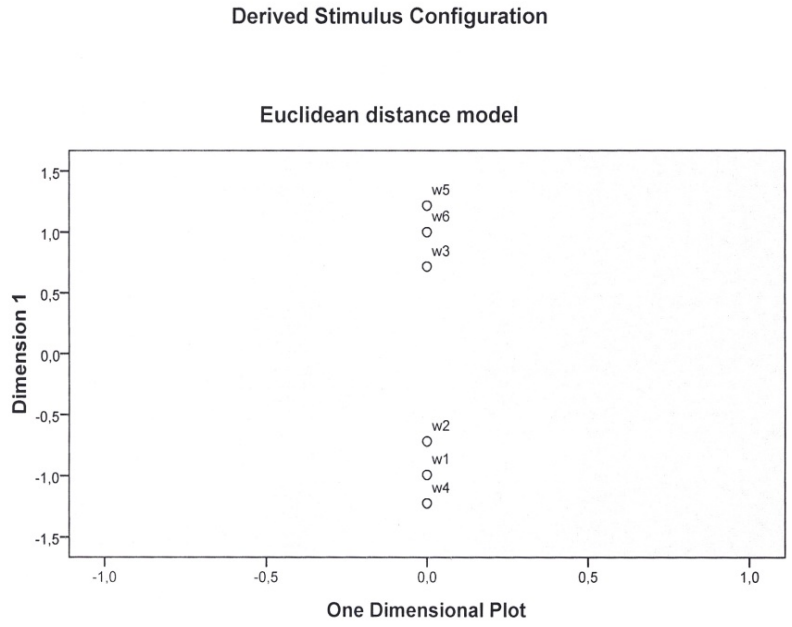


Figura 1.4.4.2.1. Representação geométrica das proximidades de palavras do texto, obtida com o SPSS. Neste caso, a dimensão de espaço é k= 1.

[2] O MDS (*Multidimensional Scaling*) procura a escala (unidimensional ou multidimensional) que melhor aproxima as perceções de proximidade de conceitos.

Como se ilustra no exemplo anterior, obtém-se uma representação geométrica correspondente às proximidades das palavras no texto sem que tenha sido necessário relacionar o significado das palavras com o significado de outros conceitos preexistentes, exteriores aos dados, como exige a teoria da aprendizagem dos significados.

No exemplo anterior, o resultado final da "aprendizagem" baseada na coocorrência de palavras seria a associação dos significados das palavras w_1, w_2, w_4 entre si e a associação dos significados das palavras w_3, w_5, w_6. Por outro lado, a quase totalidade da informação poderia ficar representada num espaço de dimensão 1 (uma linha em cujos extremos estariam os dois grupos). Isto é, neste caso, a experiência empírica observada poderia ser expressa falando dos dois grupos de palavras e da sua oposição. Em síntese, a aprendizagem ou a formação do significado das palavras baseada nas coocorrências – como aqui se ilustra – não apela ao esquema típico de aprendizagem "com professor" em que para cada nova palavra que ocorresse seria necessário que um professor associasse ou apontasse o significado de conceitos já aprendidos com os quais o significado da nova palavra ficaria associado. Esse esquema não poderia explicar a extraordinária rapidez de aquisição do significado de novas palavras pelas crianças porque exigiria comparações muito lentas entre o significado potencial da nova palavra e o de um grande número de significados alternativos de conceitos pré-existentes (Landauer & Dumais, 1997).

1.4.4.3. ASL como Instrumento de Avaliação Automática.

O modelo ASL, desenvolvido pelo grupo de Landauer (Landauer & Dumais, 1997) assenta na decomposição em valores e vetores singulares das matrizes de frequências de ocorrência de palavras em textos, assumindo que esse processo modela ou simula de modo adequado a aquisição de novos significados por parte das crianças.

Dado um conjunto de documentos designados D_1, D_2, …, D_i, …, D_p e de termos W_1, W_2, …, W_j, …, W_n, a tabela F contém as frequências de ocorrência dos termos ou palavras nos documentos. Isto é, se $F = [f_{ij}]$, $i= 1, \ldots, p$; $j= 1, \ldots, n$ representar essa tabela, então cada palavra (uma linha dessa tabela) fica representado por um vetor de frequências de ocorrência do termo nos documentos e cada documento (coluna) fica representado por um vetor de dimensão n formado pelas frequências de ocorrência das palavras nesse texto. O valor f_{ij} representa portanto a frequência do termos i no documento j.

Nesta representação que remonta a Salton, Wong, e Yang (1975) e a trabalhos relativos à indexação automática de documentos, não há qualquer referência à ordem pela

qual os termos ocorrem nos diversos textos, sendo esta matriz analisável por diferentes algoritmos conforme o problema a resolver.

Designa-se por modelo vetorial porque cada documento fica representado por um vetor com um número de dimensões igual ao número de palavras que ocorrem no *corpus* D_1, D_2, …, D_i, …, D_p, sendo os valores das componentes do vetor as frequências (ou valores que resultam da transformação das frequências) de ocorrência dessas palavras no texto. Do mesmo modo, cada palavra fica representada pelo vetor de frequências (ou valores que resultam da transformação das frequências) da ocorrência dessa palavra em todos os documentos do *corpus* e com um número de dimensões igual ao número de documentos (p).

Como pode ver-se em Landauer e Dumais (1997) e em Landauer, Foltz, e Laham (1998), a ASL é definida (em tradução livre) como "*uma teoria e uma metodologia que permite extrair e representar o significado das palavras a partir do contexto das respetivas coocorrências num grande corpus de texto"*.

Isto significa que se vê na ASL uma teoria ou modelo psicológico adequado à representação do mecanismo através do qual as pessoas (as crianças em particular) constroem o significado das novas palavras, situando-as num espaço métrico de dimensão adequada e "deduzindo" desta representação – quando têm de se expressar – proximidades psicológicas que têm correspondência em representações geométricas (espaço métrico de certa dimensão).

Para lá da componente psicológica, o modelo comporta uma componente computacional que faz uso do método dos mínimos quadrados para detetar as proximidades geométricas entre palavras, entre textos e entre palavras e textos implícitas na matriz de frequências.

Com efeito, se F (n, p) representar a tabela de contingência que contém no cruzamento da linha i (i= 1, …, n) com a coluna j (j= 1, …, p) o número de vezes que a palavra i ocorre no texto j, esta tabela expressa, no fundo, uma série de restrições (impostas pelas coocorrências de palavras sobre os diversos textos) do tipo daquelas que foram exemplificadas no número anterior.

Estas dependências não são imediatamente evidentes a partir da inspeção visual da tabela de frequências. Por exemplo, é difícil intuir diretamente, por inspeção visual, a cadeia de dependências que resultam para os significados das palavras W_1, W_2, W_3, W_4, W_5, W_6 no exemplo do número anterior.

Contudo, se construirmos com os elementos da **tabela 1.4.4.2.2.** um grafo cujos vértices são as palavras, existindo um arco entre duas palavras quando a frequência das coocorrências é maior que zero, a estrutura subjacente à **tabela 1.4.4.2.2.** e portanto às ocorrências que lhe deram origem pode ser apresentada na **figura 1.4.4.3.1.**

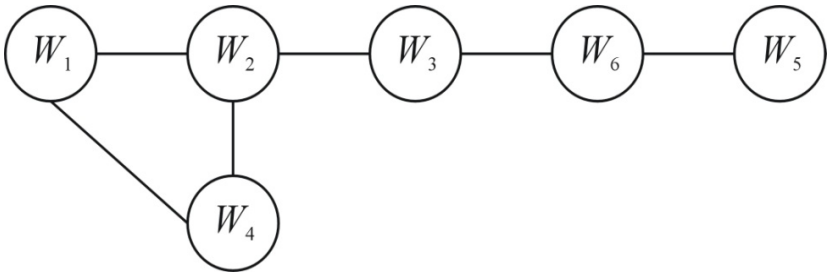


Figura 1.4.4.3.1. Grafo correspondente à **tabela 1.4.4.2.2.** depois de ter eliminado os arcos correspondentes à frequências mais baixas.

Na **figura 1.4.4.3.1.** foram eliminados os arcos com baixa frequência (correspondentes a coocorrências mais fracas). Vê-se que a estrutura é, grosso modo, linear, correspondente à fornecida pelo processo automático do MDS constante na **figura 1.4.4.2.1.**

O método computacional eleito por Landauer e Dumais (1998) e Landauer, Foltz, e Laham (1998) para simular os processos mentais que têm lugar no processo de aprendizagem humana do significado das palavras é a chamada decomposição em valores e vetores singulares (*Singular Value Decomposition* - SVD), tendo em conta que a SVD se baseia no método dos mínimos quadrados para obter uma representação métrica dos dados (Eckart & Young, 1936; Landauer, Foltz, & Laham, 1998).

Historicamente, este método computacional precede no tempo em muito o seu uso para o fim específico que lhe é atribuído na formulação da ASL uma vez que aparece já em Eckart e Young (1936).

De acordo com Eckart e Young (1936), dada a matriz F de frequências de ocorrência de palavras num conjunto de textos T_1, T_2, …, T de um certo *corpus*, a decomposição $F = U\,D\,V^T$ da matriz F num produto de três matrizes U, D, V é a solução (única) do seguinte problema de otimização: buscar as matrizes U, D, V tais que o quadrado da distância (no sentido dos mínimos quadrados) entre F e o produto $U\,D\,V^T$ seja o mínimo possível. Isto é, se U, D, V forem matrizes variáveis, pretende-se encontrar aquelas matrizes específicas tais que $||F - U\,D\,V^T||^2$ tem o menor valor possível.

Mostra-se em Eckart e Young (1936) que U é uma matriz de dimensões $n \times d$; D é uma matriz diagonal de dimensões $d \times d$; V é uma matriz de dimensões $t \times d$ em que t é o número de textos e d é a dimensão do espaço de representação.

As linhas de U correspondem às palavras mas agora representadas numa dimensão d que é menor ou igual ao número de documentos (p). Por sua vez, a matriz V que representa agora nas suas linhas os documentos mas na dimensão d, não já na dimensão n original. Por sua vez, D tem os elementos da diagonal ordenados por ordem decrescente, representando estes valores as importâncias decrescentes das dimensões finais do nosso espaço de representação (1, 2, …, d).

O valor de d é, em geral, inferior ao valor mínimo (n, p).

Isto é, se F for uma matriz com n palavras (linhas) e p colunas (textos) então, d em geral é ≤ min (n, p).

Estes conceitos podem ser ilustrados no seguinte exemplo numérico.

$F=$

	d_1	d_2	d_3	d_4
W_1	1	0	0	3
W_2	0	1	0	0
W_3	0	1	1	0
W_4	0	0	0	1
W_5	0	0	0	1
W_6	0	1	0	0
W_7	1	0	0	0
W_8	0	0	1	0
W_9	0	1	0	0
W_{10}	2	0	1	0

Tabela 1.4.4.3.1. Frequências de ocorrência das "palavras" w_1 … w_{10} nos textos d_1 … d_4.

Neste exemplo as palavras W_1, W_2, …, W_{10} foram detetadas sobre os documentos (textos) D_1, D_2, …, D_4 com as frequências indicadas. Neste caso, n= 10, p= 4. Isto significa que é possível representar tanto as palavras (linhas) como as colunas (textos) num espaço de dimensão, quando muito d= min (10,4)= 4. Isto é, nunca será necessário mais do que 4 dimensões para expressar o significado quer das palavras quer dos textos que estão representados na matriz F.

Isto significa que há um espaço métrico de dimensão d= min (10, 4) = 4 no qual tanto as palavras como os textos podem ser representados, tendo em conta os pesos dados pelos valores da diagonal de *D*. Ver no **Capítulo II** os pormenores no modo de obter esta representação.

Se apenas considerarmos as suas duas primeiras coordenadas, vemos na **figura 1.4.4.3.2.** as posições relativas ocupadas pelas 10 palavras anteriores e pelos 4 textos. Esta representação contém apenas uma percentagem $\frac{(2.873+2.124)}{2.873+2.124+2+1.12}=62\%$ da informação (ou variabilidade total) contida na tabela *F* (Landauer, Foltz, & Laham, 1998).

Realizando a decomposição, verifica-se que:

$F = U\,D\,V^T$

F=

	t_1	t_2	t_3	t_4
W_1	1	0	0	3
W_2	0	1	0	0
W_3	0	1	1	0
W_4	0	0	0	1
W_5	0	0	0	1
W_6	0	1	0	0
W_7	1	0	0	0
W_8	0	0	1	0
W_9	0	1	0	0
W_{10}	2	0	1	0

U=

-0.447	-0.023	-0.289	0.097
-0.070	-0.325	0.289	0.348
-0.184	-0.245	0.577	-0.307
-0.184	-0.245	-0.289	-0.307
-0.184	-0.245	-0.289	-0.307
-0.070	-0.325	0.289	0.348
-0.447	-0.023	-0.289	0.097
-0.114	0.080	0.289	-0.656
-0.254	-0.570	0.000	0.041
-0.641	0.525	0.289	0.153

D=

2.873	0.000	0.000	0.000
0.000	2.124	0.000	0.000
0.000	0.000	2.000	0.000
0.000	0.000	0.000	1.112

V^T=

	T_1	T_2	T_3	T_4
1	-0.758	0.473	0.000	0.449
2	-0.201	-0.690	0.577	0.387
3	-0.327	0.169	0.577	-0.729
4	-0.528	-0.521	-0.577	-0.342

Tabela 1.4.4.3.2. Componentes da SVD da matriz F.

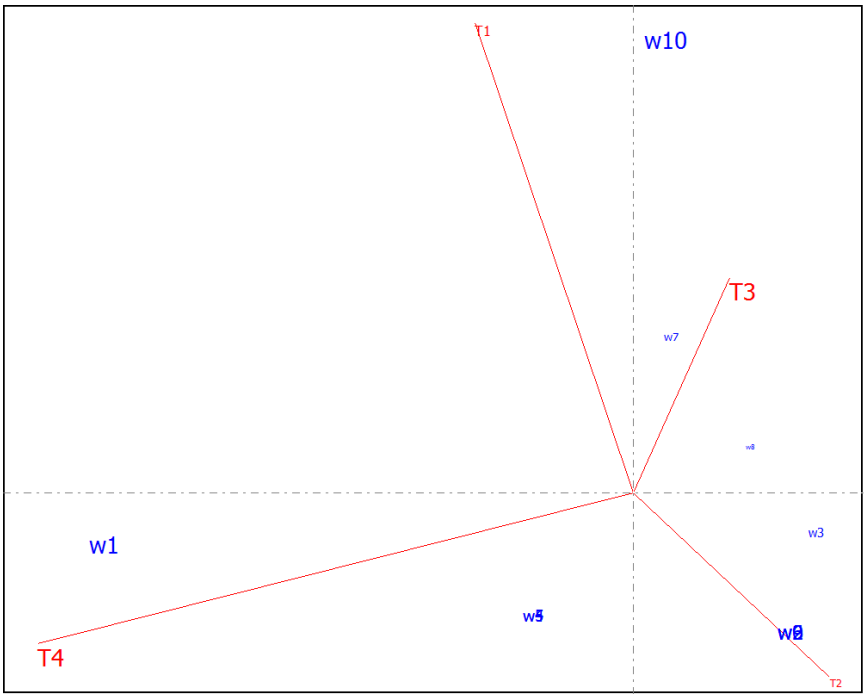


Figura 1.4.4.3.2. Proximidades entre as palavras e os textos da tabela F através da análise SVD. A azul as palavras, a vermelho (linhas) os textos.

Nessa figura estão ainda posicionadas linhas que representam os textos ou documentos T_1, T_2, T_3, T_4 usados no exemplo. Conforme se verá de modo mais pormenorizado no **Capítulo II**, os ângulos entre as posições das palavras (mais precisamente os cossenos destes ângulos) são tanto mais pequenos (cossenos tanto maiores) quanto mais associados estão ao significado das palavras. O mesmo sucede para os significados dos documentos entre si e dos documentos com as palavras. Assim, por exemplo, na **figura 1.4.4.3.2.** os textos T_1 e T_3 fazem um ângulo relativamente pequeno (cosseno elevado) entre si – o que significa que os respetivos significados estão relacionados. Já o ângulo entre os textos T_1 e T_4 é próximo de 90º o que significa que os respetivos significados não têm relação (são muito afastados). Os textos T_1 e T_2 produzem um ângulo acima de 90º - cosseno negativo, o que sugere que os seus significados tendem a opor-se. Com efeito, é isso que sucede na tabela de frequências: quando uma palavra ocorre num dos textos, não corre no outro.

Nesse gráfico nota-se a sobreposição das palavras $\{w_2, w_6, w_9\}$ no símbolo w_0 do canto inferior direito e as palavras $\{w_4$ e $w_5\}$ no ponto central da parte inferior. Isto resulta de na tabela de frequências essas palavras serem representadas por vetores iguais. Isto é: $\{w_2, w_6, w_9\}$ têm o mesmo significado – o mesmo sucedendo com $\{w_4$ e $w_5\}$ sendo este fator captado no gráfico dando a mesma posição a esses termos sinónimos (do ponto de

vista de T_1, T_2, T_3, T_4). Note-se que a sobreposição no plano de mais do que um objeto pode resultar não da igualdade de coordenadas mas do facto de objetos de coordenadas diferentes se projetarem no mesmo ponto. Não é esse o caso presente. Nesse gráfico nota-se ainda o pequeno ângulo formado pela palavra w_7 e o texto T_3 que chama a atenção para o facto de o significado da palavra w_7 ter uma grande contribuição para o significado do texto T_3. Nesta linha de raciocínio, atente-se ainda que o grupo de sinónimos $\{w_2, w_6, w_9\}$ fazem um ângulo quase nulo com T_2, o que chama a atenção para o facto de que o significado do texto T_2 está muito ligado ao significado do grupo $\{w_2, w_6, w_9\}$.

É de notar, ainda, que as palavras w_3, w_8, w_7 estão praticamente alinhadas segundo uma linha reta, o que sugere que há uma relação linear entre os significados destas palavras no conjunto de textos do *corpus*.

Verifica-se assim que a análise do resultado da ASL deste *corpus* põe em evidência pelo menos três relações (lineares) entre as palavras, sugerindo assim que não são necessárias 10 dimensões para representar o significado nem das palavras nem dos textos. O mero facto de se ter constatado que $\{w_2, w_6, w_9\}$ são sinónimos, elimina duas dimensões (no lugar de w_2, w_6, w_9 poderíamos considerar um novo termo que simbolizasse o significado como de $\{w_2, w_6, w_9\}$. O mesmo sucede com $\{w_4, w_5\}$. No conjunto seriam eliminadas três dimensões. A relação linear $\{w_3, w_7, w_8\}$ permitiria eliminar mais uma dimensão (grau de liberdade de variação) reduzindo a dimensão inicial de 10 para 6 necessária à representação dos textos.

Em síntese: o número máximo de dimensões do espaço métrico necessário para "falar" das relações entre as palavras seria, como se viu, não 10 mas apenas 4, quando muito.

No caso de se considerar um "*corpus*" formado por centenas de textos com milhares de palavras, o espaço métrico necessário para uma representação fiável do significado das palavras é de algumas centenas (Landauer, Foltz, & Laham, 1998).

Numa série de experiências de simulação para comparar o comportamento do ser humano e o da metodologia ASL, citada em Landauer, Foltz, e Laham (1998) concluiu-se que a qualidade dos resultados obtidos depende de modo crítico do número d de dimensões retida para o espaço de representação. Em certas situações em que a dimensão do espaço original era de p= 1000, os melhores resultados obtinham-se com cerca de d= 300 dimensões, diminuindo essa qualidade drasticamente até valores muito baixos quando o

número de dimensões se afastava muito (para baixo ou para cima) desse valor aparentemente ótimo.

Em experiências por nós próprios realizadas (ver **Capítulo III**) esta sensibilidade da taxa de erros da ASL em função da dimensionalidade retida para o espaço de representação é amplamente confirmada, sendo pois um elemento importante a reter, se bem que fique em aberto a questão de determinar o valor exato desta dimensão ótima ou da existência de alguma expressão matemática para o efeito.

As nossas próprias experiências parecem sugerir que essa dimensionalidade ótima está ligada a uma variância acumulada de cerca de 85% e a dimensão que garante 85% da informação é cerca de ½ p.

Isto é, no caso de p= 300, por exemplo, a dimensionalidade ótima sugerida pela experiência é de cerca de d= 150, correspondente a cerca de 80% de variância acumulada.

Dada a aparente semelhança com que o método ASL permite representar em espaços métricos o significado das palavras que ocorrem em grandes *corpus* e o correspondente processo mental usado pelos seres humanos para adquirirem e representarem no seu cérebro o significado das novas palavras, desde logo se tornou claro para os autores citados que a ASL poderia ser usada para avaliação de conhecimentos através da análise de textos produzidos pelos estudantes ao responderem a questões abertas sobre certos temas (Landauer & Dumais, 1997; Landauer, Foltz, & Laham, 1998).

Conhecer um certo tema – ou melhor, obter conhecimento válido (em correspondência com a realidade) sobre um certo tema e expressá-lo através de palavras com certo significado – implica possuir na mente uma série de palavras ligadas a esse domínio e seus significados, bem como as proximidades psicológicas inerentes a esses significados. Se o estudante responde com um texto a uma certa pergunta, é natural que as ocorrências de palavras no texto produzido traduzam as proximidades (ou ausência delas) dos significados psicológicos das palavras na sua mente; proximidades estas que podem ser captadas pela representação geométrica construída pela ASL através da decomposição em vetores e valores singulares da matriz de frequências.

Cinco métodos alternativos para proceder a esta avaliação são sugeridos informalmente em Landauer e Foltz (1997), alguns dos quais serão considerados na parte experimental deste trabalho a apresentar no **Capítulo III**.

O significado prático dos resultados obtidos por ASL pode ser apreciado de modo experimental através da consulta do original índice remissivo do manual em Landauer et al., 2007.

O referido índice remissivo foi construído usando a metodologia ASL para analisar os textos das páginas do livro em questão e as palavras consideradas no índice remissivo. Nesta análise, cada página é um texto e as palavras consideradas estão no índice remissivo. Deste modo, quando nesse índice remissivo se procura, por exemplo, a palavra "*semantics*" o que esperaríamos encontrar num índice remissivo clássico seria a indicação das páginas do livro onde poderíamos encontrar, localizar essa palavra. No referido manual, para a palavra "*semantics*", por exemplo, encontramos a referência "*p.355, 0.70, UEMs, familiarity*", o que significa que na representação geométrica gerada pela ASL, a palavra "*semantics*" faz um ângulo de cosseno 0.7 com a página 355. Portanto, o significado de "*semantics*" tem um grau de associação bastante elevado com o significado da página 355. Contudo, a palavra "*semantics*" não aparece nessa página. As expressões cuja semântica mais se relaciona com o significado da página 355 são "UEMs" e *familiarity*.

Ver o índice da obra referida para outras regras de associação palavras/páginas a considerar como exercício de entendimento do significado prático desta técnica.

1.4.5. Validade e Fiabilidade dos Programas de Avaliação Automática de Textos.

Os dois principais critérios de qualidade em investigação na Educação e Psicologia são a validade e a fiabilidade (APA, 2003). Do ponto de vista da classificação automática de textos convém precisar o significado desses dois conceitos quando a atribuição de classificações/pontuações e a avaliação de caraterísticas humanas é realizada por uma máquina: o programa de avaliação automática (SAAT).

A **fiabilidade** tem a ver, grosso modo, com a consistência dos resultados: em que medida são coerentes os resultados obtidos em sucessivas aplicações do processo de classificação automática e que fatores condicionam as eventuais inconsistências (Keeves, 1988).

A **validade** tem a ver com os riscos ou consequências das decisões a tomar com base nas classificações ou pontuações atribuídas pelo processo de classificação automática de textos.

O que se segue baseia-se largamente em Shermis e Burstein (2003), em Cizek e Page (2003) e Keith (2003).

Fontes importantes sobre este tema no contexto dos SAAT's são Bridgeman, Trapant, e Attalli (2012) e Bejar (2011), bem como a normalização da APA (2003).

De acordo com esta última referência, a Fiabilidade tem a ver com a consistência das medições quando o procedimento de testagem é repetido numa população de indivíduos ou grupos. Inerente a todo o processo de medição está sempre associado um erro de medição, de caráter aleatório.

Quando as pontuações/classificações são atribuídas por juízes humanos, seria desejável que uma pessoa ou grupo de pessoas, tendo respondido mais do que uma vez ao mesmo teste, obtivesse sempre a mesma pontuação – o que corresponderia a uma fiabilidade muito elevada. Contudo, esta perspetiva é irrealista uma vez que, para lá da variabilidade inerente ao comportamento dos seres humanos respondentes, as pontuações atribuídas em ocasiões diferentes pelo mesmo juiz humano ou por um conjunto de vários juízes humanos têm variações imputáveis aos erros cometidos por esses juízes.

De acordo com Cizek e Page (2003), expressando um ponto de vista hoje comummente aceite e vertido nas normas APA – ver APA (2003) – a fiabilidade tem a ver com a consistência, a confiança e a reprodutibilidade das pontuações/classificações atribuídas por um certo procedimento de medição, sendo, por isso, uma caraterística do processo de medição e não uma caraterística do teste. Isto é, o mesmo teste aplicado a pessoas diferentes gera pontuações/classificações diferentes.

Mais especificamente (Cizek & Page, 2003) a fiabilidade observada é uma propriedade dos dados gerados pelo processo de classificação; a fiabilidade é uma propriedade da população formada pela totalidade desses valores, sendo pois um valor constante e inacessível. Um valor populacional que tem de ser estimado a partir dos dados observados.

Uma vez que a fiabilidade está relacionada com o erro de observação σ^2 e também com a concordância ou consistência das medições, é natural que na respetiva estimação se considere o valor σ^2 e também o coeficiente de correlação entre dois conjuntos de pontuações/classificações. Sejam: $X = (x_1, x_2, \ldots, x_n)$ e $X' = (x'_1, x'_2, \ldots, x'_n)$ as classificações obtidas por duas aplicações sucessivas do mesmo teste às mesmas pessoas, usando o mesmo processo de medição.

Assim, por exemplo, se os resultados/pontuações/classificações obtidos em dois momentos diferentes pelos mesmos 10 examinandos – numa escala de 0 a 5 – forem

(5, 4, 4, 3, 5, 2, 1, 4, 4, 5) e

(2, 4, 5, 4, 4, 3, 1, 5, 4, 5)

a fiabilidade vista na perspetiva da coerência medida pelo coeficiente de correlação observado seria $r_{XX'}$= 0.91, que é apenas a estimativa do valor inacessível (populacional) de $\rho_{XX'}$.

Esta medição da fiabilidade, através do coeficiente de correlação, tem contudo alguns defeitos, um dos quais ilustrado pelo exemplo seguinte, inspirado em Cizek e Page (2003). Se, voltando à ilustração anterior, os resultados obtidos tivessem sido

X= (5, 4, 4, 3, 5, 2, 1, 4, 4, 5) e

X'= (4, 3, 3, 2, 4, 1, 0, 3, 3, 4)

o coeficiente de correlação observado seria agora $r_{xx'}$= 1. Apesar deste valor (coerência perfeita entre as duas medições), há certamente uma questão de fiabilidade uma vez que não há concordância ou consistência entre os dois conjuntos de resultados. Isto mostra que o coeficiente de correlação não é totalmente adequado como estimador da fiabilidade – o que tem levado a considerar, em complemento, o chamado **coeficiente de concordância** que mede, para cada examinando, o grau de coincidência entre as pontuações atribuídas. Esta concordância pode ser **exata** (quando as pontuações atribuídas coincidem) ou **adjacente** quando, não coincidindo, não diferem, contudo, por mais do que uma unidade na escala que está a ser usada.

Quando em vez de juízes humanos se consideram programas de classificação automática de textos (SAAT's), há que adaptar os conceitos anteriores à nova realidade resultante de um juiz humano ter sido substituído por uma máquina (o *software* de classificação).

Se um programa de classificação automática for aplicado sucessivamente aos mesmos estudantes, o erro, se o houver, não pode ser imputado ao programa uma vez que este funciona sempre do mesmo modo, realizando sempre os mesmos cálculos. Agora (Cizek & Page, 2003) o erro aleatório depende apenas das seguintes componentes:

a) Caraterísticas pessoais dos examinandos (competência, por exemplo);
b) Caraterísticas da linguagem e modo de apresentação do teste;
c) Condições de aplicação do teste.

Em Cizek e Page (2003) pode ver-se uma análise pormenorizada das manifestações destes erros e dos procedimentos estatísticos a aplicar para analisar as componentes destas contribuições bem como das respetivas interações, tanto na perspetiva clássica como na perspetiva do TRI.

Embora não faça sentido proceder a análises de fiabilidade com dados gerados por duas aplicações do mesmo programa, faz no entanto sentido avaliar os efeitos no erro (fiabilidade) ao aplicar aos mesmos estudantes em épocas diferentes versões diferentes do mesmo programa ou versões do SAAT produzidas por grupos diferentes de especialistas ou classificar as mesmas respostas por programas diferentes. Em Cizek e Page (2003) e também Powers, Burnstein, Chodorov, Fowles, e Kukich (2001) são descritas as experiências de validação e cálculos de fiabilidade em relação ao programa *E-rater* do ETS.

O conceito clássico de validade em Ciências da Educação aparece extensamente tratado em Keeves (1998). Em particular, Zeller (1998) define um processo de medição válido como aquele que mede o que pretende medir. Isto é, a validade de um processo de medição tem a ver com o grau com que um indicador empírico mede o que se propõe medir, entendendo-se por medição o processo de ligar um conceito abstrato a indicadores empíricos. Nesta perspetiva não é o indicador em si mesmo que está a ser validado mas sim os objetivos para os quais o indicador está a ser usado. A validade pode ser vista em três perspetivas (Zeller, 1998):

Validade de conteúdo. Em que medida a validade de um indicador corresponde ao conceito teórico que se propõe medir. Ou seja: em que medida um indicador expressa ou reflete a variável latente que pretende representar? Por exemplo, o comprimento de um texto (número de palavras num texto) não é um indicador válido da respetiva qualidade literária.

Validade referida a critérios. Em que medida as pontuações atribuídas por um processo automático de classificação se correlacionam com as pontuações atribuídas por juízes humanos? Este tipo de validade pode ser medida, por exemplo, através do coeficiente de correlação entre o critério (neste caso as pontuações atribuídas por um juiz humano) e as classificações atribuídas pelo classificador automático.

Validade de constructo. Tem a ver com as relações entre os indicadores observados e construções teóricas derivadas a partir de certas hipóteses. Por exemplo, se se admitir que a hipótese teórica de que a língua tem influência sobre os resultados de uma

classificação automática de textos produzidos por falantes de diversas línguas, então numa análise fatorial dos dados obtidos deveria aparecer a evidência de um fator que refletisse essa hipótese: por exemplo, um fator altamente correlacionado com a língua.

Em Bejar (2011) define-se classificação automática (*automated scoring*) como sendo a "*atribuição de uma pontuação (nota, classificação, graduação) - usando um algoritmo - a uma resposta construída por um respondente, em resposta a instruções contidas num teste*".

A validade dos resultados de um processo deste tipo é uma propriedade das inferências produzidas a partir destes resultados. Isto é, a validade destas pontuações tem a ver ou deve ser medida pelas consequências do uso a dar às referidas pontuações. Por outras palavras, ao risco associado ao seu uso em processos de decisão como, por exemplo, consequências para as pessoas testadas e para a sociedade (APA, 2003).

Em Keith (2003) pode ver-se um tratamento bastante extenso e completo das questões de validade inerentes à utilização dos resultados dos processos automáticos de classificação de testes.

Um dos critérios "naturais" usados nos processos de validação dos sistemas SAAT são as classificações atribuídas aos mesmos textos por juízes humanos. Implícito neste procedimento está pois a assunção de que as classificações atribuídas por juízes humanos conduzem a inferências válidas (Keith, 2003). Mais explicitamente, isto significa que a correlação entre as pontuações produzidas pelos SAAT's e produzidas por juízes humanos conduz a inferências válidas sobre a capacidade da escrita, por exemplo.

É claro que quando se usa mais do que um juiz humano, a validade do processo automático medido deste modo depende por sua vez das correlações entre as classificações atribuídas pelos juízes humanos, o que significa (Keith, 2003) que o aumento das correlações entre juízes humanos tem consequências na validade do processo automático.

Keith (2003) apresenta ainda resultados da validação medida em função desta correlação para diversos sistemas em 2003 (PEG, *Intellimetrics* e *E-rater*) observando-se frequentemente correlações entre os resultados obtidos por esses programas e os juízes humanos da ordem ou acima de 0.8.

Em Attali e Burstein (2004), Chodorow e Burnstein (2004), Rudner, Garcia, e Welch (2006), Bejar (2011) e Bridgeman, Trapant, e Attalli (2012) são apresentados estudos de validade que, basicamente, confirmam as perspetivas apresentadas em Shermis

e Burstein (2003), confirmando assim a validade (referida a este critério) crescente dos programas de classificação automática de textos.

Em particular, Chodorow e Burnstein (2004) apresentam evidências convincentes relativas à aplicação do *E-rater*® do ETS na classificação do teste TOEFL (*Test of English as a Foreign Language*), notando a sensibilidade dos resultados da classificação automática às linguagens nativas (Espanhol, Árabe e Japonês), comportamento que é semelhante ao observado com juízes humanos.

Em Attali e Burstein (2004), num estudo de validação de uma nova versão do *E-rater*® (*K*2.0) conclui-se, entre outras coisas, que o comprimento do texto (em número de palavras) é uma variável que contribui significativamente para as classificações holísticas obtidas.

Uma das conclusões mais importantes do estudo citado é o valor 0.93 entre os resultados obtidos com juízes humanos e os obtidos com o programa *E-rater*.

Existem óbvias relações entre qualidade e validade. Em Bejar (2011) analisam-se estas relações sendo identificados os fatores a considerar nos processos de garantia e controlo de qualidade dos SAAT's de modo a melhorar a validade dos resultados a obter com esses sistemas.

A título de exemplo cita-se a necessidade de uma interface homem-máquina adequada, que garanta que os examinandos expressam as suas respostas sem estarem sujeitos a constrangimentos desnecessários.

Em Bridgeman, Trapant, e Attalli (2012) num estudo do ETS abrangendo textos produzidos no contexto da aplicação do TOEFL, GMAT (*Graduate Management Admission Test*), GRE (*Graduate Record Examination*) e avaliados tanto por juízes humanos como pelo *software E-rater*, abrangendo centenas de milhares de testes e milhares de estudantes, confirmam-se os elevados valores de fiabilidade e validade obtidos pelos programas de classificação automática de textos.

Especificamente, no estudo envolvendo dados gerados pelo teste TOEFL, usando testes com a duração de 30 minutos, pontuados tanto pelo *software E-rater* como por dois juízes humanos e envolvendo 132347 textos (num máximo de 300 palavras) conclui-se que as correlações entre dois juízes humanos (HH) variam entre 0.61 e 0.70; as correlações entre as classificações atribuídas pelo sistema automático, o *E-rater* e um juiz humano oscilam entre 0.64 e 0.78 quando se considerou o efeito da língua de origem das pessoas testadas.

De um modo geral, a conclusão do estudo é a de que, confirmando estudos anteriores, embora a língua de origem não produza efeitos muito marcados entre os resultados atribuídos pelos seres humanos e pelo programa de classificação, notou-se que para certos grupos linguísticos as ordenações resultantes das classificações (e portanto com impacto na vida das pessoas) seriam diferentes conforme as classificações fossem produzidas pela máquina ou pelos juízes humanos. Este impacto tende a desaparecer quando o resultado final resulta de uma combinação das duas classificações: humana e automática.

De acordo com a informação recolhida na literatura que é possível consultar no *site* do ETS - *Educatiuonal Testing Service* (2013) e apesar dos estudos que consistentemente apontam para elevados índices de fiabilidade e validade na utilização dos SAAT's, continua a ser política daquela instituição, nos exames de alta responsabilidade (*high-stakes*) usar a classificação simultânea por um programa e por um juiz humano. Quando há discrepâncias significativas entre as duas classificações isso implica a adoção de um procedimento especial que em geral consiste em classificar o teste por um segundo juiz humano.

1.4.6. Experiência Acumulada com a Avaliação Automática. Tendências.

Neste número procura-se sintetizar a experiência colhida com a utilização de *software* para avaliação automática, identificar e caraterizar algumas tendências que se julga detetar.

Aspetos Metodológicos.

De um modo geral, em todos os sistemas considerados (PEG, IEA, *Intellimetrics*, *E-rater*, *Betsy*) nota-se que apesar das óbvias diferenças na tecnologia usada (regressão múltipla, inteligência artificial, redes bayesianas, análise da semântica latente) todos os sistemas são implementados de acordo com a metodologia conhecida como modelo estatístico de aprendizagem (*Learning Machine*).

Definido o algoritmo teórico para atribuir pontuações aos textos elaborados pelos estudantes (redação ou "*essays*") torna-se necessário uma fase de aprendizagem inicial em que são estimados os parâmetros do modelo, usando para o efeito informação anterior.

Esta informação tem a seguinte origem: textos anteriormente classificados por juízes humanos (em todos os casos), textos do domínio do conhecimento que se pretende

avaliar, escritos por *espertos* (professores, autores de livros e manuais) no caso da ASL e ainda outra informação – nomeadamente informação linguística.

No caso de textos anteriormente classificados por juízes humanos, esta informação de treino é designada por "amostra de treino (AT)".

Uma vez "treinado" o algoritmo com estes dados de treino (estimados os respetivos parâmetros), obtém-se uma versão provisória do sistema capaz de realizar classificações automáticas de novos textos. Contudo, isso nunca é feito sem validar o sistema sobre a chamada "amostra de teste (T)". Trata-se também de testes que anteriormente foram classificados por juízes humanos e que agora vão ser também classificados – automaticamente – pelo sistema já "treinado" que resultou da primeira fase. Ver **figura 1.4.6.1.**

No esquema a seguir apresentado – comum aos sistemas estudados – a amostra inicial de textos classificados, com $n= q+p$ textos é dividida em duas partes (A + B), ficando a primeira (A) com q testes e a segunda (B) com p. Por exemplo, uma possibilidade é $q= p = n/2$.

Os $n= q+p$ textos da amostra inicial foram previamente classificados por juízes humanos que lhes atribuíram as classificações $H= (h_1, h_2, \ldots, h_q, h_{q+1}, \ldots, h_{p+q})$.

Suponhamos que o modelo matemático usado para atribuir automaticamente uma classificação a um texto genérico T é dado por uma expressão matemática f (texto, θ) que envolve o texto e certos atributos observados do texto (número de palavras, número de pontos, as palavras do texto, …) e um conjunto de parâmetros ou pesos $\theta= (\theta_1, \theta_2, \ldots, \theta_k)$ cujos valores são desconhecidos e necessitam de ser estimados (ou aprendidos) a partir dos q textos pré-classificados.

Só para fixar ideias, admita-se que $\theta= (\theta_1, \theta_2, \theta_3)$ – com $k= 3$ – são as distâncias de um texto a classificar aos textos classificados mais próximos (num certo sentido a definir) e que a fórmula

f (texto, $\theta = (\theta_1, \theta_2, \theta_3)$) é

$s = \theta_1 d_1 + \theta_2 d_2 + \theta_3 d_3$ com $\theta_1 + \theta_2 + \theta_3= 1$.

Isto é, os parâmetros a estimar, neste caso, seriam os pesos (de soma 1) pelos quais seria necessário multiplicar as distâncias observadas aos textos mais próximos para obter

um valor aproximado da classificação atribuída pelo juiz humano ao texto T tendo em conta a atribuição h_1, h_2, h_3 aos três testes mais próximos.

Usando a informação da amostra de treino (A) os valores $\hat{\theta}=(\hat{\theta}_1,\hat{\theta}_2,\hat{\theta}_3)$ são estimados a partir dos valores h_1, …, h_p atribuídos pelos juízes humanos a testes anteriormente classificados.

O resultado é um classificador "treinado" mas que seria imprudente usar de modo operacional sobre novos textos a classificar.

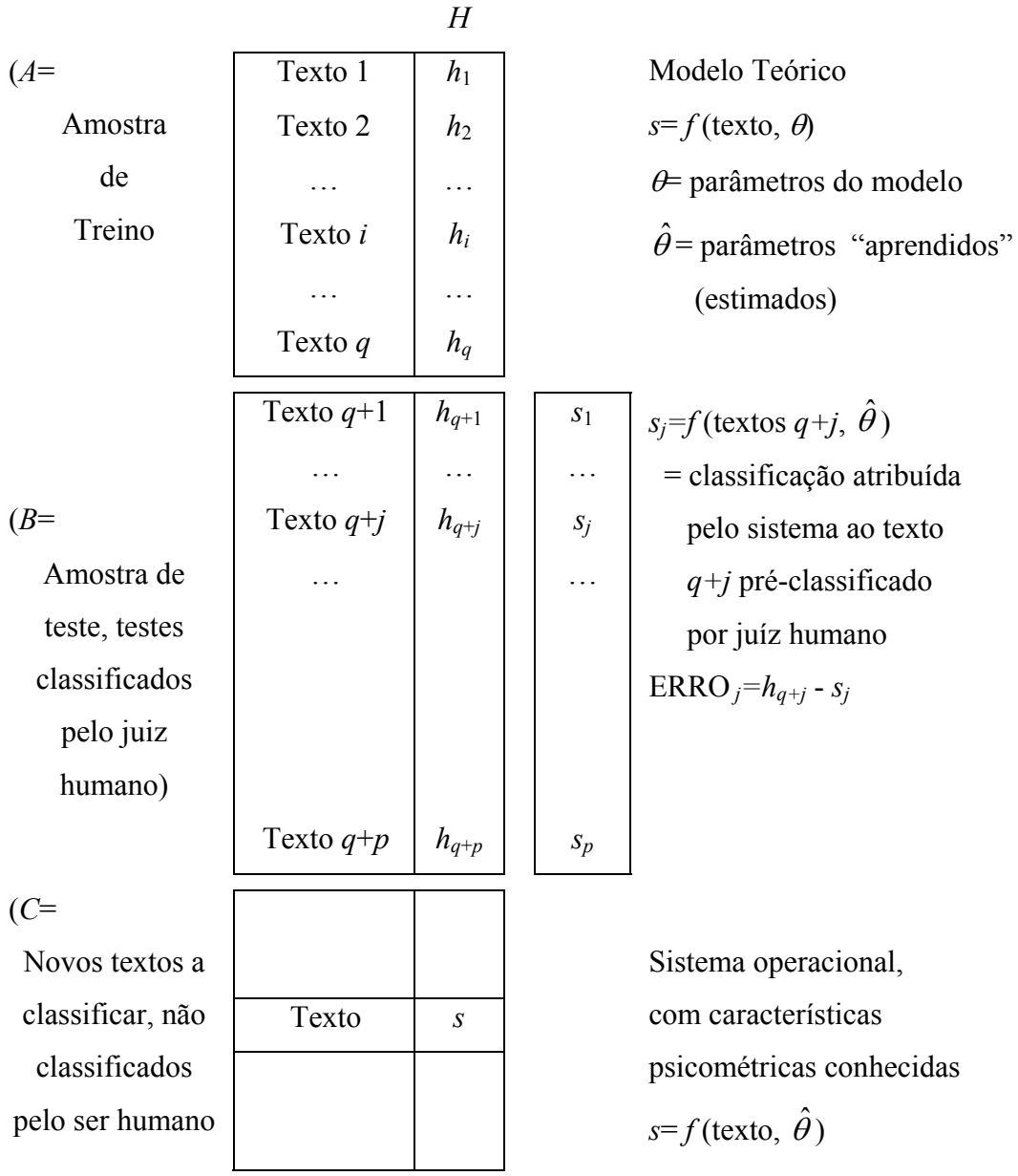


Figura 1.4.6.1. Esquema de treino de um sistema de classificação automática de textos.

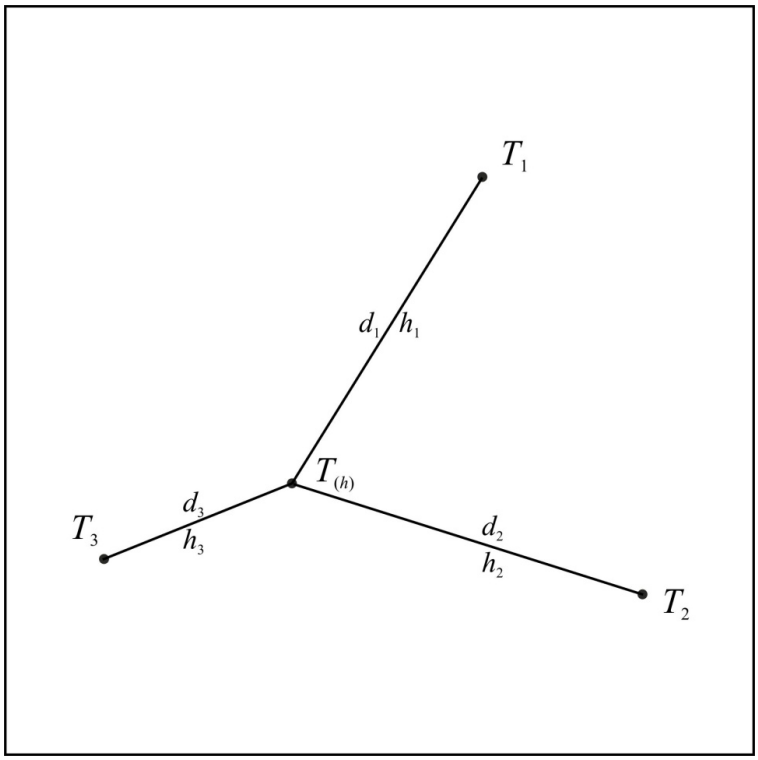


Figura 1.4.6.2. Pretende-se, neste caso, explicar a classificação h a atribuir (mediante o modelo) ao texto T em função das k= 3 distâncias (d_1, d_2, d_3) desse texto aos 3 vizinhos mais próximos T_1, T_2, T_3 aos quais os classificadores humanos atribuíram previamente as classificações h_1, h_2, h_3.

Usa-se então a 2ª parte da amostra de treino (B) – a amostra de teste – para obter estimativas do erro de classificação do classificador treinado e estimar caraterísticas psicométricas (fiabilidade e validade) do classificador.

Se T_{q+j} for o texto número j da amostra de teste (B), a que um juiz humano atribuiu a classificação h_{q+j} (j= 1, …, p) então podemos comparar os valores h_{q+1}, h_{q+2}, …, h_{q+p} com as pontuações atribuídas automaticamente pelo sistema já treinado (ver **figura 1.4.6.1.**):

s_1, s_2, …, s_j, …, s_p, obtendo-se os erros

$$e_j = \left(h_{q+j} - s_j\right) \; j = 1, \ldots, p.$$

Estes erros permitem desde logo obter uma estimativa da taxa de erro de classificação do sistema treinado e estimar também a validade referida ao critério: "comparação com a classificação do juiz humano", usando para isso o coeficiente de correlação entre as classificações atribuídas automaticamente e as obtidas do classificador humano (r_{hs}).

Quando o erro de classificação e este índice de validade forem aceitáveis, poderá então encarar-se a possibilidade do uso operacional do classificador sobre textos novos cuja classificação se desconhece. Veja no **Capítulo III** resultados obtidos com o nosso sistema ao ser aplicado a dados reais.

Esta metodologia tem óbvias limitações na sua aplicação ao contexto do apoio às atividades de um professor isolado, dada a necessidade de classificar previamente uma fração significativa dos testes. Valores usuais apontam para mínimos de 300 ou mais testes. Poderia pensar-se que se trata de um modelo que só tem interesse para grandes instituições como por exemplo os Ministérios de Educação dos países ou projetos em que estejam envolvidos milhares de testes. Contudo, os dados experimentais do **Capítulo III** resultantes da aplicação a textos produzidos no âmbito de uma avaliação contínua mostram resultados animadores neste sentido.

Questões Psicométricas

Como se viu em números anteriores tanto a fiabilidade como a validade dos sistemas automáticos de classificação de textos atingem, de acordo com a literatura disponível, valores muito elevados. Se *H* significar o juiz humano e *S* significar um sistema automático de classificação, as correlações *HS* superam frequentemente os valores *HH*. A literatura aponta frequentemente – ver, por exemplo, Burnstein (2003) ou Rudner, Garcia, e Welch (2006) – valores *HS* da ordem de 0.9 ou mais. Persistem contudo alguns problemas que recomendam uma atitude prudente quanto à utilização generalizada e sem restrições destes sistemas – principalmente em exames sumativos nacionais, com grandes implementações para a vida das pessoas.

De um modo geral pode dizer-se que, do ponto de vista experimental e de acordo com os estudos publicados pelo ETS (*Educational Testing Service*) em particular, é atualmente difícil distinguir as classificações produzidas pelos sistemas automáticos e pelos juízes humanos quando apenas estão envolvidos os chamados traços superficiais (por oposição aos traços profundos subjacentes aos textos).

Utilização Operacional dos Sistemas de Avaliação Automática de Textos (SAAT)

Antes de mais deve-se pôr em evidência que toda a investigação experimental e utilização operacional dos SAAT's incidem sobre textos produzidos por examinandos

utilizando meios informáticos, excluindo-se textos manuscritos posteriormente transcritos ou lidos por sistemas de leitura ótica de caracteres manuscritos.

Apesar de toda a investigação experimental e operacional apontar tendencialmente para a indistinção – segundo critérios da fiabilidade e validade – entre as pontuações atribuídas por sistemas SAAT e juízes humanos, constata-se que, um organismo como o ETS (*http://www.ets.org/research/contact.html* (consultado em 26-04-2013) não admite ainda a utilização exclusiva destes sistemas em provas a nível nacional. As razões podem ver-se em Williamson, Xi, e Breyer (2012) e Zhang (2013).

Este organismo distingue assim entre testes de alto risco (*high-stakes tests*) e testes de baixa responsabilidade (*low-stakes tests*).

Os testes de alto risco são os que estão associados a consequências muito importantes para a vida dos examinandos e das sociedades, como por exemplo exames nacionais em que o risco de uma atribuição errada ou enviesada de classificação pode ter consequências desastrosas para a vida das pessoas e funcionamento da sociedade.

Para este tipo de teste admitem-se várias soluções mas nunca a utilização exclusiva dos SAAT's (Williamson, et al., 2010; Williamson, Xi, & Breyer, 2012; Zhang, 2013).

Para os testes de baixo risco, como por exemplo os testes de diagnóstico ou os testes para controlo da avaliação contínua admite-se a utilização dos SAAT's.

As referências Williamson, Xi, e Breyer (2012) e Zhang (2013) são especialmente reveladoras da situação presente e das tendências que se estão a manifestar. Se por um lado se reconhecem inegáveis vantagens dos SAAT's, já constatadas experimental e operacionalmente (como objetividade, consistência, reprodutibilidade, explicação dos resultados), por outro lado, na avaliação dos traços latentes subjacentes aos textos (aspetos de estilo, conteúdo, pensamento crítico) os avaliadores humanos têm vantagem sobre os SAAT's, se bem que sejam por vezes altamente inconsistentes, sujeitos ao efeito *halo* subjetivos, influenciáveis.

Destas considerações resultam implicações práticas relativas à utilização dos SAAT's em testes de alto risco. Para estes testes, o ETS usa simultaneamente os SAAT's e o juiz humano, segundo duas alternativas (Williamson, Xi, & Breyer, 2012; Zhang, 2013).

1ª Alternativa

A classificação final baseia-se numa combinação da classificação *S* atribuída pelo SAAT e da classificação *H* atribuída pelo juiz humano. Quando a diferença entre as duas classificações excede um certo limiar e – isto é, quando $|S - H| > c$ – então é chamado um novo juiz humano.

A classificação final é:

$c = w_1 \times S + w_2 \times H$

em que $w_1 + w_2 = 1$. Por exemplo, $w_1 = w_2 = ½$.

2ª Alternativa

A classificação final é atribuída por um classificador humano (*H*). A pontuação (*S*) do SAAT é usada apenas para controlo de qualidade.

Ver Zhang (2013) e Williamson, Xi, e Breyer (2012) para outras alternativas.

Tendências

A literatura recente – ver, por exemplo, Zhang (2013) – coincide na ideia de que num futuro próximo se assistirá ao uso generalizado dos SAAT's para a classificação dos testes sejam eles de alto risco ou de baixo risco (sumativos ou formativos). A situação atual (uso dos dois sistemas) tem mais a ver com questões psicológicas, sociais e políticas do que com questões de validade.

No modelo do "*bag of words*" em que se baseia a ASL, estudos citados em Landauer, McNamara, Dennis, e Kintsch (2007) atribuem 80% do significado dos textos ao significado das próprias palavras e apenas 20% à ordem das mesmas e a outros aspetos como a estrutura das frases. Contudo, estudos mais recentes relacionados com os trabalhos de Kintsch (2001) e Dennis (2005), citados na referida obra, mostram que o modelo ASL pode ser generalizado de modo a abranger esses aspetos.

1.5. EDM - Educational Data Mining.

O conceito de *Educational Data Mining* (EDM) que no que se segue traduziremos por Mineração de Dados Educacionais (MDE) é a manifestação nas Ciências de Educação de um fenómeno global ligado à acumulação, em bases de dados digitais e outros repositórios menos organizados, de enormes quantidades de dados.

Estes dados têm naturezas muito diversas – desde dados científicos resultantes da observação organizada e planeada da natureza até dados resultantes do funcionamento normal das organizações de qualquer natureza.

Os dados representam factos expressos por números, textos, imagens e sons.

De acordo com estudos recentes – ver, por exemplo, Manyika, et al. (2011) – esta acumulação longe de abrandar tende a acelerar, tornando muito difícil, mesmo face às capacidades atuais de análise, proceder à extração atempada de informação útil à tomada de decisão.

Este facto tem contribuído não só para uma alteração significativa da investigação científica – parte da qual incide hoje sobre os dados acumulados em bases de dados – e para a criação de possibilidades novas como a investigação em Ciências da Educação (Romero, et al., 2011).

Num artigo gerado na área da crítica literária (Moretti, 2005) constata-se que, face à taxa de produção literária – em particular em romances de língua inglesa – é praticamente impossível a um especialista dessa área ler até aquele relativamente pequeno número de livros (face ao volume da produção) que lhe permitiria manter-se minimamente atualizado. Mesmo que dedicasse todo o tempo disponível a ler novas obras ocorridas num só ano, esse especialista necessitaria de centenas de anos de trabalho para ler só os trabalhos desse ano.

Daí a constatação natural (Moretti, 2005) de que só é possível exercer a atividade de crítico literário da área em questão com apoio de programas de gestão de dados textuais e de programas de análises estatísticas destes textos.

Segundo Moretti (2005), o apoio necessário não é apenas o inerente às contagens e cálculo de estatísticas básicas (quem, quando, onde, quantos textos, de que temas, etc…) mas passa também pelas sínteses automáticas dos conteúdos, feitas com a rapidez compatível com as imposições da atividade. Isto é, implica a "leitura automática" de milhares de textos em curtos períodos de tempo de modo a permitir sintetizar o respetivo conteúdo, classificar esses textos em categorias homogéneas relevantes para a atividade de crítico e que permitam, eventualmente, escolher ao acaso um pequeno número de textos que o crítico literário possa ler "pessoalmente".

De um modo geral *Data Mining* (mineração de dados) (Fayyad, Piatetsky, Shapiro, Smyth, & Uthurusamy, 1996) consiste na aplicação de algoritmos (sumarização, classificação, regressão e outros) para a extração de padrões a partir dos dados (factos).

Este processo de mineração de dados integra o processo de extração de conhecimento a partir da base de dados e que consiste, segundo o mesmo autor, na descoberta de informação útil em base de dados. O que fica para lá da mineração de dados (*data mining*) são as tarefas de interpretação, em domínios de atividade específica, dos padrões ou regularidades descobertas pela aplicação das técnicas computacionais e estatísticas.

A mineração de dados (*text mining*) é um caso particular destas atividades quando os dados são textos (sequências de caracteres). Ver, por exemplo, Feldman e Sanger (2007).

O conceito de "*big data*" – Manyika, et al. (2011) – significa pois o reconhecimento do problema do crescimento exponencial da informação digital e da quase impossibilidade, com os meios existentes (*software* e *hardware*) de extrair dessa montanha de dados informação útil e adequada à tomada de decisão.

O conceito de "*big data*" tem implícito o reconhecimento de que se torna necessário desenvolver com urgência novos algoritmos, meios de análise e teoria que permitam extrair dessas montanhas de dados, conhecimento válido para a tomada de decisão em tempo útil e modelos de predição de eventos futuros assentes nesse conhecimento empírico – o que corresponde ao conceito mais recente de "analítica" (*analitics*) abrangendo a conceção, validação e operacionalização de modelos de predição combinando estatística, informática e modelos matemáticos (investigação operacional) e outras ciências do domínio ou domínios envolvidos nos dados, tendo por objetivo produzir predições empíricas (baseadas nos dados) e métodos que permitam avaliar a qualidade dessas predições (Shmueli & Koppins, 2011).

É no contexto atrás sintetizado que surge o conceito de EDM, por volta de 1995. Ver Baker, Ryan, e Yacef (2009) que procuram fazer o estado da arte deste tema em 2009.

Basicamente EDM é a mineração dos dados resultantes da atividade do ensino, o que justifica a tradução por Mineração de Dados Educacionais (MDE). Estes dados podem ter origem na sala de aula (com a observação da aula, resultados de testes formativos, por exemplo), na interação de estudantes com *software* educativo como as plataformas de *e-*

leaning (*Moodle*, por exemplo), nos resultados de testes sumativos a nível nacional ou local, nos dados gerados pela atividade de gestão de sistemas de ensino, entre outras fontes.

Como se observa através destes exemplos, os dados referidos não são recolhidos com o fim de apoiar a investigação nas ciências da educação. Resultam simplesmente do funcionamento normal dos sistemas existentes, na prossecução dos respetivos objetivos. Contudo – à semelhança do que está a ocorrer em quase todas as áreas de atividade científica ou não – o facto é que estes dados permitem não só extrair conhecimento útil à gestão como – é este o facto novo – realizar investigação empírica sobre os sistemas de ensino e alicerçar em dados objetivos novos modos de encarar o ensino, redefinir o papel dos professores e os modos de avaliação – esbatidas, muitas vezes, estas distinções (Rupp, Nugent, & Nelson, 2012).

Embora possa parecer questionável o estabelecimento de uma especialização da atividade de mineração de dados ligada exclusivamente aos dados da educação – com o argumento de que estes têm características específicas como a natureza hierárquica e ausência de independência – o facto é que o EDM tem-se desenvolvido rapidamente e tem servido para agregar investigadores e a investigação, antes dispersa, nesta área.

Os resultados desta atividade de investigação estão acessíveis em repositórios digitais como os do *site* do *Journal of Educational Data Mining* – http://www.educationaldatamining.org/JEDM/index.php consultado em 14 de Maio de 2013 – e também os artigos publicados nos *proceeding* dos sucessivos congressos anuais. Ver o *site* http://www.educationaldatamining.org/ consultado em 12 de Maio de 2013.

Passada uma fase inicial em que se tratou principalmente de aplicar as técnicas tradicionais de *data mining* a dados do domínio das ciências da educação, assiste-se agora ao aparecimento de investigação em que são explorados aspetos e problemáticas específicas deste domínio. A título de exemplo cita-se Antunes (2010).

Uma área que está a receber atenção crescente tem a ver com os aspetos motivacionais no âmbito da aprendizagem e cujo tratamento pode ser objeto de abordagens por vezes surpreendentes e só possíveis no ambiente de superabundância de dados atual. Como por exemplo, em D'Mello e Graesses (2010) em que é relatada uma experiência destinada a investigar aspetos emocionais e afetivos da interação dos estudantes com uma plataforma de ensino eletrónica. Em vez do estudo das expressões faciais dos estudantes ao

interatuarem com as diversas partes do curso, os investigadores referidos obtêm resultados interessantes estudando os registos da postura do corpo expressa pela distribuição da pressão exercida pelo corpo do estudante sobre o assento e as costas da cadeira usada, mediante sensores de pressão especiais e o registo simultâneo dos ecrãs correspondentes do curso. Foi assim possível relacionar atitudes de entusiasmo e aborrecimento, por exemplo, com as partes relevantes do curso, com óbvias consequências para a melhoria do mesmo.

A superabundância de sensores de baixo custo, permitindo o registo das mais diversas variáveis e guardar os dados destas medições em bases de dados, bem como a disponibilidade crescente de metodologia estatística que permite o tratamento dos dados gerados de forma a autorizar inferências válidas sobre o comportamento dos estudantes, está na base de um desenvolvimento muito interessante – e sob certos aspetos surpreendente – na área do MDE. Trata-se do conceito de *Evidence Centered Design* (ECD) que poderíamos traduzir por Conceção de Sistemas de Avaliação Centrados na Evidência, atendendo ao contexto em que foi criado e desenvolvido, no ETS (Mislevy, 1992; Mislevy, Almond, & Lukas, 2003; Rupp, Nugent, & Nelson, 2012).

O conceito ECD – mantendo a sua designação original em inglês – aparece já, implicitamente, em Mislevy (1992) no âmbito de um estudo realizado pelo ETS relacionado com a questão da comparação de resultados (classificações) obtidos por grupos diferentes, avaliados por métodos distintos de avaliação e medição.

Segundo Mislevy, Almond, e Lukas (2003), o ECD pode ser descrito como uma abordagem empírica (assente nos dados) à questão da conceção e implementação de sistemas de avaliação (*assessment*). Os referidos autores constatam que o que há de comum entre todos os sistemas de avaliação em educação é "*o desejo de raciocinar a partir de coisas particulares que os estudantes dizem ou fazem para realizar inferências acerca dos seus conhecimentos e competências*".

É o que sucede num contexto limitado, por exemplo, no âmbito do TRI em que se procura inferir o comportamento dos itens e dos estudantes observando casos particulares de respostas a itens integrados em testes já aplicados (Bergner, Dröschler, Kortmeyers, Rayyan, Seaton, & Pritchard, 2012).

Estes esquemas clássicos de avaliação têm sido considerados, cada vez mais, inadequados para avaliar a evolução dos estudantes – principalmente face a avanços na psicologia e ciências da educação. Este facto é tanto mais evidente face ao avanço

tecnológico, nomeadamente quanto à possibilidade de observar aspetos inacessíveis aos testes clássicos e de tratar os dados resultantes da interação dos estudantes com sistemas de *e-learning* e da monitorização do processo de aquisição de conhecimentos usando técnicas que combinem modelos de decisão, psicometria, biometria e mineração de textos e outros dados.

Capítulo II: Metodologia da Investigação.

2.1. Introdução.

Como se disse na **Introdução à Tese**, a principal questão que este trabalho procura investigar tem a ver com a identificação e avaliação dos problemas associados ao desenvolvimento de instrumentos de ensino / aprendizagem, a serem usados por professores e estudantes, baseados em análises estatísticas de textos.

De entre esses problemas, ocupam papel central as questões ligadas à monitorização, tanto por professores como por estudantes, dos processos de aquisição de conhecimentos naquelas matérias em que a língua nativa – o português em particular – detém a maior importância. Em particular, as questões ligadas às situações em que os estudantes expressam em português as respostas a questões abertas.

Se bem que durante a elaboração deste trabalho tenham sido identificados importantes projetos de investigação ligados à língua portuguesa – veja-se, por exemplo, a investigação associada à atividade do Instituto de Linguística Teórica e Computacional, ILTEC (*http://www.iltec.pt/* consultado em 22 de Maio de 2013), os trabalhos do grupo NLX - *Natural Language and Speech Group* da Faculdade de Ciências da Universidade de Lisboa (*http://nlx.di.fc.ul.pt* – consultado em 22 de Maio de 2013) ou os trabalhos de outros investigadores como Cláudia Antunes, ligada ao Instituto Superior Técnico (*http://web.ist.utl.pt/claudia.antunes/* consultado em 22 de Maio de 2013) – constata-se que na área que nos ocupa (a da aquisição e avaliação de conhecimentos usando o português como instrumento) relativamente poucos trabalhos têm sido publicados

Dentro desta vasta área concentrámo-nos no problema da análise dos textos produzidos por estudantes em resposta a questões abertas, questão que está na interseção de muitos dos problemas antes identificados.

A preocupação inicial foi a de entender a razão pela qual a Análise da Semântica Latente (ASL) aparece na literatura técnica desta área como modelo natural de aquisição da linguagem. Para isso começou por se estudar o artigo Landauer e Dumais (1997), relativo ao chamado "*Problema de Platão*" que designa a aparente contradição entre a rapidez com que as crianças obtêm novos termos face à relativa pouca quantidade de informação que recebem durante esse processo.

Feito isto, procurou-se em seguida reproduzir, no domínio do Português, experiências relatadas na literatura para outras línguas e que dão conta da possibilidade de, usando a metodologia da ASL, construir sistemas de avaliação automática de conhecimentos com base nessa metodologia.

Neste sentido considera-se muito útil o acesso às facilidades do sítio da Universidade do Colorado (*http://lsa.colorado.edu/*) em que elementos do grupo de investigação ligado a Landauer têm publicado os resultados dessa investigação, sendo ainda possível utilizar "*on-line*" um sistema *Intelligent Essay Assessor* (IES) que permite avaliar textos submetidos por utilizadores.

Contudo, sendo o nosso objetivo o de estabelecer a possibilidade de construir sistemas utilizáveis, de modo rotineiro, por professores e estudantes portugueses com recursos muito limitados, decidimos construir um protótipo – descrito no que se segue – que permitisse, com total flexibilidade, responder às questões formuladas.

No que se segue ao longo deste capítulo, descreve-se, sucessivamente, em **2.2.** o modelo que serviu de base à implementação do protótipo usado na experimentação, em **2.3.** as decisões tomadas relativas à estruturação de uma amostra que permita validar o modelo e realizar as experiências que permitiram responder às questões colocadas e, em **2.4.**, definir a metodologia para recolher os dados. Finalmente, em **2.5.** descreve-se a metodologia geral a usar para analisar os dados nas diversas experiências a realizar com uma implementação desse modelo.

2.2. Formulação do Modelo.

2.2.1. Estrutura.

Neste número descreve-se o modelo usado como instrumento desta investigação e com base no qual foi concebido o *software* Programa de Análise Estatística de Textos (PAET) descrito em **3.4.** e objeto do Manual do Utilizador que constitui o Anexo A deste trabalho.

A **figura 2.2.1.1.** descreve a estrutura do sistema que permite, usando os textos dos professores e manuais escolares, produzir os chamados espaços semânticos, instrumento fundamental da Análise da Semântica Latente descrito em **2.2.2.**

A **figura 2.2.1.2.** descreve a estrutura do sistema experimental a desenvolver na avaliação de conhecimentos de uma certa matéria com base em textos produzidos pelos estudantes ao responderem a questões de resposta aberta e nos espaços semânticos construídos a partir dos textos dos manuais usados no ensino desse tema.

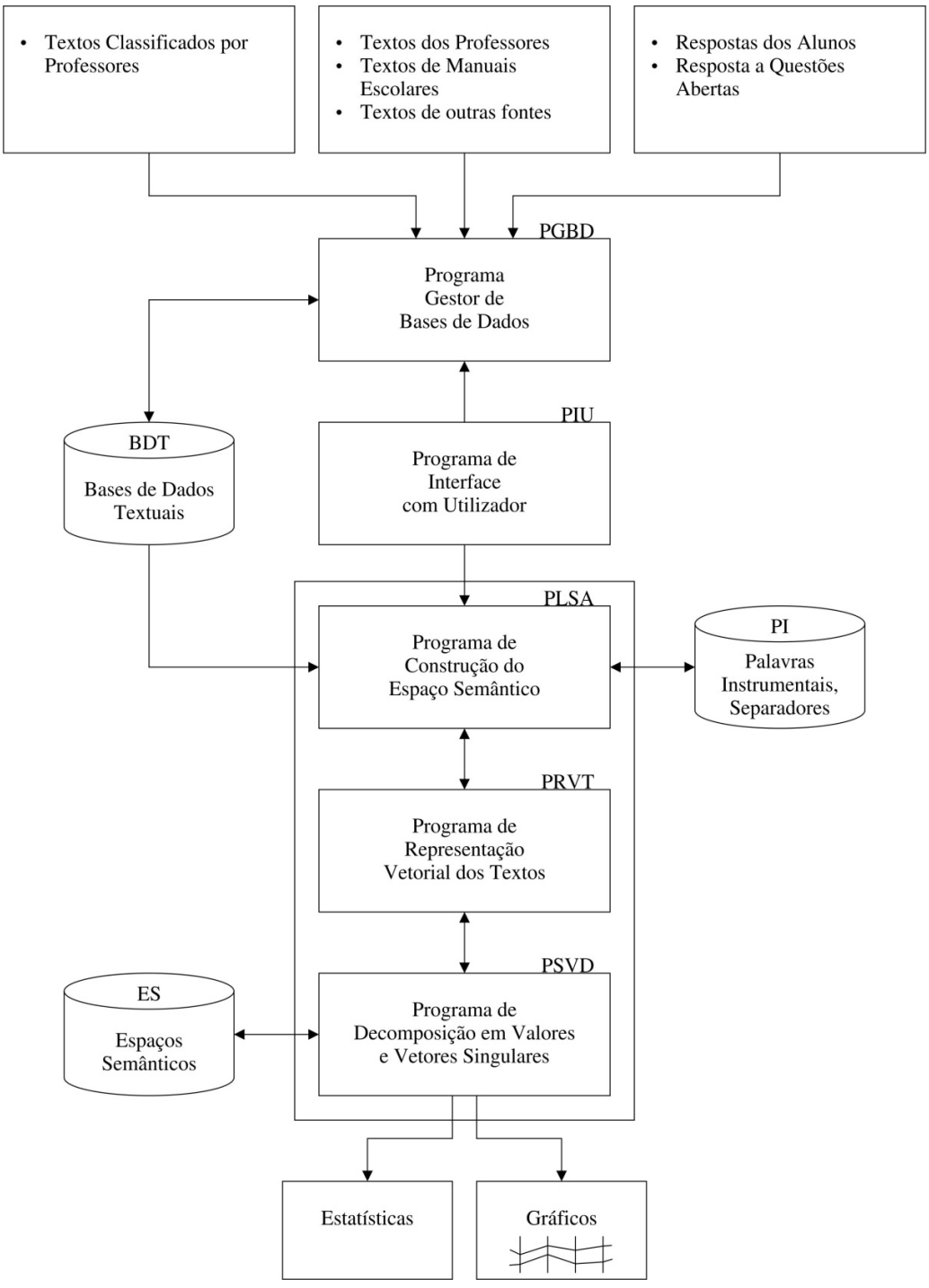


Figura 2.2.1.1. Armazenamento da informação de base e construção de espaços semânticos a partir de textos produzidos pelos professores e manuais escolares relativos ao ensino de certas matérias.

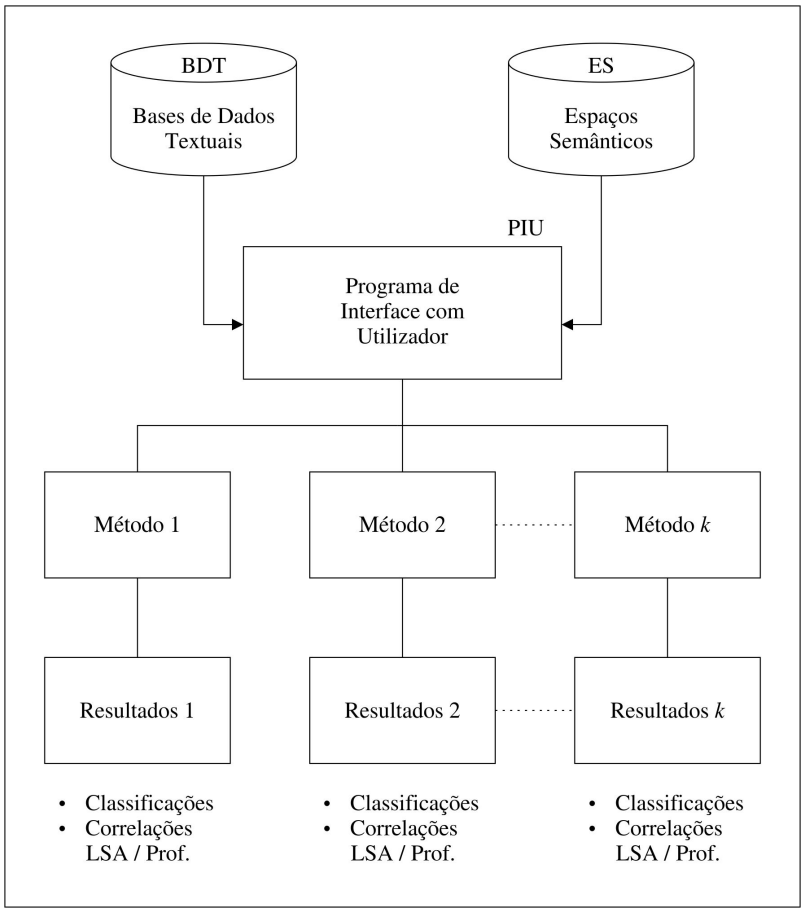


Figura 2.2.1.2. Estrutura do sistema de avaliação de conhecimentos com base em textos de resposta a questões abertas.

Os significados dos símbolos da **figura 2.2.1.1.** são as seguintes:

PGBD – Programa de Gestão da Base de Dados

BDT – Base de Dados Textual

PIU – Programa de Interface com o Utilizador

ES – Espaços Semânticos

PLSA – Programa de Construção do Espaço Semântico Latente (ASL)

PRVT – Programa de Representação Vetorial dos Textos

PSVD – Programa de Decomposição em Valores e Vetores Singulares

Os textos relevantes para uma análise são os seguintes:

- Manuais e outros documentos recomendados para a aprendizagem da matéria em avaliação.

  Exemplo: Manuais escolares usados no ensino do Português e obras de autores portugueses usados num certo nível de ensino.

- Textos – apontamentos produzidos pelos professores de uma certa matéria.

- Textos com as respostas dos estudantes aos itens de resposta aberta, contidos nos testes de exames globais ou formativos.

  Exemplo: Resposta dos estudantes do 12º ano a itens de resposta aberta dos exames nacionais.

- Textos com respostas dos estudantes a itens de resposta aberta classificados pelos professores usando os métodos tradicionais. Estes textos encontram-se em suportes de papel ou digitais – por exemplo, em textos produzidos com processadores de texto, em plataformas de ensino ou em formato *.pdf*, em sítios da Internet, entre outros. Exclui-se a utilização de programas de reconhecimento de carateres, Optical Character Recognition (OCR), pelo que todos os textos manuscritos têm de ser previamente digitalizados antes da sua introdução no sistema ou então deixados de fora, face ao custo dessa operação.

Todos os textos relevantes são carregados numa BDT – Base de Dados Textual – através de um Programa de Gestão de Base de Dados (PGBD). Para a estrutura da base de dados veja o número **3.3.** Uma vez na BDT, todos estes textos ficam disponíveis para os diversos tipos de análise e produção de espaços semânticos especializados.

As análises podem consistir, simplesmente, em contagens dos termos que integram os textos, em cálculos de índices úteis a diversos usos ou então na criação de representações vetoriais dos textos e palavras (RVT) que sirvam de base à construção de Espaços Semânticos (ES) (Landauer, McNamara, Dennis, & Kintsch, 2007) em que assenta o funcionamento de sistemas de classificação dos textos.

A escolha dos textos específicos a serem usados numa certa análise bem como daqueles que vão integrar um dado espaço semântico a ser usado na avaliação de conhecimentos de certa matéria é realizado através do Programa de Interface com o Utilizador – o programa PIU.

Escolhidos estes textos, o programa PLSA começa por identificar todas as palavras existentes em todos os textos escolhidos para análise (usando um algoritmo de identificação de "*tokens*") elimina as palavras instrumentais ou funcionais (consultando a tabela PI) e, eventualmente, identificando / extraindo as raízes das palavras através de um programa de lematização (Rocha & Coelho, 2009; Orengo & Huyck, 2001, & Alvares 2005).

O resultado é uma tabela de frequências que, para cada "forma" – raiz das palavras retida para análise – dá a frequência de ocorrência dessa forma em cada um dos textos.

Esta tabela de frequências constitui a chamada representação vetorial dos textos (RVT), a usar no seguimento da análise. Trata-se de uma tabela de contingência contendo as frequências de ocorrência das palavras ou formas nos textos do *corpus*.

Esta tabela – de contingência – é em seguida sujeita a várias transformações, antes de ser submetida à decomposição em vetores e valores singulares através do programa PSVD (Eckart & Young, 1936; Landauer, et al., 2007).

O resultado da decomposição em valores e vetores singulares (depois de outras transformações) constitui o chamado Espaço Semântico (Landauer, et al., 2007).

A cadeia de transformações (acima resumidas) que permite passar de um certo conjunto de textos para o espaço semântico necessário às operações de treino dos avaliadores de conhecimentos envolve operações computacionalmente demoradas mas cujos tempos de execução estão a tornar-se cada vez mais reduzidos face à potência crescente dos computadores pessoais, os únicos que importa considerar neste projeto. O tempo de processamento depende, entre outras variáveis, do número de textos e da respetiva extensão expressa em número de palavras, que têm de ser identificadas através da operação de tokenização, uma das mais demoradas de todo o processo. Deste modo, convém guardar o resultado desta fase numa base de dados auxiliar de Espaços Semânticos (ES) – o que permite, no início de uma sessão, retomar o trabalho já realizado, sem ter de recalcular, de cada vez, o espaço semântico que interessa.

A **figura 2.2.1.2.** descreve, por sua vez, o processo que, partindo da BDT (Base de Dados Textual) e ES (Espaços Semânticos) permite obter classificações ou avaliações dos conhecimentos, acerca da matéria em avaliação, expressos nos textos das respostas dos estudantes a itens de resposta aberta contidas nos diversos testes. A partir destes espaços semânticos podem realizar-se, sobre as respostas aos itens usados nos testes, vários estudos

úteis à atividade do professor, como por exemplo comparações estatísticas dos textos produzidos pelos vários estudantes, agrupamentos de textos semelhantes, classificações dos testes, identificação de grupos de palavras usadas por grupos de estudantes e outros.

Se bem que estas comparações possam ser realizadas diretamente sobre as tabelas de contingência, considera-se que é mais útil realizá-las sobre os "espaços semânticos" uma vez que as operações de SVD permitem, como se disse, captar semelhanças semânticas que escapam completamente à tabela de frequências, ainda que presentes nessas tabelas de modo implícito. É isto o essencial da abordagem designada por ASL.

Um Programa de Interface com o Utilizador (PIU) permite que este não só identifique quais são os testes e os itens que vão ser objeto de avaliação como alguns parâmetros relevantes para essa avaliação, como sejam o Método a usar (Método 1, Método 2, …, Método *k*). De entre os parâmetros relevantes para uma experiência, destaca-se o número de dimensões *d* a usar nos cálculos subjacentes à classificação (Landauer & Dumais,1997; Landauer, Laham, & Foltz, 2003; Landauer, et al., 2007).

Os resultados da aplicação desses métodos expressam-se em estatísticas e em gráficos que dependem do método a usar.

No número **2.5.** – metodologia – podem ser vistos os pormenores de dois desses métodos implementados neste trabalho.

2.2.2. Obtenção da Representação Vetorial dos Textos (Tabelas Léxicas).

Dado um conjunto de textos a analisar estatisticamente (corpus), um passo preliminar consiste em obter a respetiva representação vetorial (Salton, Wong, & Yang, 1975; Lebart, Salen, & Berry, 1998; Osuna, 2006) sob a forma de uma tabela de contingência ou tabela léxica que cruza "formas" (palavras ou coocorrências de palavras) e textos ou documentos, tendo no cruzamento da forma *i* com o texto ou documento *j* a frequência f_{ij} com que a forma *i* ocorre no texto D_j. Ver **tabela 2.2.2.1.**

Formas	D_1	D_2	…	D_j	…	D_p	…
F_1							
F_2							
…							
F_i				f_{ij}			
…							
F_n							

Tabela 2.2.2.1. Representação vetorial de um corpus formado pelos documentos D_1, D_2, … D_p e pelo vocabulário formado pelas formas F_1, F_2, …, F_n (Osuna, 2006).

A obtenção da tabela de contingência ou tabela léxica – elemento essencial da análise estatística de textos – pressupõe a realização das seguintes tarefas prévias:

1. Identificar e definir as unidades de texto a considerar na análise - em geral palavras, isto é, sequências de símbolos delimitados à esquerda e à direita por espaços em branco. Mas podem ser outras formas; por exemplo, duas ou mais palavras consecutivas (coocorrência de palavras).
2. Eliminar todas os delimitadores que não sejam os espaços em branco: "?", "!", ".", ",", ".", "…" …
3. Eliminar as palavras funcionais ou instrumentais (PI) – preposições, pronomes, artigos, pronomes e outras, conforme o contexto do estudo.
4. Lematização. Isto é, substituição de palavras por palavras normalizadas, obtidas por extração das raízes comuns a um conjunto de formas, reduzindo assim o número de palavras do vocabulário e aumentando a frequência das formas retidas.

   A lematização (Rocha & Coelho, 2009; Orengo & Huyck, 2001; Alvares, 2005) implica usar uma "peça" de *software* específico: o lematizador. Existem lematizadores gerais sendo o ideal usar um lematizador específico para o português. Neste trabalho usa-se o lematizador descrito em Porter (1980) sendo escassa a literatura portuguesa acerca deste tema. Ver os trabalhos do grupo da Faculdade de Ciências de Lisboa NLX (http:// *nlx.di.fc.ul.pt/* consultado em 22 de Maio de 2013).

A fim de se poder seguir com maior facilidade os conceitos descritos neste número, e nos números seguintes, considere-se o seguinte exemplo simples, baseado em algumas respostas abertas à questão: "O que é a Análise da Semântica Latente (ASL)?"

Exemplo 2.2.2.1. (O que é a ASL?)

Considere-se as seguintes respostas (R_1 … R_5) de cinco estudantes à questão anterior.

R_1 – O conceito de ASL pode ser visto como a aplicação da SVD a uma tabela de contingência resultante da representação vetorial de um conjunto de textos.

R_2 – A ASL pode ver-se como uma metodologia capaz de representar geometricamente o significado de um conjunto de textos.

R_3 – O essencial do conceito de ASL está no facto de que a técnica ou metodologia em que se baseia permite expressar geometricamente – através de ângulos e distâncias – as proximidades psicológicas entre significados dos termos.

R_4 – Em termos psicológicos pode dizer-se que a ASL é um modelo das representações mentais do significado dos termos envolvidos num conjunto de textos.

R_5 – Se a ASL dá conta das associações e proximidades psicológicas existentes na mente das pessoas entre os significados dos termos e dos textos, então é difícil resistir à tentação de usar essa metodologia como instrumento de avaliação de conhecimentos. No fim de contas, o conhecimento que alguém porventura possua acerca de uma certa matéria deve poder ser expresso em termos de palavras – representando conceitos e sua relações.

Considerando um espaço semântico – obtido sem lematização – formado apenas pelos primeiros três textos (R_1, R_2, R_3), a tabela de contingência correspondente é a seguinte (matriz X):

ID	**Palavras**	R_1	R_2	R_3
1	angulos	*0*	*0*	*1*
2	aplicacao	*1*	*0*	*0*
3	atraves	*0*	*0*	*1*
4	baseia	*0*	*0*	*1*
5	capaz	*0*	*1*	*0*
6	como	*1*	*1*	*0*
7	conceito	*0*	*0*	*1*
8	conjunto	*1*	*1*	*0*
9	contingencia	*1*	*0*	*0*
10	distancias	*0*	*0*	*1*
11	entre	*0*	*0*	*1*
12	essencial	*0*	*0*	*1*
13	expressar	*0*	*0*	*1*
14	facto	*0*	*0*	*1*
15	geometricamente	*0*	*1*	*1*
16	metodolofia	*0*	*0*	*1*
17	metodologia	*0*	*1*	*0*
18	permite	*0*	*0*	*1*
19	pode	*1*	*1*	*0*
20	proximidades	*0*	*0*	*1*
21	psicologicas	*0*	*0*	*1*
22	representaçao	*1*	*0*	*0*
23	representar	*0*	*1*	*0*
24	resultante	*1*	*0*	*0*
25	significado	*0*	*1*	*0*
26	significados	*0*	*0*	*1*
27	tabela	*1*	*0*	*0*
28	tecnica	*0*	*0*	*1*
29	termos	*0*	*0*	*1*
30	textos	*1*	*1*	*0*
31	vectorial	*1*	*0*	*0*
32	ver-se	*0*	*1*	*0*
33	vista	*1*	*0*	*0*

Tabela 2.2.2.2. Tabela de frequência correspondente aos textos do **exemplo 2.2.2.1.**

Esta tabela, adiante designada por X, constitui a representação vetorial – tabela de contingência Palavras X Textos – que vai ser submetida às operações de decomposição em vetores e valores singulares, a seguir descritas.

2.2.3. Construção e Interpretação de Espaços Semânticos. Biplots.

A representação vetorial *X* das palavras versus textos – veja número anterior – é o ponto de partida para a construção do espaço semântico correspondente (Landauer, Laham, & Foltz, 2003; Landauer, et al., 2007).

O espaço semântico resulta da decomposição em valores e vetores singulares da matriz *X* que contém as frequências das palavras (formas) nos textos do corpus, depois de eliminadas as palavras funcionais e do processo de lematização, não obrigatório.

Observe-se que cada linha da matriz *X* é um vetor (de frequências) correspondente a uma palavra (ou forma), num espaço com tantas dimensões quantos os textos do *corpus*. Isto é, X é uma matriz $X\,(n, p)$ de n linhas por p colunas.

Por exemplo, no caso descrito no número anterior, a palavra "*conjunto*" fica representada pelo vetor [1, 1, 0] significando que ocorre uma vez tanto no texto da resposta R_1 como no texto da resposta R_2. Isto é, cada palavra fica representada num espaço vetorial de dimensão p= número de textos do corpus. Neste caso, p= 3.

De modo semelhante, cada texto, por exemplo, o texto da resposta R_3 fica representado por um vetor de frequências de ocorrência das diversas palavras, espaço este que tem tantas dimensões quanto as palavras (depois da lematização) que ocorrem no conjunto dos textos. No caso, cada texto (R_1, R_2, R_3) fica representado por um vetor de 33 componentes ou dimensões.

Dada a matriz da representação vetorial dos textos *X* com n linhas por p colunas, contendo no cruzamento de uma palavra com um texto o número de vezes com que essa palavra ocorre nesse texto – a operação de decomposição em valores e vetores singulares expressa *X* como o produto de 3 matrizes:

$$\underset{(n,p)}{X} = \underset{(n,r)}{U}\ \underset{(r,r)}{D}\ \underset{(r,p)}{V}^{T}$$

Estas matrizes componentes U, D, V^T obtêm-se pelo método dos mínimos quadrados (Landauer, et al, 2007). As dimensões destas matrizes são, respetivamente,

$U\,(n \times r)$ $V\,(p \times r)$ e D $(r \times r)$

Em que n é o número de palavras, p é o número de textos e r é a chamada característica de X: o número de linhas ou colunas linearmente independentes – um número que, quando muito, é igual ao mínimo de n e p. No caso da situação de exemplo descrito no número anterior, $r = p = 3$.

A matriz U tem tantas linhas quantas as palavras e um número de colunas = r.

A matriz V tem tantas linhas quantos os textos no corpus e um número de colunas igual a r.

A matriz D (simétrica e com valores apenas na diagonal) tem um número de linhas e colunas igual a r, contendo na diagonal os chamados valores singulares $s_1, s_2, \ldots, s_p$. Isto é:

$$D = \begin{bmatrix} s_1 & & & \\ & s_2 & & \\ & & \ldots & \\ & & & s_r \end{bmatrix}$$

Mostra-se – ver, por exemplo, Gabriel (1971) – que os quadrados destes valores correspondem às variâncias dos fatores (ortogonais) subjacentes à matriz de dados. Isto é, a soma de $s_1^2 + s_2^2 + \ldots + s_p^2 =$ variância (ou informação) da matriz de dados.

Isto significa que U e V representam, respetivamente, através dos vetores nas respetivas linhas, as palavras e os textos por vetores com a mesma dimensão: $r \leq \min(n, p)$.

Por exemplo, na situação do exemplo anterior, se r fosse igual a 2, então tanto as palavras como os textos ficariam representados por vetores de dimensão = 2 e, portanto, poderiam ser representados simultaneamente por pontos de um plano.

Se $r = 3$ (é o caso) então tanto as palavras como os textos ficam representados por pontos de um espaço de 3 dimensões que se apresenta na **figura 2.2.3.1.**

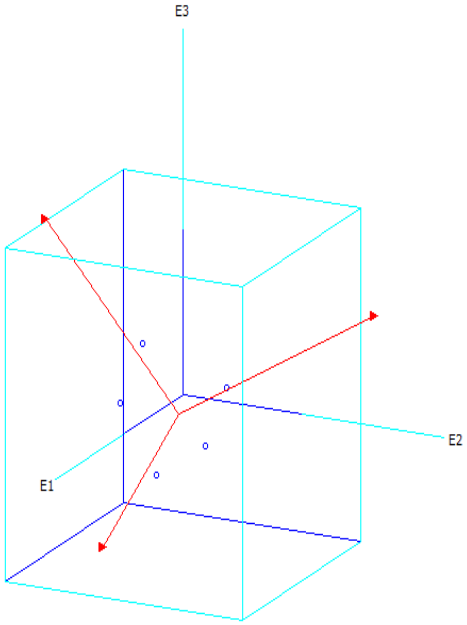


Figura 2.2.3.1. Os vetores representam os textos R_1, R_2, R_3 e os pontos ○ representam os cinco locais em que se sobrepõem as 33 palavras. Neste caso, r= min (n, p) = min (33, 3)= 3.

Considerando as projeções sobre o plano formado pelos eixos E_1 e E_2, obtém-se a **figura 2.2.3.2.**

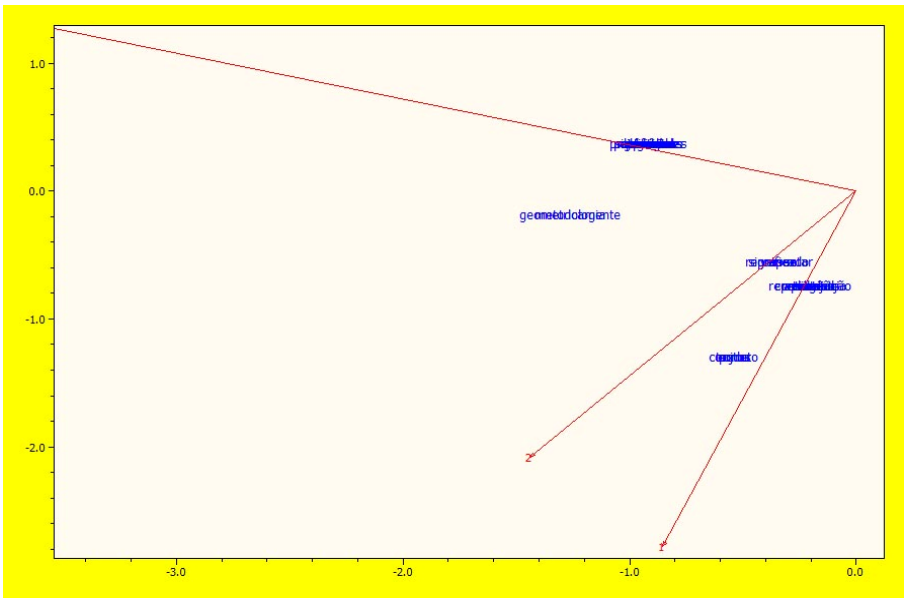


Figura 2.2.3.2. Os textos R_1, R_2, R_3 estão representados por vetores e as 33 palavras estão projetadas no plano E_1, E_2. Face à sobreposição de palavras com as mesmas coordenadas, o resultado é pouco legível.

As **figuras 2.2.3.1.** e **2.2.3.2.** são exemplos de biplots associados à SVD da matriz de dados original X.

As coordenadas para estas representações obtêm-se diretamente dos vetores que resultam da decomposição em valores e vetores singulares, de acordo com as metodologias – definidas por Gabriel (1971) e Galindo (1985) – para a construção de biplots.

Os biplots de Gabriel (1971) obtêm-se usando para as coordenadas das palavras as linhas da matriz

$$A = U\,\mathrm{D}^{\alpha}$$

e para as coordenadas dos textos (colunas) as linhas da matriz

$$B = U\,\mathrm{D}^{1-\alpha}$$

em que α é um número entre 0 e 1.

Observe-se que a decomposição em valores e vetores singulares pode ser apresentada como

$$X = U\,\mathrm{D}\,V^{T} = U\,\mathrm{D}^{\alpha}\,\mathrm{D}^{1-\alpha}\,V^{T} = U\,\mathrm{D}^{\alpha}\,(V\,\mathrm{D}^{1-\alpha})^{T} = A\,B^{T}$$

em que $A = U\,\mathrm{D}^{\alpha}$ e $B = V\,\mathrm{D}^{1-\alpha}$.

As n linhas de A e as p linhas de B contêm vetores que representam, respetivamente, as palavras e os textos num espaço comum de dimensão $r \leq \min(n, p)$.

O papel de D - e, portanto, dos valores singulares da diagonal respetiva – é o de definir as escalas dos eixos do gráfico (plano 3-dimensional) em que vão ficar representados tanto palavras como textos.

Estas representações geométricas em que tanto palavras como textos (linhas e colunas) de X aparecem na mesma representação gráfica (gráfico de dispersão) em posições perfeitamente definidas, podem generalizar-se a quaisquer dimensões 1, 2, 3, …, r – embora apenas possam ser visíveis para $r \leq 3$.

Exemplo 2.2.3.1. (O que é a ASL? - continuação)

No caso da matriz X com a representação vetorial correspondente aos três textos R_1, R_2, R_3 tem-se (usando o *software* referido em Vairinhos, e Galindo (2004)) a seguinte decomposição:

A=	1	2	3
1	-0.324	-0.720	-0.614
2	-0.843	0.514	-0.158
3	-0.843	0.514	-0.158
4	-0.429	-0.466	0.774
5	-0.753	1.186	0.160
6	-1.167	-0.207	-0.772
7	-0.753	-1.186	0.160
8	-0.324	-0.720	-0.614
9	-0.843	0.514	-0.158
10	-1.272	0.048	0.615
11	-0.843	0.514	-0.158
12	-0.843	0.514	-0.158
13	-0.843	0.514	-0.158
14	-0.843	0.514	-0.158
15	-1.272	0.048	0.615
16	-1.272	0.048	0.615
17	-0.843	0.514	-0.158
18	-0.843	0.514	-0.158
19	-0.753	-1.186	0.160
20	-0.843	0.514	-0.158
21	-0.843	0.514	-0.158
22	-0.324	-0.720	-0.614
23	-0.429	-0.466	0.774
24	-0.324	-0.720	-0.614
25	-0.429	-0.466	0.774
26	-0.843	0.514	-0.158
27	-0.324	-0.720	-0.614
28	-0.843	0.514	-0.158
29	-0.843	0.514	-0.158
30	-0.753	-1.186	0.160
31	-0.429	-0.466	0.774
32	-0.324	-0.720	-0.614
33	-0.324	-0.720	-0.614

B=	1	2	3
1	-0.144	-2.683	-1.648
2	-0.195	-1.736	2.078
3	-3.763	1.913	-0.425

Empilhando as matrizes *A* e *B* numa matriz *C* – **figura 2.2.3.3.** – as coordenadas que permitem a representação gráfica das palavras ocupam as *n* primeiras linhas da matriz *C* e as coordenadas que permitem a representação gráfica dos textos ocupam as posições *n*+1, …, *n*+*p* da matriz *C*.

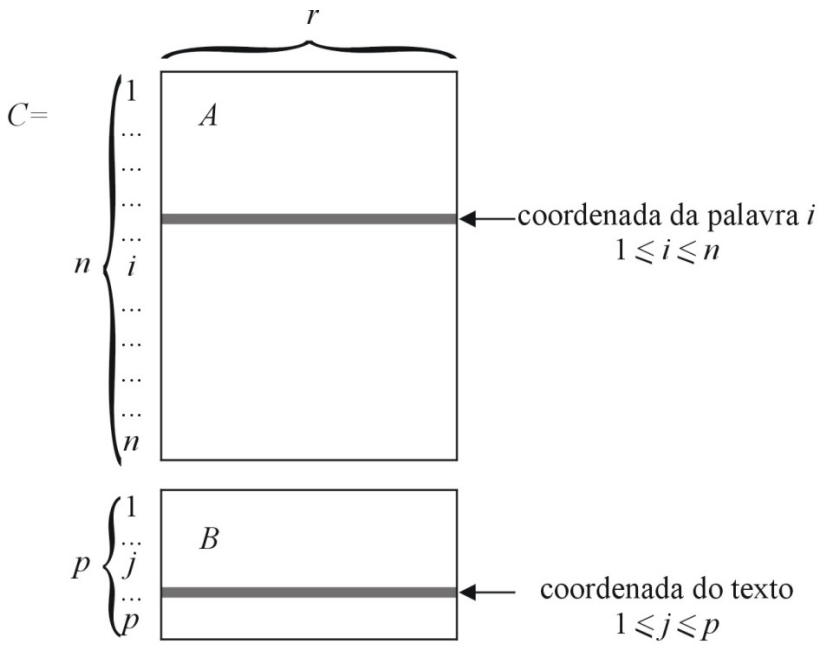


Figura 2.2.3.3. A matriz *C* que se obtém empilhando as matrizes *A* e *B* contém os marcadores de palavras e textos no espaço de dimensão *r*.

Verifica-se pela metodologia indicada anteriormente que, escolhendo duas ou três colunas da matriz $C = \begin{bmatrix} A \\ B \end{bmatrix}$ - veja **figura 2.2.3.3.** - podem obter-se representações gráficas em que aparecem simultaneamente as palavras e os textos que foram usados para construir a matriz *X*.

Dado que estas representações mostram simultaneamente os dois tipos de informação (linhas e colunas) designam-se biplots (Gabriel, 1971). Contudo, como se vê, este conceito pode generalizar-se a qualquer dimensão acima de 2, embora os gráficos correspondentes não possam visualizar-se.

Fica também claro das considerações anteriores que toda a metodologia computacional correspondente à ASL (Análise da Semântica Latente) que seja baseada na SVD, fica captada pela metodologia dos biplots, definida em 1971 por Gabriel e, muito especialmente pelos biplots de Galindo (Galindo, 1985).

Em geral, na expressão da decomposição usada em Gabriel (1971).

$$X = U\,\mathrm{D}\,V^T = U\,\mathrm{D}^{\alpha}\,\mathrm{D}^{1-\alpha}\,V^T = U\,\mathrm{D}^{\alpha}\,(V\,\mathrm{D}^{1-\alpha})^T = A\,B^T$$

faz-se $\alpha = \frac{1}{2}$, pondo assim, em pé de igualdade os pesos das representações das palavras e dos textos. Para $\alpha \neq \frac{1}{2}$, palavras e textos ficam representados em referenciais com escalas distintas.

Galindo (1987) mostra que quando se definem as matrizes *A* e *B* através de $A = U\,\mathrm{D}$ e $B = V\,\mathrm{D}$ então as linhas da matriz $C = \left[\frac{A}{B}\right]$ - veja **figura 2.2.3.3.** - contêm os chamados marcadores (no sistema de coordenadas dos vetores – variáveis latentes e textos) ficando tanto uns como outros representados com a máxima qualidade possível e sendo agora a representação uma genuína representação conjunta (num gráfico cujos eixos têm as mesmas escalas) de palavras (linhas) e textos (colunas).

Mostra-se também em Galindo (1985) que, ao contrário dos biplots de Gabriel (1971), em que

$$X = U\,\mathrm{D}\,V^T = A\,B^T, \text{ com } A = U\,\mathrm{D}^{\alpha},\ B = V\,\mathrm{D}^{1-\alpha}\ (0 \leq \alpha \leq 1)$$

nos biplots de Galindo isto não sucede, tendo-se

$$Y = (U\,\mathrm{D})\,(V\,\mathrm{D})^T = U\,\mathrm{D}^2\,V^T$$

em que a matriz D^2 contém, agora, os valores próprios (variâncias dos eixos fatoriais). É claro que $Y \neq X$ e, portanto, esta definição de biplot não permite recuperar a matriz de dados inicial.

Mostra-se, em Galindo (1985), que essas representações têm as seguintes propriedades:

1. O cosseno do ângulo entre duas colunas (textos) de *X* nessa representação significa o coeficiente de correlação entre as colunas correspondentes. Em síntese: o cosseno do ângulo entre dois textos é o coeficiente de correlação entre os vetores representativos dos textos. Isto significa que se dois textos fazem no gráfico um ângulo muito pequeno, os respetivos significados estão muito próximos.
2. Nessa representação, a distância entre os pontos representativos de duas linhas é a distância de Mahalanobis. Isto significa que se duas linhas (palavras) de *X* estão representadas como pontos próximos no biplot, então correspondem a

linhas semelhantes (palavras) e reciprocamente, linhas muito distintas estarão nessa representação, muito afastadas.

3. Uma vez que nesta representação palavras e textos estão representados num sistema referencial (formado pelos eixos fatoriais dotados de escalas idênticas para linhas/palavras e colunas/textos) então tem sentido definir ângulos (e cossenos dos ângulos) entre palavras e textos. O cosseno deste ângulo implica que o significado de uma palavra contribui tanto mais para o significado de um texto quanto menor o ângulo entre a palavra e o texto. Ou seja: o significado de uma palavra está tanto mais associado ao significado de um texto – contribui tanto mais para o significado do texto – quanto menor for o ângulo entre a palavra e o texto no gráfico. Esta interpretação não é admissível nos biplots de Gabriel.

Mostra-se em Bellegarda (2007) que os cossenos dos ângulos envolvidos entre as representações das palavras e textos podem ser calculados como a seguir indicado.

Seja $C = \left[\frac{A}{B}\right]$ (veja **figura 2.2.3.3.**), se considerarmos a linha genérica a_i (i= 1 … n) da matriz A e a linha genérica b_j (i= 1 … p) da matriz B, tem-se:

$a_i = u_i$ D (palavras i)

$b_j = v_j$ D (texto j)

em que u_i e v_j são as linhas genéricas das matrizes U e V, respetivamente (resultantes da SVD).

O cosseno do ângulo entre os dois vetores representantes de uma palavra de um texto é $Cos(a_i, b_j) = \frac{\langle a_i, b_j \rangle}{\sqrt{\langle a_i, a_i \rangle \langle b_j, b_j \rangle}} = \frac{a_i\, b_j^T}{\sqrt{a_i\, a_i^T\, b_j\, b_j^T}} = \frac{u_i\, D^2\, v_j^T}{\sqrt{(u_i\, D^2\, v_j^T)(v_j\, D^2\, v_j^T)}}$.

O cosseno do ângulo entre duas palavras é:

$$Cos(a_i, a_k) = \frac{u_i\, D^2\, u_k^T}{\sqrt{(u_i D^2\, u_i^T)(u_k\, D^2\, u_k^T)}}.$$

O cosseno do ângulo entre dois textos é:

$$Cos(b_j, b_e) = \frac{v_j D^2 v_e^T}{\sqrt{\left(v_j D^2 v_j^T\right)\left(v_e D^2 v_e^T\right)}}.$$

Na terminologia de Laudauer, Foltz, e Laham (1998) e Landauer, et al. (2007) a matriz $C = \left[\frac{A}{B}\right]$ representa então o "espaço semântico" associado a um certo conjunto de textos (corpus).

Como pode verificar-se, do ponto de vista computacional, quando a ASL se baseia na SVD da matriz X, todos os conceitos da ASL podem, com vantagem, ser interpretados em termos dos biplots de Galindo. A representação conjunta de palavras e textos no mesmo gráfico tem vantagens consideráveis do ponto de vista da interpretação intuitiva do que está em causa, uma vez que as proximidades angulares entre os dois tipos de conceitos em análise (palavras e textos) permite interpretar intuitivamente o significado dos textos em função do significado das palavras que lhe ficam próximas (no sentido angular); as relações entre palavras (semelhanças) em função das proximidades angulares e distâncias entre elas; as proximidades angulares entre textos representando proximidades entre os respetivos significados. Esta interpretação gráfica intuitiva da operação de decomposição da matriz de frequências, mantêm-se mesmo que a utilização de grande número de dimensões – por exemplo 300 dimensões - não permita visualizar os gráficos dos biplots correspondentes. Contudo, o conceito de ângulo, correlação e toda a interpretação anteriormente apresentada continua a fazer sentido.

É claro que, de acordo com Landauer (1997) o essencial da ASL está na análise psicológica do processo de aquisição do significado de novas palavras por parte das crianças que, conforme resultados experimentais e de simulação desses autores, parece ser bem modelado pelo mecanismo em que se baseia a operação de SVD, tendo-se concluído que as proximidades mentais entre os significados das palavras ficam bem representados por distâncias geométricas obtidos através da SVD – e que no presente trabalho se mostra que são bem captados através do conceito de biplot, como se pode ver experimentalmente no **Capítulo III**.

2.2.4. Métodos de Avaliação de Conhecimentos baseados em ASL/Bipolts.

De acordo com Landauer, Foltz, e Laham (1998, 2003) o processo de aquisição de novos conceitos por parte das crianças parece ser satisfatoriamente modelado usando os princípios subjacentes à construção das decomposições em valores e vetores próprios de matrizes retangulares que traduzem a representação vetorial dos textos (Salton, Wong, & Yang, 1975).

Esta conclusão foi obtida experimentalmente a partir de simulações realizadas usando esse modelo. Daqui decorre, naturalmente, a tentativa de usar essa metodologia computacional para avaliar conhecimentos acerca de uma certa matéria que tenham sido adquiridos através da língua natural e expressos através de relações entre os significados de palavras relevantes, obtidos durante a aprendizagem.

Isto é, se uma pessoa/estudante obteve conhecimento de algo de relevante acerca de certa matéria, esse conhecimento – expresso na mente do sujeito através de proximidades semânticas aprendidas pelos métodos anteriormente descritos, deve revelar-se ou manifestar-se no que escreve, na escolha das palavras e suas associações em textos de resposta a questões abertas sobre o tema em avaliação.

A construção da resposta a uma questão aberta implica que o respondente se esforce por identificar, escolher, as palavras cujos significados estão associados, na sua mente, ao tema em apreço. Estas associações – corretas ou não – têm a ver com a qualidade do conhecimento adquirido durante a aprendizagem.

Em síntese (Landauer, & Dumais, 1997; Landauer, Foltz, & Laham, 1998, 2003) as semelhanças psicológicas entre os significados das palavras associados ao conhecimento de uma certa matéria tendem a manifestar-se nos textos redigidos pelos estudantes através de proximidades entre as palavras cujos significados são o suporte desse conhecimento. Isto significa que o conhecimento adquirido pelo estudante acerca de uma certa matéria pode, em princípio, ser avaliado examinando as proximidades das palavras usadas nos textos de resposta (redação/*essays*) a questões de resposta aberta usando, na análise dos textos produzidos, exatamente o mesmo modelo computacional que parece estar envolvido na respetiva aquisição e organização mental – o SVD, nas perspetiva de Landauer.

Como se viu atrás, os resultados de uma análise SVD podem ser expressos de modo gráfico (biplots) em duas, três ou mais dimensões, se bem que com uma certa perda quantificável de informação.

Landauer, Foltz, e Laham (1998) identificam cinco métodos para atribuir classificações holísticas a textos de resposta a questões abertas, com base no espaço semântico construído com os textos usados para a aprendizagem da matéria em causa.

Neste trabalho analisam-se dois métodos principais, relacionados (mas não idênticos) com o Método 1 (Landauer, Foltz, & Laham, 1998, p. 30) e com o Método 5 (Landauer, Foltz, & Laham, 1998, p. 30) e que passamos a descrever em seguida. Nos números **3.5.** e **3.6.** deste trabalho serão apresentados os resultados experimentais acerca da respetiva utilização.

2.2.4.1. Método 1.

Os passos a usar para implementar este método são os seguintes:

1.º Construir um Espaço Semântico (ES) com base nos textos usados para a aprendizagem dos estudantes e outros textos escritos por peritos da matéria a ensinar.

2.º Professores classificam uma certa percentagem das respostas dos estudantes, constituindo os testes assim classificados uma amostra de aprendizagem do classificador a construir.

3.º A amostra classificada pelo método tradicional – referida em 2 – é posicionada no espaço semântico usando um processo de cálculo descrito em Bellegarda (2007).

4.º Os textos não classificados são, por sua vez, posicionados no ES e classificados em função da respetiva semelhança – medida pelos cossenos dos ângulos – em relação aos textos já classificados.

O processo é exemplificado usando a tabela, produzida pelo *software* experimental desenvolvido:

Textos NO Classificados		CL.SYS	CL.PROF	Correlaçã	Txt 11	Txt 12	Txt 14	Txt 15	Txt 17	Txt 18		
					39	20	36	17	43	25		
Txt. Nº: 20		39	37	0,908	0,98	0,71	0,97	0,91	0,95	0,64		
Txt. Nº: 21		25	28		0,51	0,92	0,51	0,72	0,61	0,97		
Txt. Nº: 23		39	49		0,94	0,58	0,92	0,78	0,85	0,52		
Txt. Nº: 24		25	15		0,54	0,91	0,56	0,71	0,62	0,97		
Txt. Nº: 26		39	43		0,96	0,61	0,88	0,81	0,85	0,54		
Txt. Nº: 27		25	18		0,59	0,96	0,61	0,79	0,70	0,99		

Figura 2.2.4.1.1. Ilustração do Método 1. Os cossenos dos ângulos entre textos a classificar (lado esquerdo) e a da amostra de treino (topo) estão no corpo da tabela.

Referindo-nos à figura anterior, e às **figuras 2.2.1.1.** e **2.2.1.2.**, suponhamos que já foi construído o ES com base nos textos usados na aprendizagem dos estudantes e que estão por classificar os textos 11, 12, 14, 15, 17, 18, 20, 21, 23, 24, 26, 27 produzidos pelos estudantes.

Pelo método tradicional – professor – são classificados os textos 11, 12, 14, 15, 17, 18 aos quais foram atribuídas, por esses professores, respetivamente, as classificações 39, 20, 36, 17, 43, 25.

Em função desta informação e do ES pretende-se atribuir classificações aos textos (coluna do lado esquerdo) 20, 21, 23, 24, 26, 27.

No passo número 3, os textos classificados pelo professor são posicionados no ES, usando a metodologia definida, por exemplo, em Bellegarda (2007).

Consideremos, por exemplo, o texto número 23. Esse texto faz com os textos já classificados – 11, 12, 14, 15, 17, 18 – ângulos cujos cossenos são, respetivamente, 0.94 (com 11), 0.58 (com 12), 0.92 (com 14), 0.78 (com 15), 0.85 (com 17) e 0.52 (com 18). A classificação 39 foi, neste caso, obtida como sendo igual à classificação do texto com o qual o texto 23 tem o menor ângulo (maior cosseno), maior semelhança. Isto é: 39.

Outras alternativas são possíveis, como por exemplo, a média dos k textos classificados mais próximos do texto a classificar (k= 1, 2, 3). Ver a **figura 1.4.6.2** do **Capítulo I**.

As classificações dos restantes textos são obtidas de modo semelhante.

A fim de ter uma ideia da validade do método, neste caso também os testes 20, 21, 23, 24, 26, 27 foram classificados pelo Professor e, no final, foi calculado o coeficiente de correlação entre as classificações do professor e as atribuídas automaticamente pelo algoritmo descrito, tendo-se obtido, neste caso específico, o valor 0.908, também, apresentado na **figura 2.2.4.1.1.**

O significado geométrico do processo descrito pode ver-se no esquema da **figura 2.2.4.1.2.** a seguir apresentada.

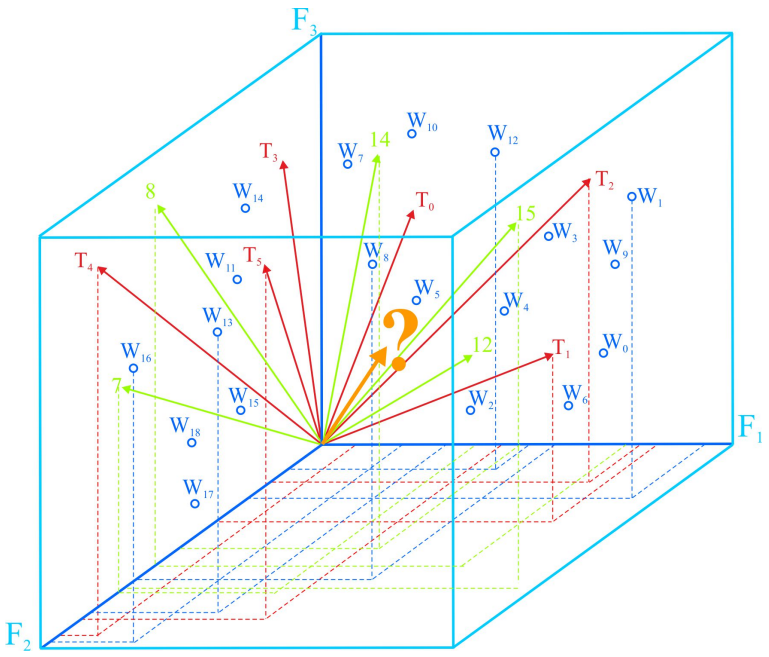


Figura 2.2.4.1.2. Biplot 3D correspondente ao Método 1. Na figura, T_1, T_2, …, T_{15} são textos e W_1, W_2, …, W_{27} são palavras do ES.

O símbolo "?" corresponde a um texto cuja classificação é desconhecida.

A **figura 2.2.4.1.2.** é apenas um esquema explicativo para veicular o significado geométrico do que está em causa. Corresponde a um biplot em 3 dimensões (assumindo neste caso – o que é geralmente muito irrealista – que essas 3 dimensões captam a totalidade da informação). As setas pintadas a verde correspondem a textos classificados com valores 12, 14, 15, …, etc.

Uma vez que o texto cuja classificação se pretende atribuir forma um ângulo de cosseno máximo com o texto classificado com 15, então, se o critério for este, a classificação seria 15.

Contudo, se a regra a implementar fosse a de atribuir como classificação a média dos k= 3 textos classificados mais próximos (com maiores cossenos), o resultado a atribuir ao texto identificado por "?" seria

$(14 + 15 + 12) / 3 \cong 14$.

Numa situação real, as dimensões envolvidas são da ordem das dezenas, pelo que um gráfico cartesiano para biplots com dimensão *n-D*, com $n > 3$ não são visualizáveis (Landauer et al., 1993). Contudo, pode-se construir um chamado biplot cilíndrico (Vairinhos & Galindo, 2012) em que, usando um gráfico de coordenadas paralelas correspondente aos eixos fatoriais subjacentes à SVD é possível visualizar um biplot em espaços de dimensão muito elevada se bem que não tão intuitivos quanto os gráficos cartesianos habituais. Nesse biplot, tanto textos como palavras são representados por trajetórias.

2.2.4.2. Método 2.

Neste Método 2 – à semelhança do que sucede com o Método 5, mencionado em Landauer, Foltz, e Laham (1998) – apenas se usam os textos dos estudantes e o texto da resposta do professor à questão posta. A ideia básica é modelar a preferência do professor pela resposta do estudante usando como critério o cosseno do ângulo entre a resposta do estudante e a resposta do professor: quanto maior esse cosseno (menor o ângulo) maior a "consonância" / preferência entre a resposta do estudante e a resposta do professor. Como variante deste método também se pode medir a distância da resposta estudante-professor e usar essa distância como critério de preferência do professor pela resposta do estudante.

Considera-se que este método pode ter interesse para apoiar o professor nos processos de avaliação formativa, objetivando comparações entre textos, estudante/professor. O que faz com que este método possa, eventualmente, ser usado na monitorização da absorção de conhecimentos transmitidos pelo professor em aulas presenciais mas não em avaliação global que tem carácter absoluto por pretender comparar os conhecimentos do estudante com um referencial absoluto (o espaço semântico formado a partir dos textos de ensino).

A vantagem consiste em poder ser facilmente operacionalizável, uma vez que exige como únicos *inputs peças* de informação obrigatoriamente presentes as respostas dos

estudantes e a resposta do professor às questões que ele próprio formulou, não implicando a classificação prévia de parte dos textos.

Posto isto, o método de classificação consiste nos passos seguintes:

Método 2.

1. Construir um espaço semântico (só de estudantes) usando a matriz X que é a representação vetorial das palavras e textos dos estudantes.
2. Sobre o espaço semântico obtido em **1** posicionar a resposta do professor, ou os textos de resposta do professor ou vários professores, usando a metodologia definida em Bellegarda (2007).
3. Obter os cossenos dos ângulos entre os textos dos professores e os textos de cada uma das respostas dos estudantes e calcular a nota do texto a classificar em função destes valores.

   A ideia subjacente é a de que, quanto menor o ângulo entre o texto de uma estudante e o texto do professor, maior a preferência do professor pelo texto do estudante.
4. Ordenar estas classificações.

Na **figura 2.2.4.2.1.** está explicada a interpretação geométrica deste método.

Ver também **figura 3.7.5.**, no **Capítulo III**.

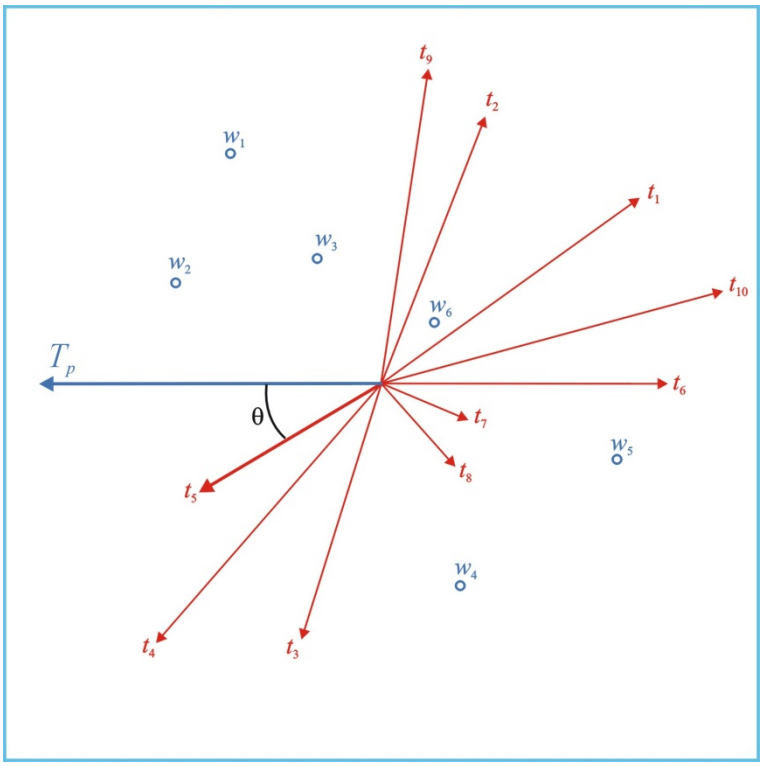


Figura 2.2.4.2.1. Os textos dos estudantes são representados pelos vetores (marcadores) t_1, t_2, …, t_{10}. O texto da resposta do professor está representado pelo vetor T_p. A classificação do estudante t_5 é função do cosseno do ângulo θ entre esse texto e o vetor T_p.

2.3. Estruturação da Amostra.

A principal questão de inferência subjacente a este estudo tem a ver com a questão da validade do uso dos métodos de classificação automática. Seguindo a literatura sobre o tema (Landauer, Foltz, & Laham, 1998, 2003; Landauer, et al., 2007; Islam & Hoque, 2010; Yang, Buckendahl, Juszkiewicz, & Bhola, 2012) como se viu na **Introdução**, a generalidade dos autores sugere como critério de validade dos métodos de avaliação automática de conhecimentos baseados em testes de resposta aberta o coeficiente de correlação entre as classificações atribuídas por professores e as classificações atribuídas pelos sistemas automáticos de classificação. Também se relaciona esse coeficiente de correlação R (sys, Prof) com o coeficiente de correlação entre as classificações atribuídas por professores diferentes R (Prof, Prof).

Para que um sistema seja considerado aceitável o coeficiente de correlação R (sys, Prof) deve ser elevado e significativamente diferente de zero; além disso, R (sys, Prof) deve ser da ordem de grandeza ou superior a R(Prof, Prof).

Isto significa que ao estruturar a amostra, o fator anterior (validade) e portanto a significância do coeficiente de correlação deve ser a principal consideração em causa; mais do que generalizar a toda a população alvo (neste caso, estudantes que usem a língua portuguesa como instrumento de aprendizagem da matéria a avaliar), a principal preocupação é garantir que os coeficientes de correlação obtidos são significativos e o mais elevado que for possível.

Tendo em conta as considerações anteriores e os recursos com que esta investigação contou – apenas os recursos do investigador – foi considerado que uma amostra de cerca de 600 respostas a questões abertas constantes dos exames nacionais de Português de três escolas escolhidas ao acaso poderiam gerar a evidência suficiente para a maior parte dos estudos a realizar.

Para os testes em apreço teriam de ser obtidos os enunciados dos exames, os critérios de classificação dos testes, as respostas dos estudantes a cada uma das alíneas de resposta aberta e classificações atribuídas pelos professores contratados para avaliarem as respostas e ainda os manuais de ensino usados no estudo dos estudantes para se prepararem para esses exames específicos.

Além deste dados deveriam ainda essas respostas serem avaliadas por outros professores – contratados agora pelo investigador – usando os mesmos critérios e critérios de avaliação holística – por forma a conhecer em que medida o critério de avaliação (holístico *versus* critério clássico) e o professor influenciavam os resultados.

2.4. Plano de Recolha de Dados.

Tendo em conta as limitações já referidas em **2.3.** e tendo ainda em conta que a principal questão de inferência (generalização) a partir dos dados se prende com a significância do coeficiente de correlação entre os resultados de uma classificação automática e as classificações a atribuir aos mesmos testes por professores humanos, o plano estabelecido foi o seguinte:

1. Escolher ao calhas três escolas secundárias a que a autorização do Ministério da Educação nos permitiu aceder e registar, em suporte magnético, os textos manuscritos dos testes disponíveis nessas escolas, garantindo pelo menos 600 respostas a questões de resposta aberta. Os dados assim obtidos são designados no

que se segue por EXMIN e incluem não só as respostas dos estudantes como os textos usados por esses estudantes no ano letivo a que se reportam esses exames.

2. Transcrição dos textos manuscritos em papel para um suporte magnético garantindo a não utilização do corretor ortográfico, por forma a que o texto a estudar pelo *software* a desenvolver corresponda ao que foi produzido pelos estudantes, sendo o mesmo com que os professores a contratar para realizar as correções têm de lidar.

Terminada a transcrição, os textos gerados pelo processador de texto – em geral o WORD da Microsoft – são carregados, usando o *software* a desenvolver no âmbito do projeto, na BDT – ver **figura 2.2.1.1.** – ficando disponíveis para os testes do *software* e para estudos de validação.

3. Nessa BDT são carregados não só os textos das respostas dos estudantes como os textos dos manuais usados no ensino dos estudantes cujas respostas são usadas no estudo.

4. Além dos textos, é carregada na BDT informação auxiliar que permita identificar os textos, seu autor e o contexto em que as respostas foram produzidas, de modo a permitir relacionar os resultados e das análises dos textos com outras variáveis. Desta informação auxiliar fazem parte as classificações atribuídas às respostas pelos professores contratados pelo Ministério da Educação e também as classificações atribuídas a essas mesmas respostas por professores a contratar pelo investigador para o mesmo efeito.

5. Por contacto pessoal junto de um professor de gestão de um instituto politécnico foi possível aceder às respostas manuscritas – no contexto de uma avaliação formativa – de cerca de duzentos estudantes na resposta a uma questão relativa ao conceito de Qualidade Total, bem como ao texto usado no estudo dessa matéria. Esses dados, de carácter puramente observacional, são também carregados na BDT, servindo essencialmente para verificar os aspetos operacionais do programa a implementar na concretização do sistema descrito em **3.2.** No que se segue, estes dados são designados por GESTÃO.

6. Por contacto pessoal junto do orientador desta tese foi também possível aceder aos textos das respostas de 17 estudantes do ensino superior relativas aos fatores de disciplina / indisciplina em sala de aula.

Trata-se também de dados observacionais úteis para validar certos aspetos funcionais do sistema a implementar. Estes dados são também carregados na BDT e serão designados no que se segue por DISCIPLINA.

Em síntese, à exceção dos dados EXMIN, que permitem, em princípio, aplicar inferência estatística ao coeficiente de correlação (sys, Prof), todos os dados restantes têm carácter observacional que, se bem que não permitam justificar teoricamente generalizações para lá dos dados, permitem contudo sujeitar o programa a implementar a uma gama variada de situações que permitam testar o seu funcionamento e ilustrar a respetiva funcionalidade em condições reais.

2.5. Metodologia Estatística de Tratamento de Dados.

Recolhidos e registados esses dados numa Base de Dados Textual (BDT) as metodologias a usar para o seu tratamento estatístico são as seguintes:

1. Estatística descritiva univariada das principais variáveis, a saber:

- Distribuição da amostra por disciplina, classificação do professor, comprimento do texto, disciplina, curso.
- Distribuição dos resultados obtidos nos testes classificados por professores ao nível de cada uma das questões de resposta aberta.
- Distribuição das classificações atribuídas pelos professores contratados para reapreciarem as provas previamente classificadas ao nível das questões de respostas abertas.

2. Estudos bivariados:

- Relacionar, usando tabelas de contingência, coeficientes de correlação ou gráficos de dispersão e os resultados obtidos pelo sistema com as classificações atribuídas por professores diferentes, calculando a significância dos coeficientes de correlação.

3. Utilização dos biplots de Galindo para expressar graficamente os Espaços Semânticos usados no estudo das diversas metodologias de classificação e para estudar as relações entre textos e palavras usadas na representação vetorial dos textos. Face à elevada dimensionalidade dos espaços de representação (sempre

superior a 3 o que torna impossível a visualização completa da informação) estes biplots são usados como auxiliares da intuição geométrica tanto na compreensão dos métodos como na interpretação dos resultados.

4. Utilização de métodos de análise classificatória (análise de "*clusters*") usando as coordenadas dos textos resultantes da decomposição SVD para detetar agrupamentos homogéneos de textos de resposta dos estudantes e que traduzam resultados das comparações desses textos dois a dois.

Capítulo III: Recolha e Análise de Dados.

3.1. Introdução.

Como se viu no **Capítulo I** – Metodologia – este estudo pode ser classificado como um estudo fundamentalmente observacional destinado a recolher evidências acerca da possibilidade de usar técnicas de análise estatística de textos em tarefas de ensino – aprendizagem num contexto escolar.

Não sendo um estudo experimental em que o investigador tem total controlo sobre as condições de observação (quando, quem, como e em que condições) procurou-se, isso sim, obter dados reais característicos dos ambientes – alvo das metodologias eventualmente saídas deste estudo – dando cumprimento às indicações em **2.3.**

Foi, pois, decidido que seriam obtidos textos de resposta de estudantes portugueses do 12º ano a questões abertas integrantes dos exames nacionais de algumas disciplinas – como Filosofia, Português, História – em que os conhecimentos adquiridos são expressos em português.

Inicialmente foi também equacionada a possibilidade de utilização de textos de resposta a questões abertas constantes de plataformas eletrónicas de ensino. A grande vantagem dessa abordagem seria o facto de os textos produzidos pelos estudantes e professores se encontrarem já digitalizados, reduzindo enormemente o esforço financeiro da investigação, o tempo necessário à preparação dos dados e aumentando a dimensão da amostra.

Infelizmente, insuspeitadas dificuldades relacionadas com a legislação sobre privacidade e outras, levaram-nos a abandonar essa hipótese.

Concentrámo-nos, pois, na possibilidade de obter textos produzidos por estudantes do ensino secundário ao responderem às provas dos exames oficiais de um ano suficientemente afastado da atualidade para reduzir as reticências ao respetivo acesso.

Obtida uma carta de apresentação do Ministério de Educação, escolheram-se três escolas ao acaso e obteve-se o acesso a fotocópias de n= 200 respostas a dois grupos de questões constantes dos exames nacionais de Português, História, Filosofia, entre outras, de 2008 dos estudantes das escolas em causa – num total de cerca de 600 textos gerados

pelas respostas dos estudantes (ver em **3.5.** a caracterização das amostras) – no que se segue designados por dados EXMIN. Também se obtiveram os dados observacionais das duas outras fontes citadas em 2**.3.**: dados de resposta a uma questão de gestão de uma turma de um curso de Gestão num instituto politécnico (dados GESTÃO) em resposta a uma questão de qualidade e ainda as respostas de 17 estudantes de um curso superior relativas aos fatores de indisciplina na sala de aula (DISCIPLINA).

No *site* do Ministério de Educação (GAVE) *http://bi.gave.min-edu.pt/exames/exames/eSecundario/524/?listProvas* (2008 | 2ª fase) pode consultar-se o enunciado desse exame, bem como as cotações da prova. Os itens com resposta aberta eram os relativos ao Grupo I partes A e B e Grupo III com cotações, respetivamente,

Grupo I A – 70 pontos

Grupo I B – 30 pontos

Grupo III – 50 pontos.

A cotação total do teste – abrangendo também questões de resposta múltipla e associação era de 200 pontos.

No número **3.5.** pode ver-se um resumo estatístico da amostra de textos obtida.

Para lá dos textos das respostas dos estudantes aos exames finais do Ministério de Educação, registaram-se ainda os resultados atribuídos por professores contratados pelo Ministério da Educação para classificar os testes.

Desde já se observa que o tipo de classificação especificada pelo Ministério da Educação nas regras de cotação, entra com fatores como a organização, a correção linguística, a estruturação temática e está relativamente afastada do tipo de observação holística que se recomenda para este tipo de metodologia (Landauer, Foltz, & Laham, 2003).

Integravam também este conjunto de dados a analisar os textos dos manuais de ensino, nomeadamente, textos de guias de ensino e interpretação dos textos literários envolvidos nos exames: os Lusíadas de Luís Vaz de Camões, o Memorial do Convento de José Saramago, poesia de Fernando Pessoa.

Usou-se para esta tarefa os manuais bastante difundidos entre os estudantes do ano letivo em causa.

Além destes elementos de informação, considerou-se ainda necessário pedir a colaboração de uma Professora de Português do Ensino Secundário, a quem foi pedido que classificasse 60 textos de Português usando um critério holístico, mais consentâneo com o recomendado para este tipo de estudos.

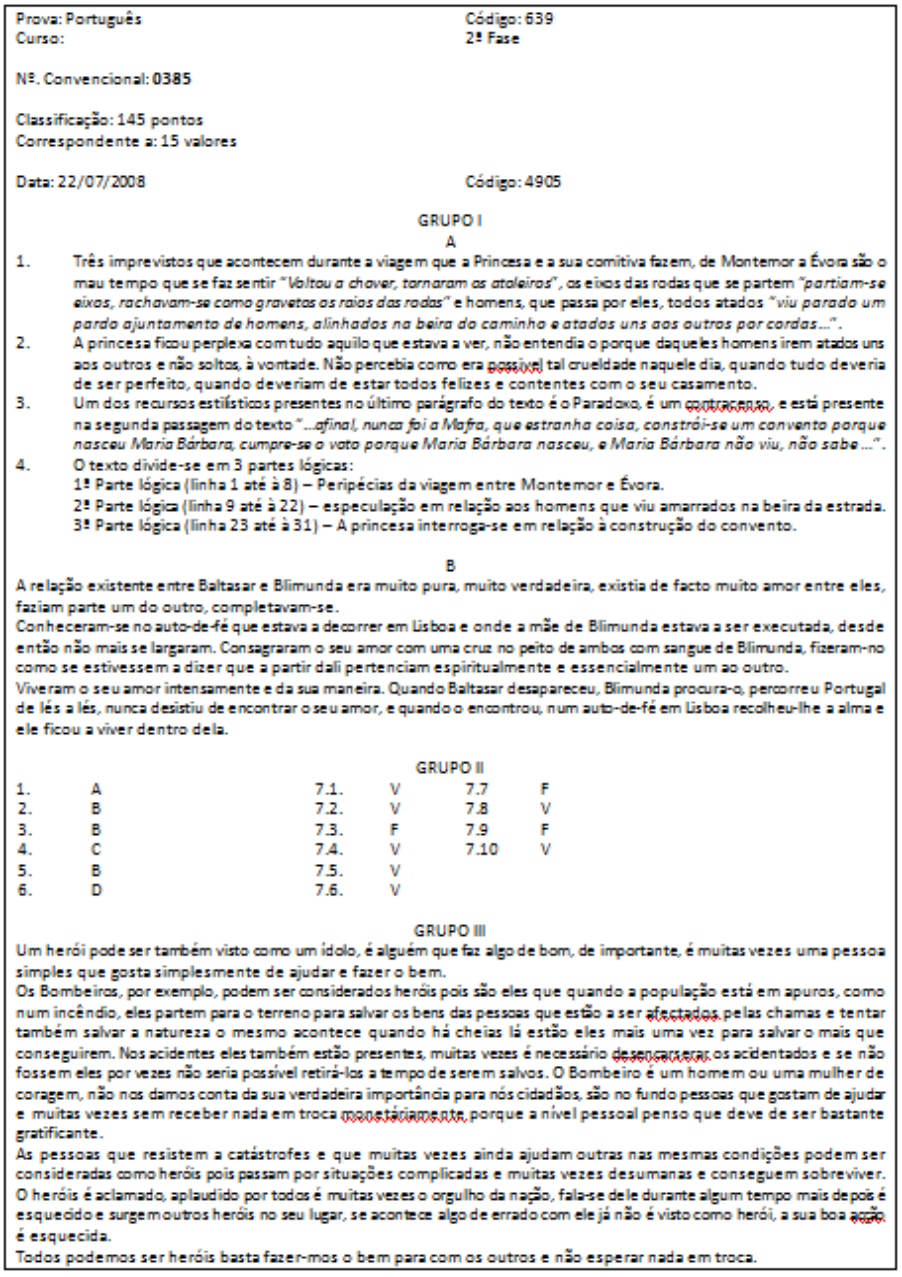
No número **3.9.6.** é apresentado um estudo comparativo das classificações atribuídas pelos professores do Ministério da Educação e pela Professora por nós contratada.

3.2. Recolha e Registo de Dados.

Sendo certo que todos os dados – pelas razões referidas – estavam em suporte de papel e eram constituídos – com exceção dos textos dos manuais de ensino – por textos manuscritos, a sua preparação para registo e posterior gravação em suporte digital revelou-se uma tarefa extremamente morosa e cara, uma vez que implicou a transcrição a partir dos textos fotocopiados e a posterior verificação para garantir uma fiabilidade total na transcrição, tendo sido necessário montar durante meses um pequeno serviço de apoio. Foi decidido não aplicar o corretor ortográfico nem corrigir erros gramaticais, por mais evidentes que fossem; procurou-se, isso sim, garantir que o texto a analisar fosse idêntico ao texto produzido pelo autor.

Inicialmente, por uma questão de facilidade e rapidez da tarefa de transcrição, foram produzidos ficheiros de texto usando um processador de texto. Numa segunda fase, os textos produzidos pelo processador de texto foram carregados, juntamente com os elementos de informação adicionais, numa base de dados desenvolvida para este projeto. Veja número seguinte.

Na **figura 3.2.1.** apresenta-se um exemplar desse tipo de ficheiro de texto.



Prova: Português — Código: 639
Curso: — 2ª Fase

Nº. Convencional: **0385**

Classificação: 145 pontos
Correspondente a: 15 valores

Data: 22/07/2008 — Código: 4905

GRUPO I

A

1. Três imprevistos que acontecem durante a viagem que a Princesa e a sua comitiva fazem, de Montemor a Évora são o mau tempo que se faz sentir "*Voltou a chover, tornaram os atoleiros*", os eixos das rodas que se partem "*partiam-se eixos, rachavam-se como gravetos os raios das rodas*" e homens, que passa por eles, todos atados "*viu parado um pardo ajuntamento de homens, alinhados na beira do caminho e atados uns aos outros por cordas…*".
2. A princesa ficou perplexa com tudo aquilo que estava a ver, não entendia o porque daqueles homens irem atados uns aos outros e não soltos, à vontade. Não percebia como era possível tal crueldade naquele dia, quando tudo deveria de ser perfeito, quando deveriam de estar todos felizes e contentes com o seu casamento.
3. Um dos recursos estilísticos presentes no último parágrafo do texto é o Paradoxo, é um contrasenso, e está presente na segunda passagem do texto "*…afinal, nunca foi a Mafra, que estranha coisa, constrói-se um convento porque nasceu Maria Bárbara, cumpre-se o voto porque Maria Bárbara nasceu, e Maria Bárbara não viu, não sabe …*".
4. O texto divide-se em 3 partes lógicas:
   1ª Parte lógica (linha 1 até à 8) – Peripécias da viagem entre Montemor e Évora.
   2ª Parte lógica (linha 9 até à 22) – especulação em relação aos homens que viu amarrados na beira da estrada.
   3ª Parte lógica (linha 23 até à 31) – A princesa interroga-se em relação à construção do convento.

B

A relação existente entre Baltasar e Blimunda era muito pura, muito verdadeira, existia de facto muito amor entre eles, faziam parte um do outro, completavam-se.
Conheceram-se no auto-de-fé que estava a decorrer em Lisboa e onde a mãe de Blimunda estava a ser executada, desde então não mais se largaram. Consagraram o seu amor com uma cruz no peito de ambos com sangue de Blimunda, fizeram-no como se estivessem a dizer que a partir dali pertenciam espiritualmente e essencialmente um ao outro.
Viveram o seu amor intensamente e da sua maneira. Quando Baltasar desapareceu, Blimunda procura-o, percorreu Portugal de lés a lés, nunca desistiu de encontrar o seu amor, e quando o encontrou, num auto-de-fé em Lisboa recolheu-lhe a alma e ele ficou a viver dentro dela.

GRUPO II

1.	A	7.1.	V	7.7	F
2.	B	7.2.	V	7.8	V
3.	B	7.3.	F	7.9	F
4.	C	7.4.	V	7.10	V
5.	B	7.5.	V		
6.	D	7.6.	V		

GRUPO III

Um herói pode ser também visto como um ídolo, é alguém que faz algo de bom, de importante, é muitas vezes uma pessoa simples que gosta simplesmente de ajudar e fazer o bem.
Os Bombeiros, por exemplo, podem ser considerados heróis pois são eles que quando a população está em apuros, como num incêndio, eles partem para o terreno para salvar os bens das pessoas que estão a ser afectados pelas chamas e tentar também salvar a natureza o mesmo acontece quando há cheias lá estão eles mais uma vez para salvar o mais que conseguirem. Nos acidentes eles também estão presentes, muitas vezes é necessário desencarcerar os acidentados e se não fossem eles por vezes não seria possível retirá-los a tempo de serem salvos. O Bombeiro é um homem ou uma mulher de coragem, não nos damos conta da sua verdadeira importância para nós cidadãos, são no fundo pessoas que gostam de ajudar e muitas vezes sem receber nada em troca monetariamente, porque a nível pessoal penso que deve de ser bastante gratificante.
As pessoas que resistem a catástrofes e que muitas vezes ainda ajudam outras nas mesmas condições podem ser consideradas como heróis pois passam por situações complicadas e muitas vezes desumanas e conseguem sobreviver.
O herói é aclamado, aplaudido por todos é muitas vezes o orgulho da nação, fala-se dele durante algum tempo mais depois é esquecido e surgem outros heróis no seu lugar, se acontece algo de errado com ele já não é visto como herói, a sua boa acção é esquecida.
Todos podemos ser heróis basta fazer-mos o bem para com os outros e não esperar nada em troca.

Figura 3.2.1. Texto relativo ao estudante identificado pelo número convencional 0385, relativo ao Exame da Português do 12º ano, Curso Científico-Humanísticos, depois de transcritos a partir do original manuscrito.

3.3. Estrutura da Base de Dados Textual (BDT).

Uma vez transcritos e verificados os textos das respostas para suporte digital, esses dados foram seguidamente carregados numa base de dados textual (BDT) de acordo com o especificado na **figura 2.2.1.2.**

Essa base de dados foi desenvolvida tendo em conta apenas as necessidades desta investigação, do desenvolvimento de um protótipo com funcionalidade pensada e

otimizada para permitir a realização das experiências necessárias ao estudo. A criação de um sistema para uso operacional exige desenvolvimentos específicos aqui não abordados.

A tabela principal em que foi armazenada a informação da amostra tem a estrutura caracterizada pelos seguintes campos (ver manual do PAET, em Anexo A):

IDTEXTO – Número sequencial identificador do texto.

TEMA – Indica a que tema se refere o texto (Português, Filosofia, História, etc.)

TITULO – Nome do exame a que se referia ou título do manual.

AUTOR – Número convencional do respondente (estudante que fez o exame ou autor do manual).

DATA – Data do exame ou data da edição do manual.

TIPTEXT – Tipo de texto

A – Reposta do Estudante

P – Texto produzido por um Professor

AP – Manual de Aprendizagem

PARTE – Grupo I, Grupo II, etc.

SPARTE – Subparte – A, B, C, etc.

TEXTO – Texto do estudante, do professor ou do manual.

CLASPROF – Classificação do professor (quando conhecida).

A BDT foi desenvolvida em Microsoft ACESS 2003, tendo-se usado campos do tipo MEMO para armazenar os textos completos das respostas dos estudantes e textos de ensino.

Com esta estrutura é possível guardar na mesma tabela as respostas dos estudantes, os textos produzidos pelos professores para fins didáticos, as respostas dos professores às questões colocadas, os manuais de ensino e, inclusivamente, textos dos escritores e outros autores, quando disponíveis.

Dado o volume de informação e a sua variedade, considerou-se útil separar, em tabelas diferentes embora com a mesma estrutura, a informação relativa às respostas dos estudantes, dos manuais, e dos textos produzidos pelos professores.

Na **figura 3.3.1.** apresenta-se o ecrã principal de consulta e registo de dados do programa protótipo desenvolvido para apoio da investigação. Ver em Anexo A o manual de utilização.

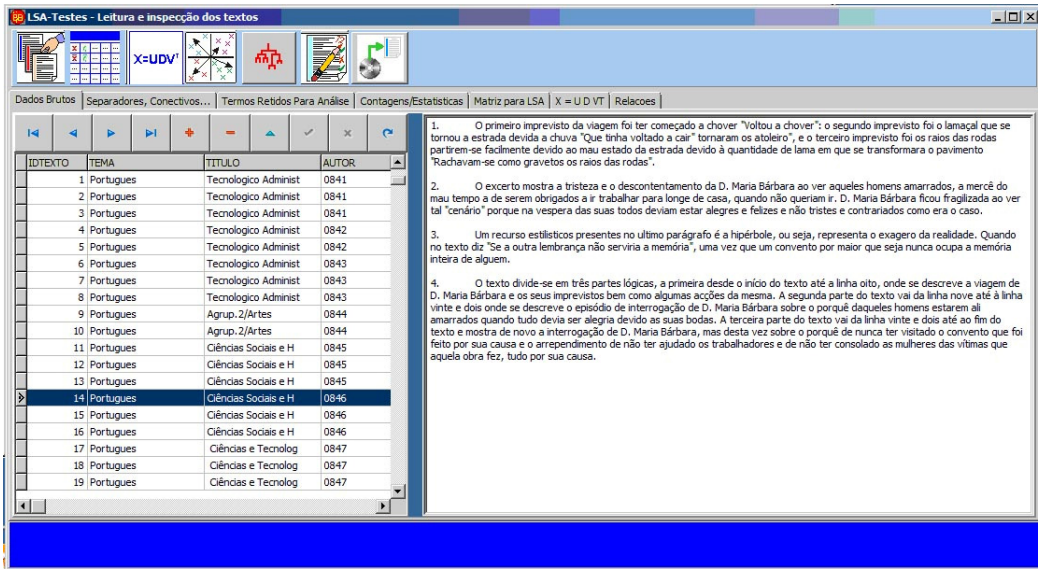


Figura 3.3.1. Consulta de um texto de resposta de um estudante. No lado esquerdo os elementos que permitem identificar o texto. Do lado direito o texto.

Além desta tabela constante na BDT, foi também criada a seguinte tabela – veja **figura 3.3.2.** – relativa às palavras a excluir (palavras funcionais) ao extrair as palavras dos textos.

NUMPAL – Identificador único da palavra.

ORIGEM – Origem da palavra – eventualmente o nome de um documento da qual foi extraído.

PALAVRA – A palavra ou símbolo a excluir.

CTGRAM – Categoria gramatical (Separador, Proposição, Pronome, Número).

QUANDO – Data da introdução.

Figura 3.3.2. Palavras funcionais, separadores e outros símbolos a excluir da análise.

Na **figura 3.3.3.** apresenta-se um exemplo com algumas dessas palavras funcionais, separadores e outros símbolos a excluir da análise lexicográfica.

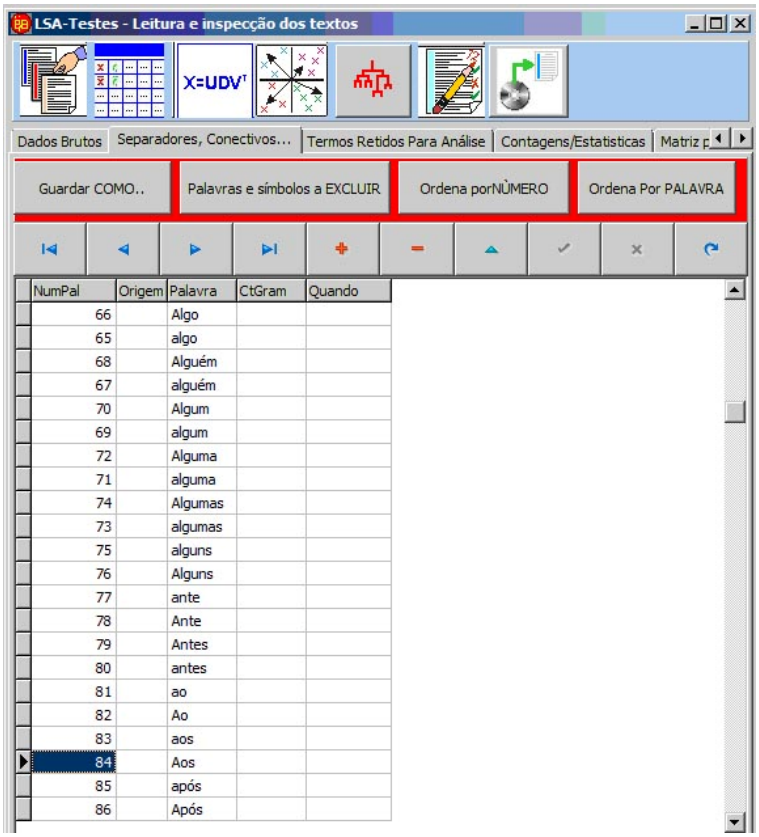


Figura 3.3.3. Exemplo de palavras funcionais, separadores e outros símbolos.

No apoio à execução destas tarefas considera-se muito útil a informação do site *http://www.iltec.pt/* do ILTEC – Instituto de Linguística Teórica e Computacional (consultado em 4 Dez 2012), bem como o site *http://www.portaldalinguaportuguesa.org/* (Portal da Língua Portuguesa).

Estes dois *sites* do ILTEC oferecem recursos – nomeadamente artigos científicos – de grande interesse para apoio de investigação futura nesta área.

3.4. Criação de Instrumentos de Análise Estatística de Textos.

Embora exista *software* de uso geral que permite realizar algumas das tarefas implícitas no esquema das **figuras 2.2.1.1.** e **2.2.1.2.** relativas ao modelo a implementar, o facto é que desde logo se considerou necessário desenvolver *software* integrado que

permitisse realizar, em ambiente interativo, com comodidade, as tarefas de análise textual seguintes:

1 – **Gestor da base de dados BDT**.

2 – **Toquenizador**.

Programa que permite, dado um texto, identificar todas as sequências de caracteres separadas por espaços em branco (ou outros símbolos separadores).

3 **– Lematizador**.

Programa que permite extrair a raiz comum de um conjunto de palavras.

4 – **Programa para construção da representação vetorial dos textos**.

Programa que permite transformar um texto numa matriz de contingência contendo no cruzamento de uma linha correspondente a uma palavra (ou forma) com uma coluna correspondente a um texto, a frequência com que essa palavra ocorre nesse texto.

5 – **Programa de decomposição SVD – *Singular Value Decomposition***.

Programa que decompõem uma matriz X contendo a representação vetorial de um texto ou conjunto de textos num produto de três matrizes $U\Sigma V^T$ em que U e V são matrizes contendo nas suas linhas – respetivamente – representações vetoriais, de dimensão reduzida, das palavras e dos textos e Σ uma matriz quadrada, diagonal, contendo os valores singulares da decomposição, por ordem decrescente.

6 – **Programa para construção do Espaço Semântico - ES**.

Programa que usando os resultados da decomposição SVD gera uma matriz cujas linhas são marcadores das palavras e dos textos num espaço comum de baixa dimensão. Ver **Capítulo I – Metodologia**.

Os marcadores usados neste trabalho correspondem ao biplot de Galindo (Galindo, 1985) e têm a expressão $U\Sigma$ para as palavras (isto é, cada palavra retida para análise (W_i, i= 1…n) é representada pelo marcador $b_i = u_i\ \Sigma$, i= 1…n, em que U e Σ resultam da SVD e em que os textos são representados pelas linhas de $V\Sigma$, sendo cada texto representado pelo marcador $b_j = v_j\ \Sigma$, i= 1…p.

Tanto a_j como b_j são representados por vetores de um espaço de baixa dimensão r, sendo r= característica de $X \leq \min(n, p)$ em que n é o número de palavras identificadas e retidas depois das operações de toquenização e lematização e p o número de textos.

7 – **Programa de Biplots**.

Programa para realizar a representação gráfica dos espaços semânticos tanto em coordenadas retangulares (cartesianas) como paralelas.

Com exceção dos programas de lematização e SVD, todo o restante *software* teve de ser desenvolvido de raiz ou adaptado no âmbito do projeto.

O programa SVD resultou de uma adaptação às necessidades deste trabalho de um programa de Sergey Bochkanov (projeto ALGLIB®). Ver *http://www.alglib.net* (consultado em 01 de Julho de 2010)

Quanto ao programa de lematização foi usado o lematizador "universal" de Porter. Ver *http://tartarus.org/~martin/PorterStemmer/* (consultado em 20 de Novembro de 2010)

Na implementação do programa foram ainda usados alguns procedimentos da biblioteca *Software Development Lohninger* (SDL) que pode ser consultado em *http://www.lohninger.com/* (consultado em 01 de Julho de 2011).

A preocupação neste projeto foi a de criar um laboratório experimental que permitisse testar algumas ideias relativas ao uso da análise estatística de testes usando textos de resposta a questões abertas, não tendo havido grandes preocupações de otimização.

Em particular, considera-se importante, em futuros trabalhos, desenvolver uma lematização mais adequada aos problemas do português.

3.5. Resumo Estatístico das Amostras.

3.5.1. Dados EX-MIN – Exames do Ministério.

Recorde-se que se trata de respostas ao exame de Português de 2008 (2ª Época) e que este exame comportava três grupos dos quais o Grupo I abrangia duas questões (A e B) relativas ao Memorial do Convento de José Saramago e um Grupo III também com uma questão de resposta aberta relativa a outros autores. Ver um exemplar no *site* do Ministério de Educação (GAVE) *http://bi.gave.min-edu.pt/exames/exames/eSecundario/524/*

?listProvas (2008 | 2ª fase). A cotação total do teste era de 200 pontos dos quais 70 afetados ao Grupo I A), 30 relativos ao Grupo I B) e 50 pontos relativos ao Grupo III.

No que se segue, o significado das variáveis usadas nos quadros e gráficos é o seguinte:

Variante – É a variante do enunciado do exame de português, conforme a origem dos estudantes.

Alínea – A) ou B) do Grupo I.

Grupo – Grupos G1, G2, G3.

ClassProf – Classificação atribuída pelo professor contratado pelo Ministério.

NCaract – Número de caracteres do texto da resposta do estudante.

Considerando a totalidade da amostra, a respetiva estrutura consta da **tabela 3.5.1.1.**

GRUPO		**A)**	**B)**	**Total**
G1	---	183	180	363
G2	198	---	---	198
Total	198	183	180	561

Tabela 3.5.1.1. Distribuição da amostra.

Dispõe-se, assim, de um total de 561 textos distribuídos por três agrupamentos praticamente da mesma dimensão.

Considerando agora apenas os textos da resposta ao grupo G1, alínea A), a **tabela 3.5.1.2.** resume os valores das classificações do professor e do número de palavras no texto para os textos de resposta à A).

	Nº Textos	**Mínimos**	**Máximo**	**Média**	**DP**
ClassProf	183	0	70	34.47	183
NCaract	183	25	4603	1767.38	699.7

Tabela 3.5.1.2. Resumo das classificações do professor e do comprimentos dos textos de resposta à A).

As **tabelas 3.5.1.3.** e **3.5.1.4.** apresentam os valores correspondentes para B) e para o grupo G3.

	Nº Textos	**Mínimos**	**Máximo**	**Média**	**DP**
ClassProf	180	0	30	14.39	8.3
NCaract	180	19	1404	712.78	166.5

Tabela 3.5.1.3. Resumo das classificações do professor e do comprimentos dos textos de resposta à B).

	Nº Textos	**Mínimos**	**Máximo**	**Média**	**DP**
ClassProf	198	0	48	25.7	11.2
NCaract	199	119	2795	1445.0	357.9

Tabela 3.5.1.4. Resumo das classificações do professor e do comprimento dos textos de resposta ao Grupo III.

A distribuição das variáveis ClassProf e NPalTexto nesses três agrupamentos aparecem, respetivamente, nas **figuras 3.5.1.1.** a **3.5.1.6.**

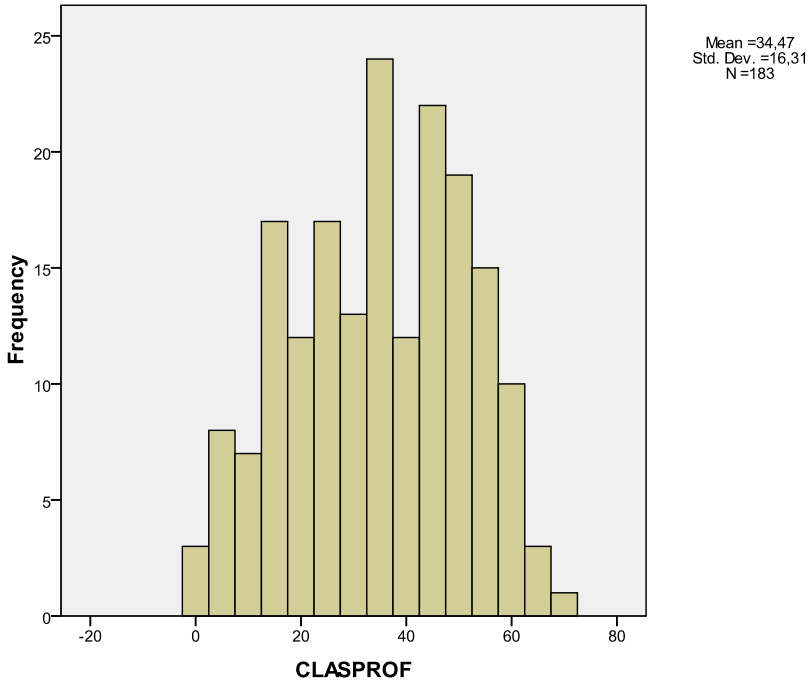


Figura 3.5.1.1. Distribuição das classificações do professor na A).

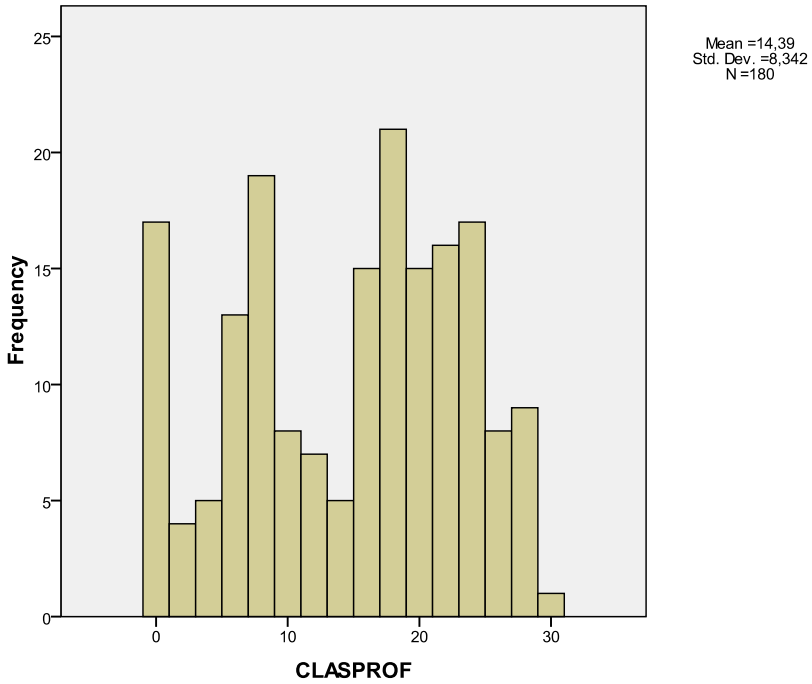


Figura 3.5.1.2. Distribuição das classificações nos textos de resposta à B).

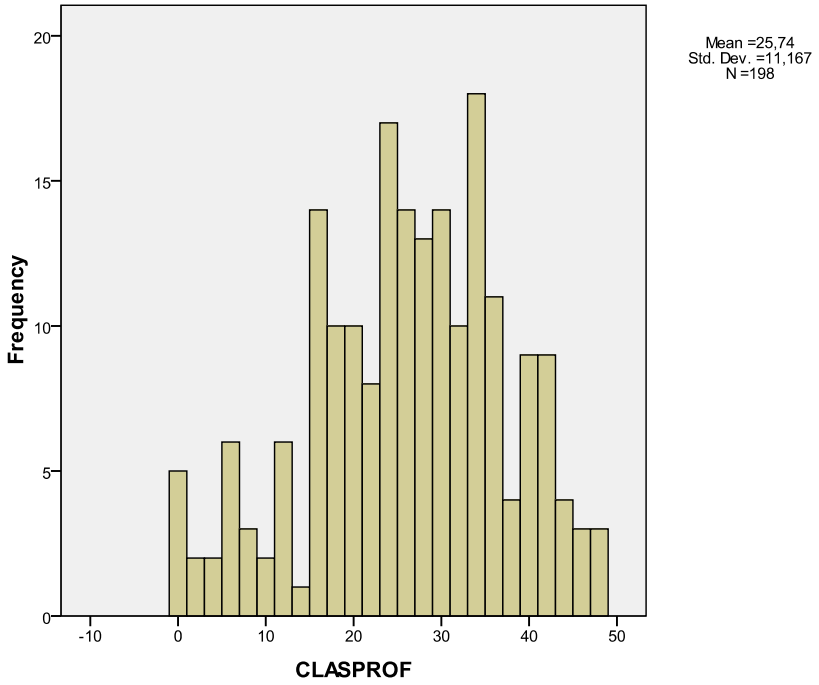


Figura 3.5.1.3. Distribuição das classificações do professor nos textos de resposta ao G3.

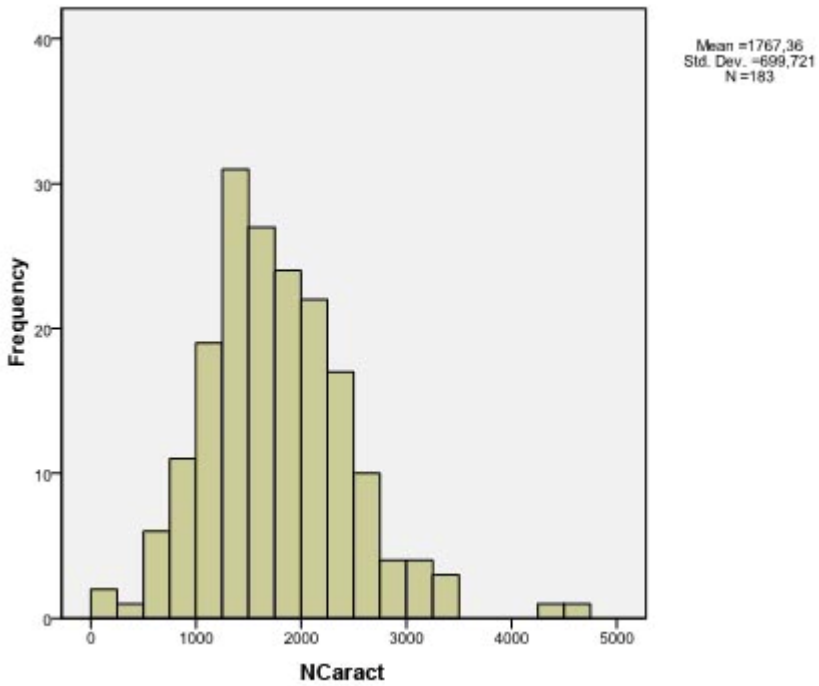


Figura 3.5.1.4. Distribuição do número de palavras no texto para as respostas à A).

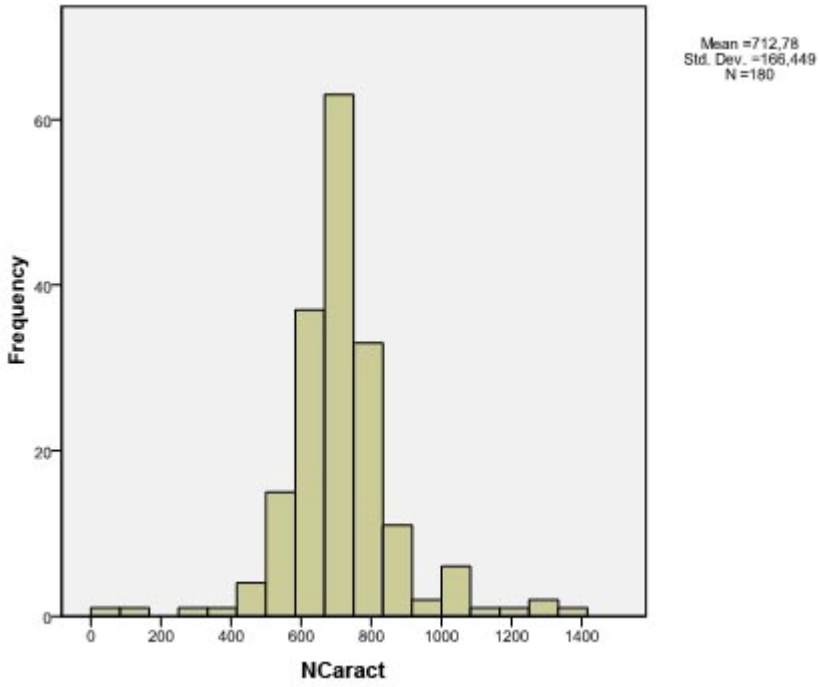


Figura 3.5.1.5. Distribuição do número de palavras no texto para as respostas à B).

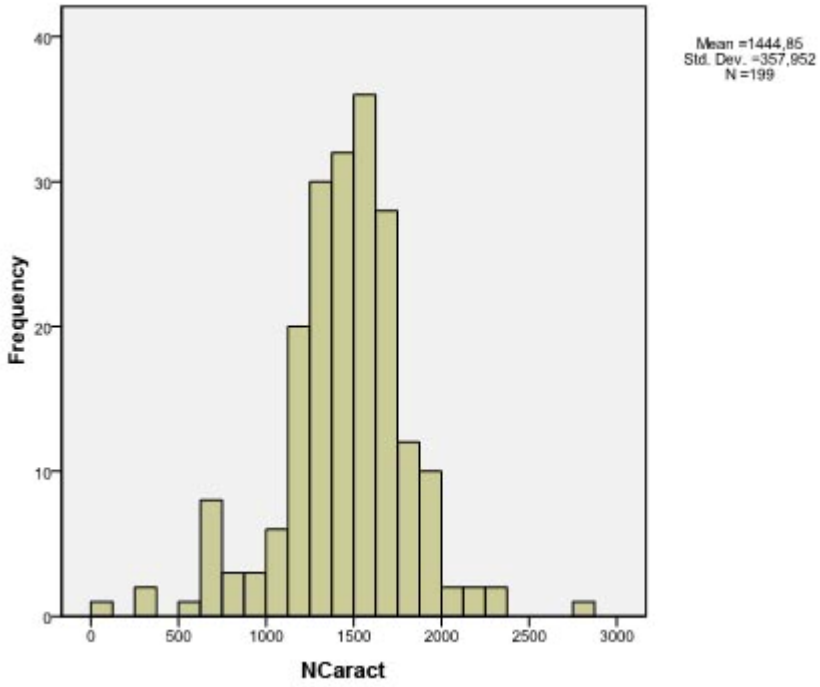


Figura 3.5.1.6. Distribuição das classificações do professor nos textos de resposta ao G3.

A fim de detetar alguma eventual relação ou influência entre as classificações atribuídas pelos professores e o comprimento do texto expresso pelo número de caracteres, número dos textos, seguidamente apresenta-se, para as respostas a cada alínea (A, B) do Grupo I e para o Grupo II, a correlação entre ClassProf e NCaract, bem como gráficos de dispersão, que mostram que no caso A) e Grupo III há efetivamente uma relação significativa entre o resultado da classificação e o comprimento da resposta. Isto deve querer dizer, simplesmente, que os estudantes que têm maior conhecimento da matéria tendem a escrever mais do que os outros.

- Para as respostas à A), obteve-se:

  Correlação (ClasProf, NCaract) = 0.464 (Significativo)
  (Significância = 0.000)

- Para as respostas à B), obteve-se:

  Correlação (ClasProf, NCaract) = 0.115 (Não Significativo)
  (Significância = 0.124)

- Para as respostas ao Grupo III, obteve-se:

  Correlação (ClasProf, NCaract) = 0.377 (Significativo)
  (Significância = 0.009)

As **figuras 3.5.1.7.**, **3.5.1.8.** e **3.5.1.9.** apresentam os gráficos de dispersão das variáveis (ClassProf, NPalTexto).

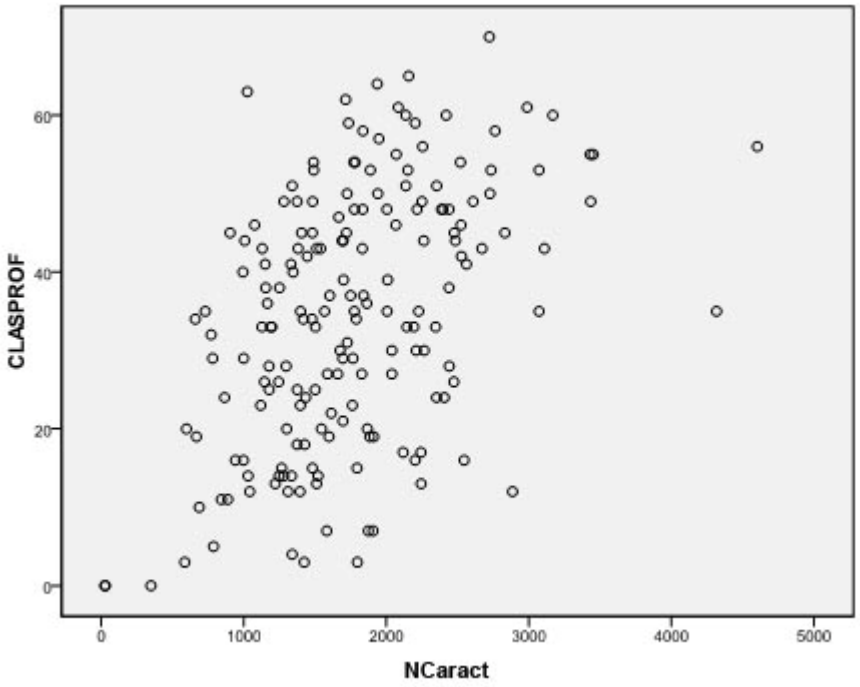


Figura 3.5.1.7. Gráfico de dispersão ClassProf, NPalTexto para A). Deteta-se uma certa tendência crescente.

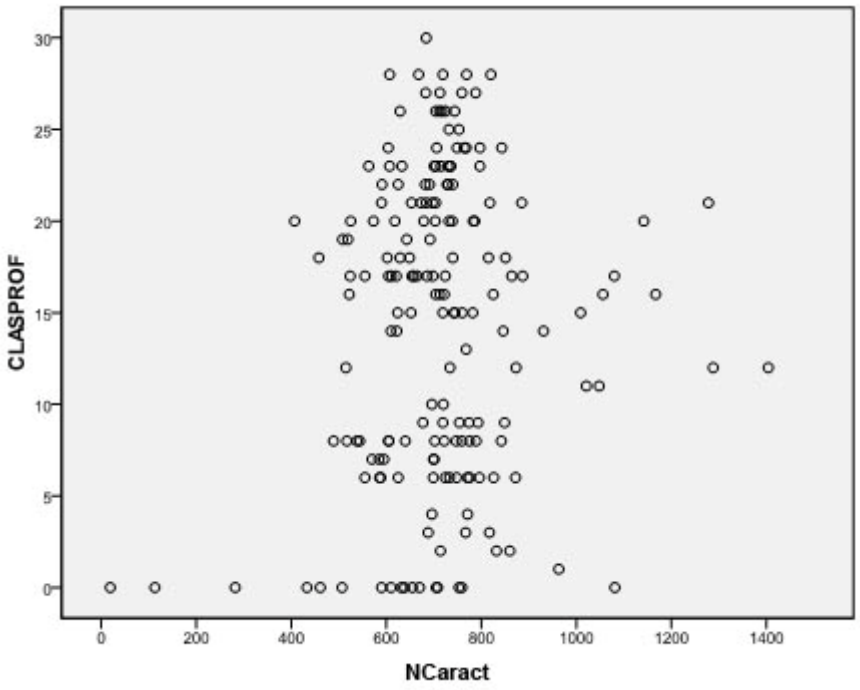


Figura 3.5.1.8. Gráfico de dispersão ClassProf, NPalTexto para B). Não há relação percetível.

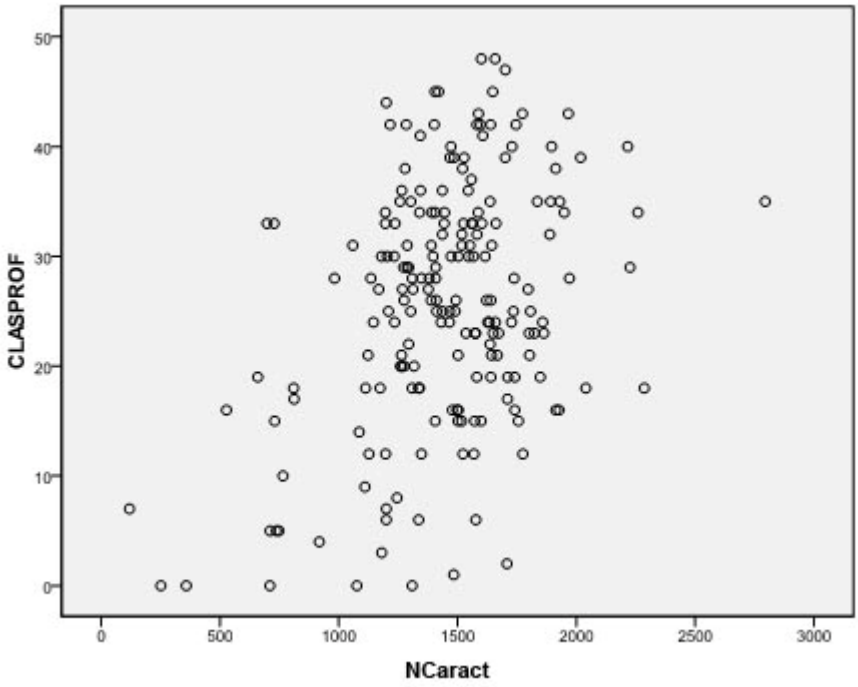


Figura 3.5.1.9. Gráfico de dispersão ClassProf, NPalTexto para o Grupo III. Tendência crescente.

3.5.2. Dados GESTÃO.

Estes dados correspondem a respostas de estudantes universitários da cadeira Gestão de um Instituto Politécnico em que era pedida a caracterização do conceito de Qualidade Total sendo o número total de respostas disponíveis (número de textos válidos)= 187.

A **tabela 3.5.2.1.** resume as variáveis Classificação do Professor e Número de palavras no texto (comprimento).

	Nº Textos	**Mínimos**	**Máximo**	**Média**	**DP**
ClassProf	187	0	100	42.8	22
NCaract	217	0	2059	584	391

Tabela 3.5.2.1. Caracterização das variáveis classificação do professor e número de palavras dos textos de resposta.

As **figuras 3.5.2.1.** e **3.5.2.2.** apresentam os histogramas das variáveis ClassProf e NPalTexto.

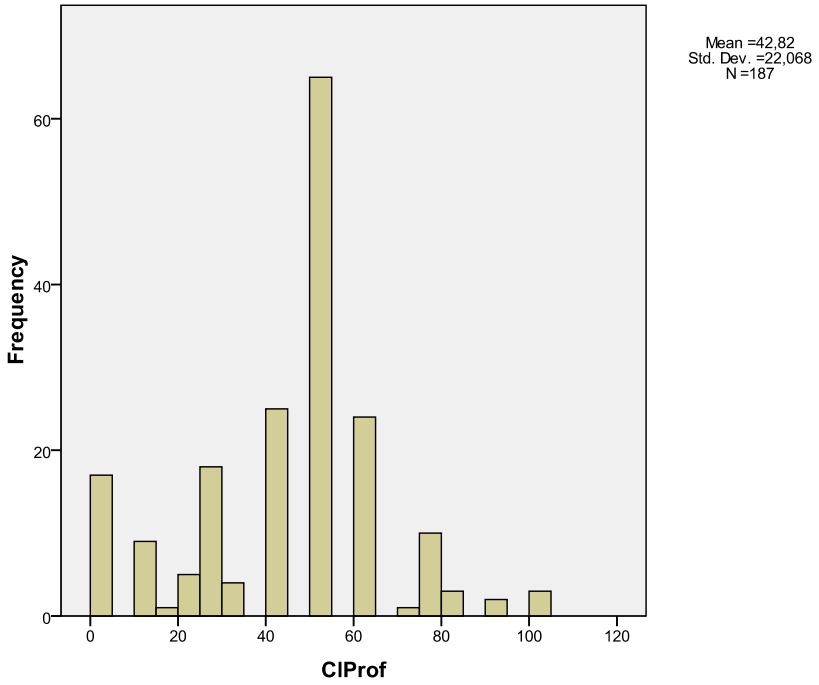


Figura 3.5.2.1. Histograma da classificações dos professores.

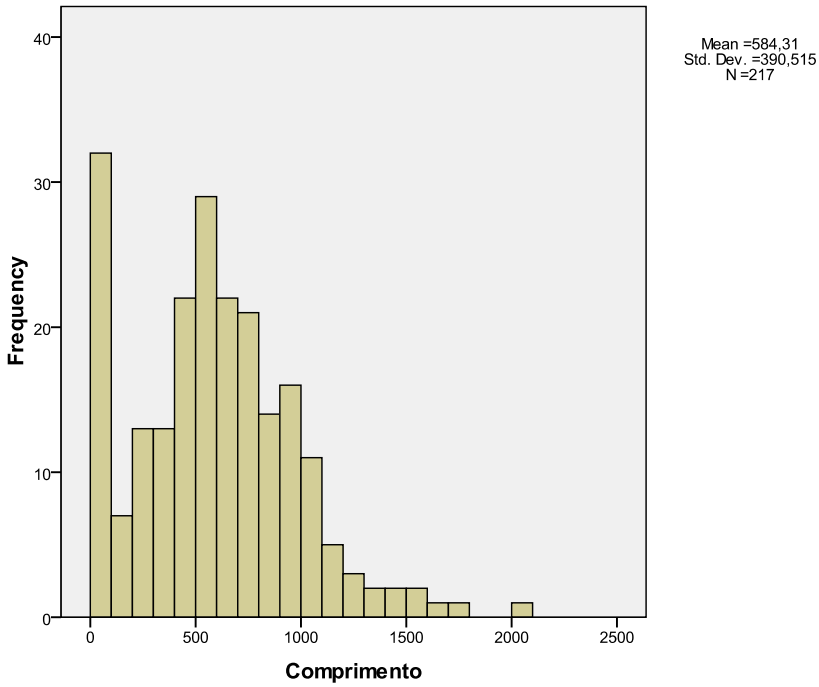


Figura 3.5.2.2. Histograma dos comprimentos dos textos em número de caracteres.

Para estes dados a correlação – significativa – entre as classificações do professor e o número de palavras nos textos (comprimento) é 0.549 (significância= 0.000).

Na **figura 3.5.2.3.** está o gráfico de dispersão respetivo.

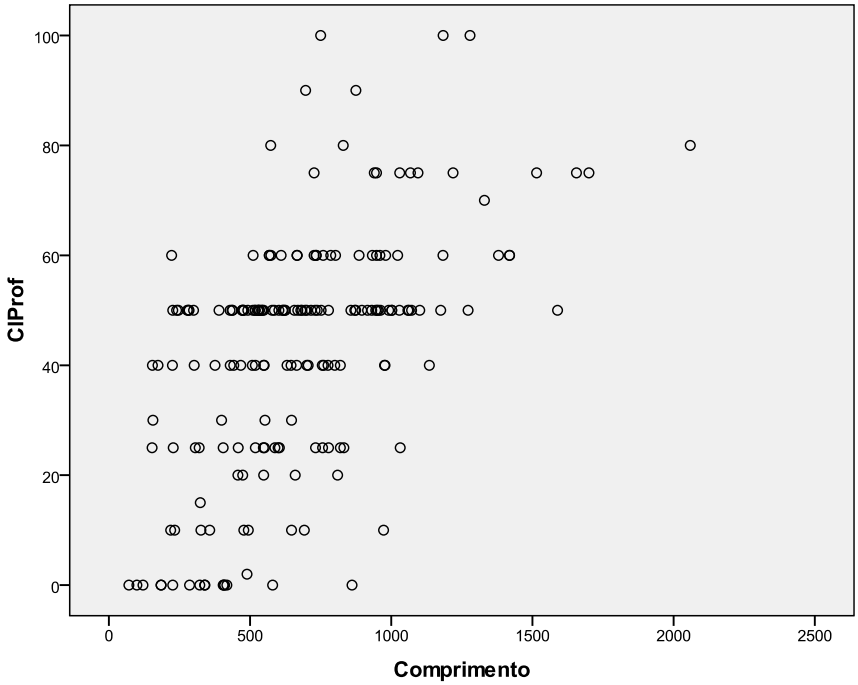


Figura 3.5.2.3. Gráfico de dispersão das classificações do professor em função do número de caracteres dos textos. Tendência crescente.

Fazem-se aqui as mesmas considerações feitas a propósito dos dados EX-MIN. Possivelmente, os estudantes que mais sabem suster a necessidade de escrever mais e, fazem-no.

3.5.3. Dados DISCIPLINA.

Trata-se de 17 textos – não classificados pelo professor – cujo histograma relativo ao número de palavras se apresenta na **figura 3.5.3.1.**

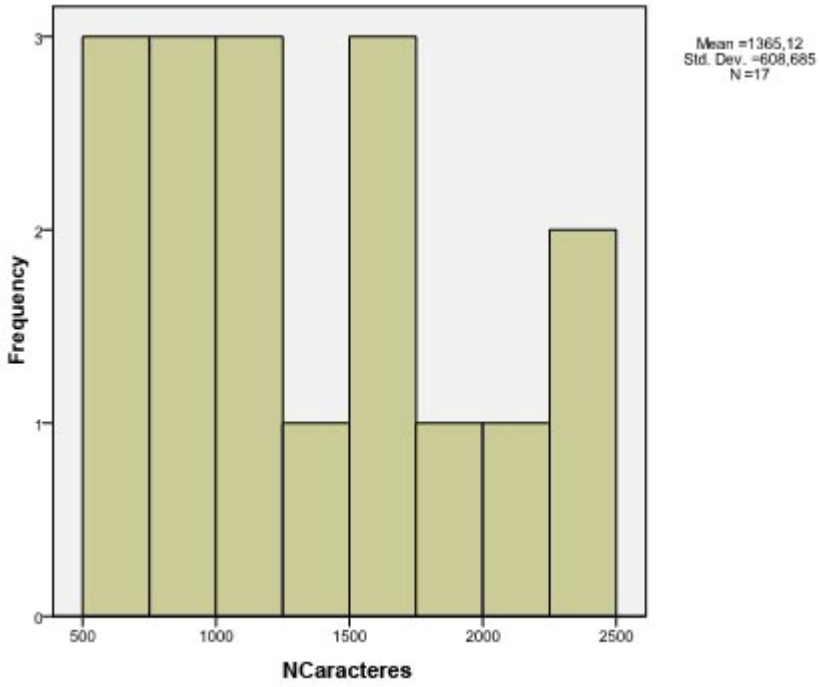


Figura 3.5.3.1. Histograma da distribuição do número de caracteres do texto (comprimento).

Verifica-se que o número médio de caracteres oscila entre 505 e 2415, sendo a média 1365 (DP= 609).

3.6. Comparação de Textos. Análise Descritiva.

Do ponto de vista do apoio aos professores no exercício das respetivas atividades letivas – nomeadamente no desempenho de tarefas ligadas à avaliação contínua baseada em respostas a questões abertas – têm muita importância as tarefas de comparação dos textos produzidos pelos estudantes, Por exemplo, a comparação dos textos produzidos por dois estudantes ao responderem à mesma questão ou a comparação de dois textos produzidos pelo mesmo estudante em instantes diferentes. A primeira situação surge frequentemente quando se pretende classificar respostas a testes de avaliação formativa.

A segunda questão quando se pretende avaliar objetivamente a evolução de um indivíduo ou conjunto de indivíduos.

Existirão diferenças no vocabulário usado pelos dois estudantes? Em que consistem essas diferenças e são elas relevantes para comparar os conhecimentos e comportamentos dos mesmos? Existe alguma evolução na linguagem usada pelo mesmo estudante ou por uma certa turma? Em que consiste essa evolução?

No que se segue, apresentam-se exemplos que documentam o uso dos instrumentos de análise de textos desenvolvidos e integrados no programa designado coletivamente por PAET (Programa de Avaliação Estatística de Textos). Ver o Anexo A o respetivo manual.

Neste número ilustra-se a sua utilização na comparação de textos.

No exemplo a seguir analisado usam-se, a título ilustrativo, dois textos de resposta a uma questão relativa ao conceito de qualidade total dos estudantes de um curso de gestão – dados GESTÃO. Ver descrição dos dados em **3.2.** (Recolha e Registo de Dados). Como se viu, o número de respostas à questão colocada é cerca de 200. A comparação destes textos dois a dois implicaria realizar 200×150= 30.000 comparações. Se cada comparação implicasse 0.5 minuto (muito pouco tempo) isso significaria cerca de 11 dias (24 horas) de trabalho contínuo.

Na prática, os professores leem um teste de cada vez e procuram – socorrendo-se da memória, da capacidade de raciocínio e da experiência – ser coerentes ao atribuir uma classificação com base nessa leitura e nas leituras das respostas já realizadas – sendo quase inevitável um processo de aprendizagem do professor que leva a que os últimos testes estejam a ser classificados com critérios "ligeiramente" diferentes dos iniciais.

Acredita-se que um professor competente e sério é inultrapassável na capacidade de realizar estas comparações tendo em conta toda a informação relevante de que dispõe (as respostas, os conhecimentos, a experiência, o conhecimento dos estudantes e do contexto de ensino) levando em conta questões inacessíveis a sistemas automáticos (qualidade da linguagem, estilo, etc.). Contudo, limitações bem conhecidas da Psicologia – impossibilidade prática de raciocinar sem erro quando o número de conceitos na memória de trabalho excede um certo limiar (cerca de 7) bem como a limitação da própria capacidade dessa mesma memória de trabalho.

No exemplo que a seguir se expõe para ilustrar estas questões apresenta-se também como os instrumentos desenvolvidos em apoio desta tese (ver programa PAET – programa de Análise Estatística de Textos, em Anexo) permitem realizar eficazmente este tipo de comparações, usando recursos ao alcance da generalidade dos professores.

Na **figura 3.6.1.** apresentam-se os dois textos das respostas de dois estudantes do curso de Gestão a que nos referimos em **3.2.** a propósito dos dados.

Pedia-se aos estudantes, no contexto de uma avaliação formativa que, sem apoio de qualquer documento de consulta, caracterizassem o conceito de gestão baseada na Qualidade Total apresentada nas aulas.

Resposta do Estudante nº. 1

A gestão pela qualidade total é uma gestão que dá muita importância á qualidade, esta pesa muito mais que a quantidade. É a sua principal preocupação. Esta gestão aposta na exclusividade e na diferenciação, fazendo vender estes seus dois valores.

Resposta do Estudante nº. 2

A gestão pela qualidade total é a de certa forma a gestão apropriada para que o produto final exceda as expectativas e a qualidade. De acordo com a cadeia reativa de Demming a relação positiva entre a qualidade e a rentabilidade, onde indica que para haver rentabilidade tem que existir também qualidade.

Figura 3.6.1. Textos das respostas de dois estudantes à questão da caracterização do conceito de qualidade total.

Usando a funcionalidade do programa PAET (ver Anexo A o respetivo manual de utilização), na **tabela 3.6.1.** são apresentadas as listas de palavras e respetivas frequências de ocorrência em cada um dos textos (já depois de eliminados os símbolos de pontuação, artigos e demais palavras funcionais) depois de eliminados os símbolos de pontuação, artigos e demais palavras funcionais que constem de uma tabela semelhante à da **figura 3.3.3.**

NºEstudante	Palavra	Frequência
134	qualidade	4
129	gestão	3
129	esta	2
129	muito	2
134	para	2
129	qualidade	2
134	rentabilidad	2
134	gestão	2
134	acordo	1
129	aposta	1
134	apropriada	1
134	cadeia	1
134	certa	1
134	demming	1
129	diferencia	1
129	dois	1
134	entre	1
129	estes	1
134	exceda	1
129	exclusividad	1
134	existir	1
134	expectativa	1
129	fazendo	1
134	final	1
134	forma	1
134	haver	1
134	indica	1
129	mais	1
134	onde	1
129	pela	1
134	pela	1
129	pesa	1
134	positiva	1
129	preocupa	1
129	principal	1
134	produto	1
129	quantidade	1
134	reativa	1
129	seus	1
129	total	1
134	total	1
129	valores	1
129	vender	1
129	importância	1
134	relação	1
134	também	1

Tabela 3.6.1. Frequências de ocorrência das palavras nos textos das duas respostas constantes da **figura 3.6.1.**

Observe-se que a mera realização desta operação envolvendo apenas dois pequenos textos é tão demorada que está fora de questão realizá-la "à mão" de modo sistemático.

Da consulta da **tabela 3.6.1.** constata-se que as palavras mais usadas são "qualidade" (quatro vezes pelo estudante 134 e duas vezes pelo estudante 129).

Observa-se, pois, que o número de palavras dos dois textos é praticamente o mesmo (26 palavras para a resposta do estudante 129 e 34 para a resposta do estudante 134).

Lendo os textos, constata-se que as respostas são bastante semelhantes notando-se, embora, na resposta 129, eventualmente, uma maior capacidade de síntese.

Observando a tabela da **tabela 3.6.1.**, é reforçada a ideia (em princípio) de que as respostas são "bastante semelhantes". Contudo, a comparação pode ser melhorada colocando as frequências das palavras usadas nos textos ao lado umas das outras e tentando ver se há coincidências e diferenças – mais uma operação que é praticamente impossível de realizar na prática por ser extremamente demorada. Ver **tabela 3.6.2.**

Nas colunas **t_129** e **t_143** estão as frequências de ocorrência, nesses textos da totalidade das palavras usadas no conjunto de todos os textos observados.

Por exemplo, para a palavra (com identificador w6) "*demming*" vê-se que essa palavra ocorre no texto do estudante 143 mas não ocorre no texto do estudante 129; a palavra de identificador "*gestão*" ocorre três vezes no texto do estudante 129 e duas no texto do estudante 143.

O grau de intensidade com que as palavras são usadas nos textos (frequências absolutas) traduz certamente "algo" da importância ou adesão dos seus autores ao conceito que essa palavra simboliza. Mas como "valorar" estas frequências do ponto de vista da atribuição de uma "nota" ao estudante que represente fielmente o seu nível de conhecimento acerca do conceito examinado?

A fim de melhorar a comparação, o *software* desenvolvido apresenta ainda (colunas **T129_Usa** e **T143_Usa** da **tabela 3.6.2.**) por cada palavra, "0", quanto ela não ocorre e "1" quando ocorre – o que permite "esquecer" a frequência para apenas nos concentrarmos no uso ou não da palavra.

Somando estas colunas vê-se facilmente que o número de palavras diferentes ou distintas usadas pelos dois respondentes é muito semelhante (21 e 25 respetivamente) havendo apenas quatro coincidências (palavras: "*gestão*", "*pela*", "*qualidade*", "*total*") das quais apenas três (gestão, qualidade, total) são relevantes do ponto de vista do tema.

Mais uma vez se reforça, com esta análise, por um lado a impossibilidade prática da sua realização manual e por outro lado a perceção, neste caso, de que as duas respostas são "**bastante semelhantes**".

SbGraf	Palavra	t_129	t_143	T129_Usa	t143_Usa	Comuns
w1	acordo	0	1	0	1	0
w2	aposta	1	0	1	0	0
w3	apropriada	0	1	0	1	0
w4	cadeia	0	1	0	1	0
w5	certa	0	1	0	1	0
w6	demming	0	1	0	1	0
w7	diferencia	1	0	1	0	0
w8	dois	1	0	1	0	0
w9	entre	0	1	0	1	0
w10	esta	2	0	1	0	0
w11	estes	1	0	1	0	0
w12	exceda	0	1	0	1	0
w13	exclusividad	1	0	1	0	0
w14	existir	0	1	0	1	0
w15	expectativa	0	1	0	1	0
w16	fazendo	1	0	1	0	0
w17	final	0	1	0	1	0
w18	forma	0	1	0	1	0
w19	gestão	3	2	1	1	1
w20	haver	0	1	0	1	0
w21	importância	1	0	1	0	0
w22	indica	0	1	0	1	0
w23	mais	1	0	1	0	0
w24	muito	2	0	1	0	0
w25	onde	0	1	0	1	0
w26	para	0	2	0	1	0
w27	pela	1	1	1	1	1
w28	pesa	1	0	1	0	0
w29	positiva	0	1	0	1	0
w30	preocupa	1	0	1	0	0
w31	principal	1	0	1	0	0
w32	produto	0	1	0	1	0
w33	qualidade	2	4	1	1	1
w34	quantidade	1	0	1	0	0
w35	reativa	0	1	0	1	0
w36	relação	0	1	0	1	0
w37	rentabilidad	0	2	0	1	0
w38	seus	1	0	1	0	0
w39	também	0	1	0	1	0
w40	total	1	1	1	1	1
w41	valores	1	0	1	0	0
w42	vender	1	0	1	0	0
	Total	**26**	**31**	**21**	**25**	**0** **4**

Tabela 3.6.2. As palavras dos textos das respostas dos estudantes 129 e 143 dispostas de forma a facilitar as comparações.

Isto significaria que as duas respostas deveriam ou ter “notas” semelhantes ou mesmo iguais.

Verificou-se, contudo, que a "nota" atribuída ao texto do estudante 129 foi 50% (numa escala de 0 a 100) e que a "nota" atribuída ao texto do estudante 143 foi 25% na mesma escala.

Usando o programa PAET já atrás, referido, as 30.000 comparações e resumos estatísticos relativos às 200 respostas levam cerca de 10 segundos.

Isto é, não só se torna possível realizar essas tarefas como elas são realizadas com um critério invariável e com elevada fiabilidade – tornando admissíveis a multiplicação dos testes com questões em que o estudante elabora a resposta e por esse mesmo facto a transformação de uma avaliação num ato formativo.

3.7. Comparação de Textos usando Biplots.

No contexto das avaliações surge muitas vezes a necessidade de comparar os textos de resposta dos estudantes a uma questão específica. Contudo, por falta de tempo, só muito esporadicamente se realizam essas comparações. Mais precisamente, pode interessar, antes de atribuir uma classificação, agrupar os estudantes do ponto de vista da resposta dada a uma certa questão. Feita a classificação, pode interessar ver se as atribuições de "notas" resultantes do processo habitual são coerentes com os grupos naturais de textos de resposta a essa questão específica.

De resto, a criação de grupos de textos com respostas consideradas semelhantes pode permitir ao professor – independentemente do critério de classificação que venha a seguir – organizar melhor o seu trabalho antes da leitura que procede a atribuição da classificação.

Para ilustrar as ideias aqui expressas considere-se as seguintes 10 respostas à A) do Grupo I dos dados EXMIN atrás considerados.

Esses 10 textos – implicitamente os 10 estudantes que os produziram – têm os identificadores a seguir indicados $C = \{11, 14, 17, 20, 23, 26, 29, 32, 35, 38\}$.

Isto é, desejamos comparar os 10 estudantes atrás indicados do ponto de vista das respostas dadas – dos textos que produziram – à alínea A) do exame de Português.

Usando o programa elaborado para apoio desta tese (PAET – ver manual em Anexo A) um biplot construído com base nesses 10 textos aparece na **figura 3.7.1.**

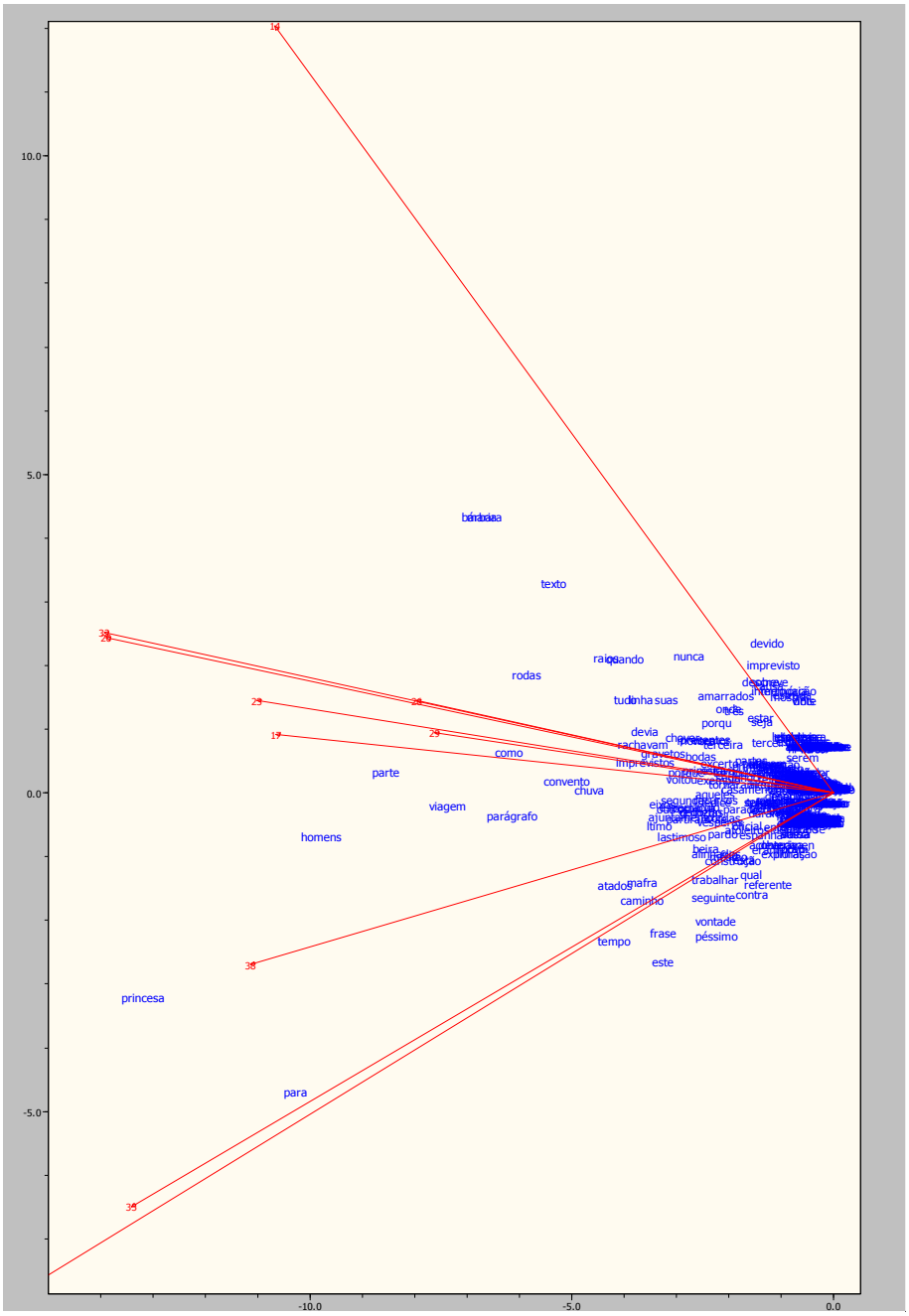


Figura 3.7.1. Biplot dos 10 textos produzidos por 10 estudantes identificados por *C*= {11, 14, 17, 20, 23, 26, 29, 32, 35, 38}.

Na **figura 3.7.1.** as setas a vermelho simbolizam os estudantes, representados pelos respetivos textos, identificados pelos números do conjunto *C*= {11, 14, 17, 20, 23, 26, 29, 32, 35, 38}. As palavras a azul representam palavras usadas nesses textos.

Através do mero exame visual desse gráfico, podem-se imediatamente detetar proximidades e afastamentos entre os textos produzidos por esses 10 estudantes. Repare-se que estes tipos de comparação de respostas dos estudantes são muito difíceis de realizar na

prática até para um pequeno número de textos como é o caso deste exemplo. O grande número de palavras envolvidas (505 palavras) dificulta a operação. Seria necessário realizar $(10\times10)/2 = 50$ comparações, cada comparação envolvendo 500 palavras – sem falar nos cálculos necessários para expressar o resultado da comparação em termos objetivos.

Neste biplot (HJ – biplot de Galindo) os ângulos – os cossenos dos ângulos – entre os textos representam correlações entre esses textos. Quanto menor é o ângulo maior o cosseno e portanto maior a proximidade entre os significados dos dois textos.

No caso da **figura 3.7.1.** detetam-se, pela mera inspeção visual, "grupos naturais" de textos: os grupos formados pelos estudantes/textos {11, 35} – ver canto inferior esquerdo – os textos/estudantes {20, 32} – ou um grupo um pouco menos homogéneo {20, 32, 19, 29}.

Nota-se, também, que a estrutura do biplot é muito condicionada pelo afastamento dos textos/estudantes 11 e 14.

Esta observação visual poderia sugerir que examinássemos os textos dos estudantes que aparecem muito próximos no biplot.

Por exemplo, conviria examinar os textos {20 e 32} que aparecem quase sobrepostos no biplot em causa.

Será essa proximidade um produto apenas do acaso? Aqui o professor pode ter informação extra que explique essa proximidade sugerida ou apontada pelo exame do biplot.

No caso presente, os textos em causa são os que aparecem, respetivamente, nas **figuras 3.7.2.** e **3.7.3.**

Texto do Estudante 20

1. Três dos imprevistos da viagem que a princesa e a sua comitiva fazem, de Montemor a Évora são: a chuva e o frio, exemplo: "voltou a chover, tornaram os atoleiros, partiram-se eixos, rachavam-se como gravetos os raios das rodas"; as rodas que se partiram por causa da chuva, exemplo ".partiram-se eixos, rachavam-se como gravetos os raios das rodas"; e por fim os homens que a princesa viu atados uns aos outros, exemplo; "viu parado um pardo ajuntamento de homens, alinhados na beira do caminho e atados uns aos outros por cordas .".
2. O excerto "turbou-se de tão lastimoso espectáculo de grilhetas, em vésperas das suas bodas, quando tudo devia ser ledice e regozijo" quer dizer que tudo devia estar a ser alegre e feliz para a princesa, mas que não estava. A princesa não estava bem nem feliz porque ia ser obrigada a casar e nem ela pode decidir algo sobre o seu casamento.
3. Um dos recursos estilísticos presentes no último parágrafo é a aliteração, ou seja a repetição de palavras, por exemplo: "constrói-se um convento porque nasceu Maria Bárbara, cumpre-se o voto porque Maria Bárbara nasceu, e Maria Bárbara não viu .". O autor utiliza este recurso estilístico para reforçar a sua ideia.
4. A primeira parte é o primeiro parágrafo, descreve a viagem da princesa e da sua comitiva de Montemor a Évora.
   A segunda parte corresponde ao segundo parágrafo que descreve o tempo que estava e a angústia da princesa.
   A terceira parte corresponde ao terceiro parágrafo que fala sobre Maria Bárbara e o Convento de Mafra mandado construir em nome do seu nascimento e que ela nunca viu.

Figura 3.7.2. Texto do estudante 20 ao responder à A).

Texto do Estudante 32

1. Três dos imprevistos da viagem que a princesa e a sua comitiva fazem são a chuva "Voltou a chover", os raios das rodas que se rachavam "partiram-se os eixos, rachavam-se como gravetos os raios das rodas", "partiram-se os eixos", e os homens parados no caminho " viu parado um pardo ajuntamento de homens, alinhados na beira do caminho e atados uns aos outros por cordas, seriam talvez uns quinze.
2. O autor quando nos diz que a princesa: "turbou-se de tão lastimoso espectáculo de grilhetas, em véspera das suas bodas, quando tudo devia ser ledice e regozijo" refere-se ao espanto e atormento que esta sentiu ao ver aqueles homens em tal situação, pois o seu casamento estava para breve e tendo ela razões para estar alegre, estava admirada com tais homens, pois não sabia o que faziam ali parados e amarrados uns aos outros.
3. Um dos recursos de estilo presentes no último parágrafo é a aliteração "as suas mãos o caldo dos pedreiros". A aliteração é a repetição de um mesmo som.
4. O texto divide-se em três partes lógicas. A primeira parte é "De Montemor a Évora (.) melhor se me casassem na primavera", a segunda parte "cavalgava à estribeira (.) como estará ela agora, passados todos estes anos" e a terceira parte "A princesa já não pensa nos homens (.) se a outra lembrança não serviria a memória".

Uma frase que sintetiza o conteúdo da primeira parte poderá ser: o espanto de Maria Bárbara ao ver tais homens amarrados, durante a sua viagem para Espanha; da segunda parte poderá ser: A explicação do oficial a Maria Bárbara, acerca dos homens amarrados, e da terceira parte: A interrogação de Maria Bárbara do porquê de nunca ter ido a Mafra e de nunca ter visto o convento.

Figura 3.7.3. Texto do estudante 32 ao responder à A).

A funcionalidade incluída no PAET permite calcular os cossenos dos ângulos mútuos entre um qualquer conjunto de objetos. Em particular: textos/textos; textos/palavras; palavras/palavras. Isto é, possibilita-se assim o exame das proximidades entre textos; proximidades de palavras a textos ou proximidades de palavras entre si.

Na **figura 3.7.4.** mostram-se parcialmente os cossenos dos ângulos entre todas as palavras envolvidas nos 10 textos que nos servem de exemplo e esses textos; neste caso, apenas são mostrados cossenos com valor acima de 0.4, sendo possível fixar qualquer limiar.

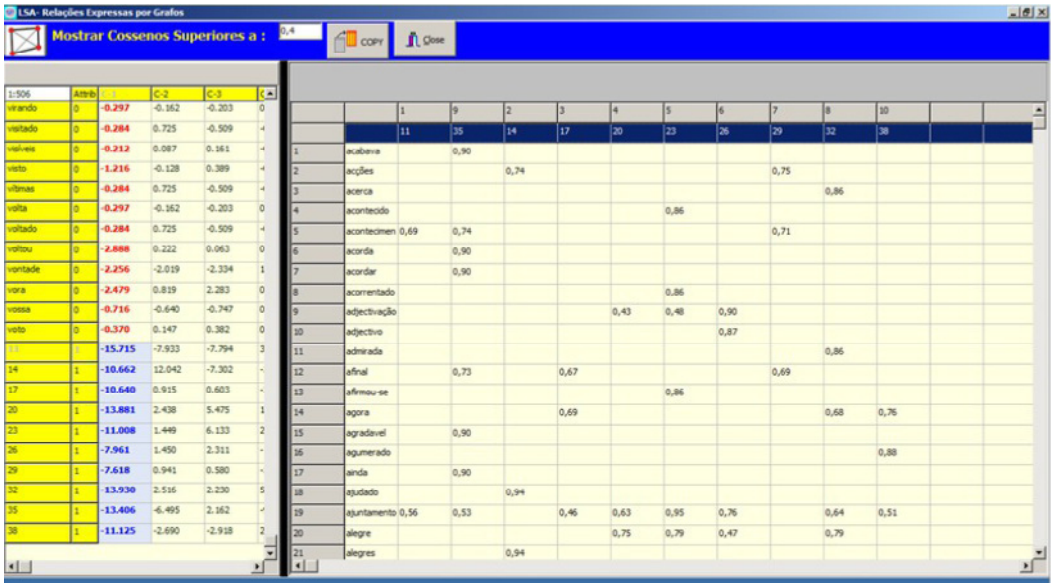


Figura 3.7.4. Cossenos (de valor superior a 0.4) entre os textos do conjunto {11, 14, 17, 20, 23, 26, 29, 32, 35 e 38} e as palavras usadas nesses textos.

Além das proximidades (expressas pelos ângulos e respetivos cossenos) podem também ver-se as palavras associadas a cada um dos grupos de estudantes/textos e que explicam essas associações. Por exemplo, ao grupo dos textos {17, 20, 23, 26, 29, 32} estão associadas palavras como {parte, viagem, convento, chuva, viagem, parágrafo, …}.

Ao texto 38 está muito associada a palavras "princesa".

Deve observar-se que os pontos que representam palavras e textos são projeções no plano em causa (2 dimensões) de objetos que estão em espaços de dimensão muito maior. Pode pois acontecer que certas associações textos/textos ou textos/palavras tenham de ser melhor esclarecidas calculando os ângulos (ou cossenos dos ângulos que tenham em conta a dimensão desses espaços, como foi feito na **figura 3.7.4.**). Pode também suceder que no plano se projetem, como sobrepostos, objetos que no espaço de dimensão > 3 estão muito afastados.

Na **figura 3.7.4.** observam-se, por exemplo, apenas os cossenos – calculados usando a dimensionalidade associada a 85% da informação (dimensão d= 5) – e não apenas

as 2 dimensões da **figura 3.7.3.** Esta figura representa apenas cerca de 60% de toda a informação.

A título de exemplo, nota-se na figura anterior que os textos dos estudantes 11 e 35 (que aparecem muito próximos na **figura 3.7.1.**) fazem ângulos pequenos (cossenos elevados) com as apalavras "acontecimento" (cosseno 0.69 e 0.74) e "ajuntamento" (cossenos 0.56 e 0.53).

Uma busca sistemática – a realizar pelo programa PAET – pode sugerir explicações para as associações de textos descobertas.

Uma possibilidade a explorar no âmbito de uma avaliação formativa é tentar perceber em que medida é que uma atribuição de "notas" aos textos é compatível com o que conhecemos a respeito dos estudantes – ou em que medida uma classificação atribuída automaticamente é compatível com uma classificação tradicional.

Pode ter interesse, por exemplo, projetar todos os textos/estudantes sobre uma direção que represente o texto do melhor estudante ou uma direção que represente os textos dos melhores estudantes – comparando em seguida a relação de ordem assim obtida com a relação de ordem definida pela classificação.

Essa possibilidade está ilustrada na **figura 3.7.5.**, obtida com os mesmos dados mas com um programa diferente para estudar biplots (Vairinhos, 2003). Ver também **Método 2**, em **2.2.4.2.**

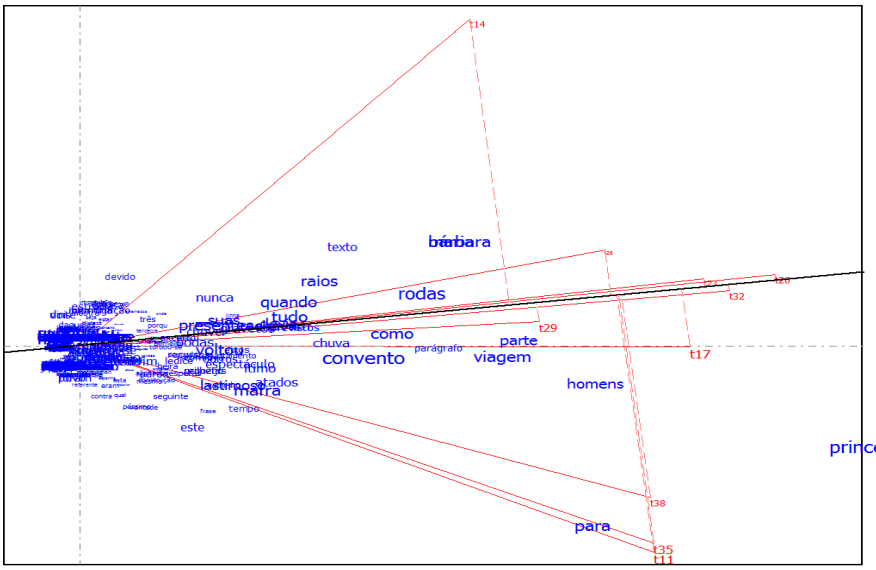


Figura 3.7.5. Projeção da totalidade dos estudantes na direção do estudante 23, com a nota máxima entre os 10 estudantes do conjunto *C*= {11, 14, 17, 20, 23, 26, 29, 32, 35, 38}.

Observando a **figura 3.7.5.** verifica-se que considerando essas projeções, se obtém uma relação de ordem entre os testes que é:

$t_{14} < t_{29} < t_{26} < t_{11} < t_{38} < t_{35} < t_{17} < t_{23} < t_{32} < t_{20}$.

Em síntese, vê-se que o biplot pode sugerir uma relação de ordem entre os textos resultantes destas comparações.

Usando não apenas os 10 textos considerados nas ilustrações anteriores mas a totalidade dos 60 textos de resposta à A), a **figura 3.7.6.** dá o biplot correspondente.

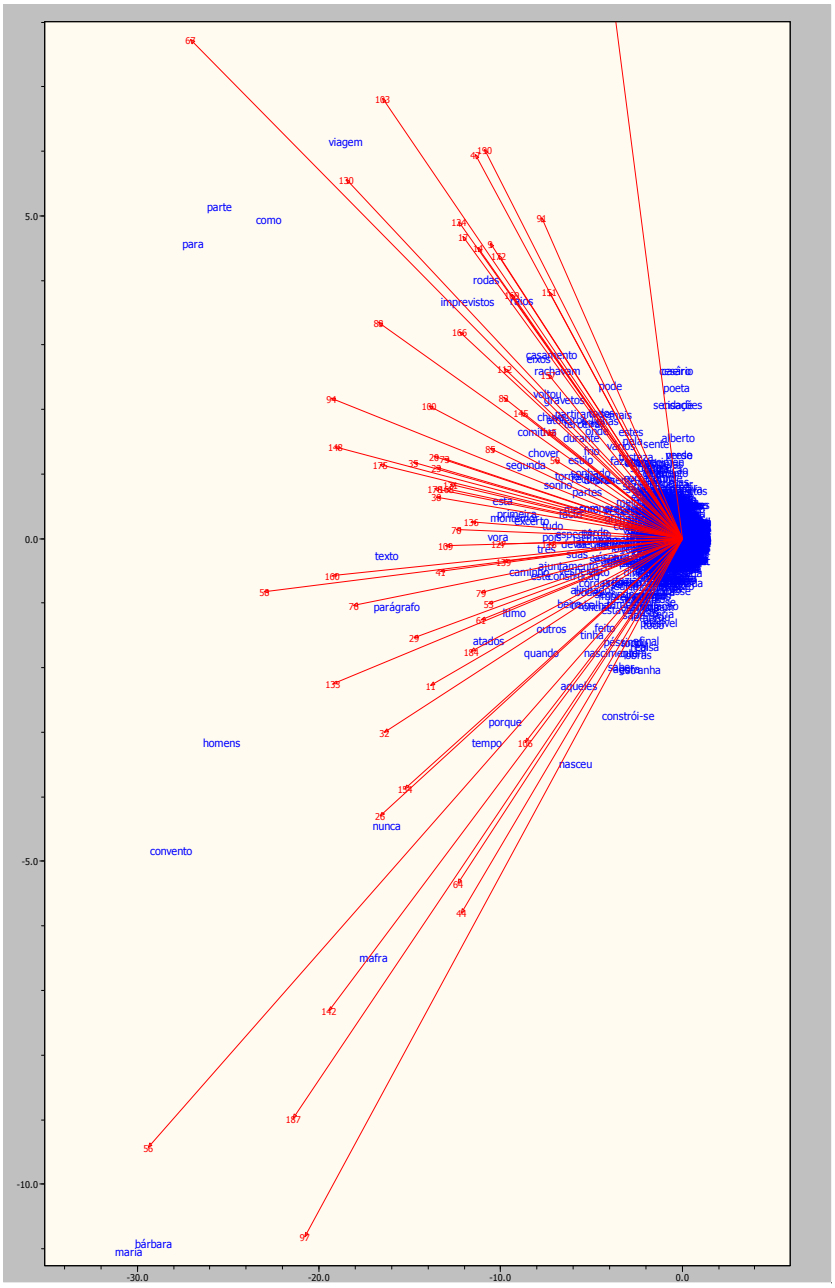
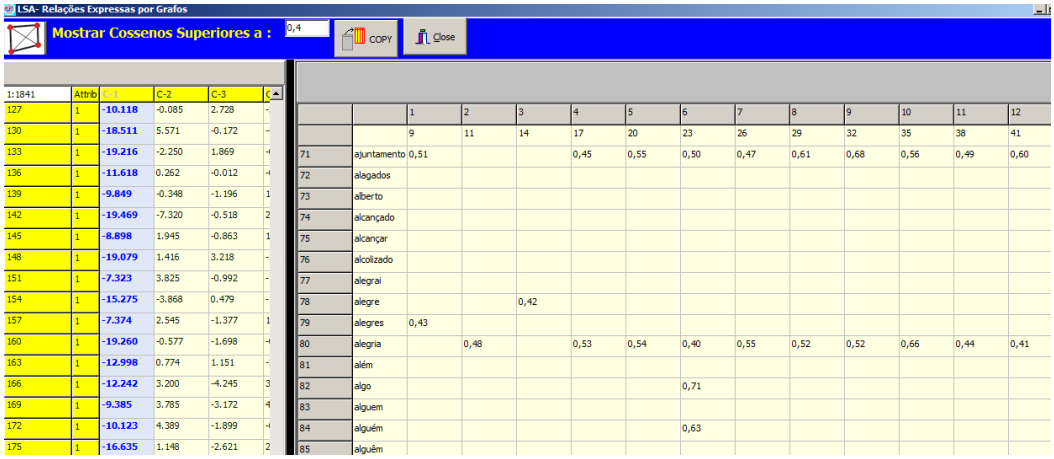


Figura 3.7.6. Biplot construído com os 60 textos de resposta à A) dos dados EX-MIN (prova 639/2ª fase, 2008).

Por exemplo, vê-se que os textos/estudantes {29, 67, 58, 47, 100, 187} são muito semelhantes e que essa semelhança é explicada pela influência de palavras como {*maria, barbara, convento, mafra, nunca, nascem, tempo*}.

Na **figura 3.7.7.** apresentam-se os cossenos dos ângulos entre palavras e textos, calculados usando a dimensionalidade (*d*= 30) correspondentes a cerca de 90% da informação total.



		1	2	3	4	5	6	7	8	9	10	11	12
		9	11	14	17	20	23	26	29	32	35	38	41
71	ajuntamento	0,51			0,45	0,55	0,50	0,47	0,61	0,68	0,56	0,49	0,60
72	alagados												
73	alberto												
74	alcançado												
75	alcançar												
76	alcoizado												
77	alegrai												
78	alegre			0,42									
79	alegres	0,43											
80	alegria		0,48		0,53	0,54	0,40	0,55	0,52	0,52	0,66	0,44	0,41
81	além												
82	algo						0,71						
83	alguem												
84	alguém						0,63						
85	alguêm												

Figura 3.7.7. Cossenos (de valor superior a 0.4) entre as palavras e os 62 textos de resposta à A) calculados com uma dimensão *d*= 30 correspondente a 89% da informação total.

Na figura anterior constata-se, por exemplo, que a palavra "alegria" tem um significado muito associado aos textos 11, 17, 20, 23 entre outros.

3.8. Categorização Automática de Textos em Apoio da Avaliação Formativa.

3.8.1. Introdução.

No contexto de um processo de avaliação contínua, a avaliação formativa tem um papel extremamente importante e neste número estuda-se a possibilidade de usar os métodos de classificação automática (análise de clusters) como auxiliar do trabalho do professor. Isto não significa usar esta metodologia para atribuir automaticamente classificações e avaliar os estudantes mas usar apenas essa metodologia nas tarefas para as quais foi concebido: obter grupos homogéneos de objetos – no caso presente: textos das respostas dos estudantes a questões abertas.

3.8.2. Lógica do Uso das Análises e Cluster.

A lógica que justifica o que se segue é a seguinte: textos de estudantes com baixo nível de conhecimento, acerca de certa matéria, devem ter características semelhantes e testes de estudantes com alto nível de conhecimento devem também ter características semelhantes. Deste modo é de esperar que quando se comparam todos os textos de um conjunto de *k* textos, os textos semelhantes devem agrupar-se em grupos homogéneos, Logo, os métodos de análise de clusters devem em princípio ter utilidade do ponto de vista de identificar, por exemplo, grupos de testes semelhantes antes de proceder à respetiva classificação.

As análises deste tipo implicam sempre o cálculo de um índice de semelhança (ou dissemelhança ou distância entre os objetos a comparar (textos das respostas no caso presente) e um **critério de agregação** que permita ir juntando, mediante certas regras, os objetos, conforme a sua semelhança (ou distância), construindo-se assim uma árvore ou dendograma.

Uma vez que existem diversos índices de agregação e de dissemelhança, as combinações dos índices de dissemelhança com os critérios de agregação conduz facilmente a dezenas de árvores e portanto à dificuldade de decidir qual o par (índice, critério) é "ótimo" para a problemática em questão.

Uma abordagem possível é uma experimentação que procure, no contexto em causa, obter o melhor par (índice, critério).

No caso presente, a metodologia de análise de clusters vai ser aplicada ao resultado da decomposição da matriz de frequências (tabelas de contingência) e portanto a uma representação geométrica dos textos em que cada texto é representado por um vetor cuja dimensão corresponde a uma percentagem significativa da informação (**variabilidade total**). Ver **Capítulo II**.

Para esses dados (que representam os textos num espaço métrico) um par (Índice de Semelhança, Critério de Agregação) muito usado é (Distância Euclidiana, Índice de Ward).

No nosso caso, a percentagem adotada é cerca de 85% que, experimentalmente, se comprovou, com os dados característicos deste tipo de estudo corresponder a uma dimensão da ordem de ½ da dimensão inicial. Isto é, se a matriz dos dados com as frequências tiver p= 100 textos, então $d \cong p/2 = 45$ garante a consideração de grande parte da informação dos dados: entre 85% a 90%.

3.8.3. Experimentação com os Dados GEST.

Considerando a totalidade dos 187 textos de resposta à questão relativa ao conceito de Qualidade Total – foram eliminados os textos em que o estudante, embora tendo entregue a folha de resposta, o texto aparece em branco – e usando o programa PAET (ver Anexo A) que foi dotado de um método de análise de clusters.

O resultado aparece na **figura 3.8.3.1.** em que se pode observar a árvore completa:

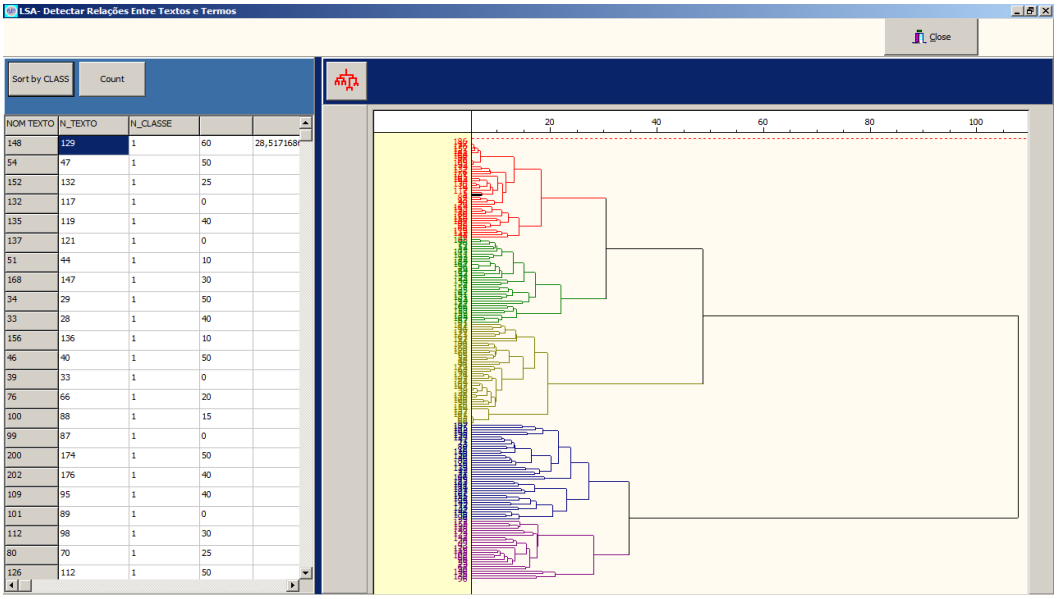


Figura 3.8.3.1. Dendograma dos textos GESTÃO. Escala de dissemelhança na parte superior da janela do lado direito.

Na **figura 3.8.3.1.** pode observar-se que a árvore correspondente aos 187 textos do conjunto de dados GESTÃO aparece cortada ao nível de dissemelhança (distância) 28.5 de forma a considerar apenas 5 classes. Nessa figura cada uma das 5 classes de textos (clusters), existentes a esse nível, aparece pintada de uma cor diferente (vermelho, verde, amarelo, azul e magenta) de forma a permitir uma fácil inspeção visual. No lado esquerdo da figura aparece, para cada uma das classes indicadas, não só o identificador do texto como a classificação atribuída pelo professor da disciplina a essas respostas. Isto é, no lado esquerdo, cada texto aparece associado a dois valores: a classificação que lhe foi atribuída pelo professor e o identificador (número) da classe a que pertence na árvore do lado direito.

A questão que se põe é a de saber se estas classes ou "categorias" obtidas automaticamente "têm algo a ver" com as classificações atribuídas pelo professor,

seguindo o chamado critério "holístico" – resultante da impressão global que a leitura da resposta provoca no professor e que o levou à classificação indicada.

No que se segue, apresentam-se os resultados da análise estatística desses resultados, obtida com o *software* SPSS (versão 17) – SPSS, 2007.

Na **tabela 3.8.3.1.** aparece a tabela de contingência que cruza as classificações atribuídas pelo professor com as classes obtidas automaticamente pela metodologia de análise de clusters, tendo no cruzamento das linhas com as colunas o número de textos classificados nessa célula.

Verifica-se que o professor concentrou as respetivas classificações nos valores 40%, 50% e 60% aparecendo 16 textos com classificação zero e 3 com a classificação máxima de 100%.

ClProf * Classe Crosstabulation

Count

		Classe					
		1	2	3	4	5	Total
ClProf	0	10	5	1	0	0	16
	2	0	1	0	0	0	1
	10	4	3	0	1	1	9
	15	1	0	0	0	0	1
	20	1	1	0	1	2	5
	25	6	8	0	2	2	18
	30	4	0	0	0	0	4
	40	6	6	3	4	6	25
	50	9	6	27	17	6	65
	60	1	4	6	8	5	24
	70	0	0	0	1	0	1
	75	0	0	2	5	3	10
	80	0	0	2	1	0	3
	90	0	1	1	0	0	2
	100	0	1	1	1	0	3
Total		42	36	43	41	25	187

Tabela 3.8.3.1. Tabela de contingência cruzando as classificações do professor com as classes do processo de análise de clusters (distancia euclidiana, critério de WARD).

A **tabela 3.8.3.2.** dá o resultado do teste do χ^2 para decidir se existe alguma relação ou dependência entre os resultados atribuídos pelo professor e as classes obtidas pelo procedimento de análise de clusters. Verifica-se claramente uma associação entre linhas e colunas, o que contraria a ideia de que essa classificação possa ser um mero resultado do

azar, conforme é mostrado através de três testes que produzem o mesmo resultado (Qui-quadrado, Razão de verosimilhança e Associação Linear-Linear). Em todos os casos é rejeitada a hipótese nula de ausência de relação, concluindo-se então que há evidência compatível com a ideia de que as duas classificações estão relacionadas.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	112,116[a]	56	0
Likelihood Ratio	121,096	56	0
Linear-by-Linear Association	33,437	1	0
N of Valid Cases	187		

a. 64 cells (85,3%) have expected count less than 5. The minimum expected count is 13.

Tabela 3.8.3.2 Resultado do teste do qui-quadrado, obtido com o SPSS.

A **tabela 3.8.3.3.** mostra, dentro de cada classe resultante da análise de clusters (1, 2, 3, 4, 5), os valores de diversas estatísticas básicas relativas às classificações dos textos atribuídos pelo professor. Verifica-se que, apesar da variabilidade, as classes aparecem associadas a médias bem definidas. As classes 3 e 4 não se distinguem bem.

Report

ClProf

Classe	Mean	N	Std. Deviation	Minimum	Maximum	Range
1	26,07	42	19,46	0	60	60
2	33,94	36	24,683	0	100	100
3	54,19	43	15,273	0	100	100
4	53,54	41	16,854	10	100	90
5	46,6	25	17,483	10	75	65
Total	42,82	187	22,068	0	100	100

Tabela 3.8.3.3. Estatísticas da classificação dos professores correspondentes às classes obtidas por análise de clusters (obtida com o software SPSS).

Em síntese, esta experiência mostrou claramente que existe uma relação bem definida entre as classificações do professor e o resultado da análise de clusters, apoiando a

sugestão de que faz sentido que o professor use o resultado de uma análise deste tipo "antes" de atribuir as classificações finais. Fica por definir o critério de escolha do nível de corte da árvore (28.5, neste exemplo).

Isto é, não se está a sugerir que de modo acrítico e automático o professor atribua uma nota única a todos os textos de uma classe mas que proceda a vários cortes (quanto mais baixo no índice de dissemelhança – eixo horizontal na **figura 3.8.3.1.**) antes de examinar os testes segundo os seus próprios critérios.

3.8.4. Experimentação com os Dados do EX-MIN.

Considerando agora as respostas à A) do grupo I do conjunto de respostas às questões relativas ao MEMORIAL DO CONVENTO de JOSÉ SARAMAGO do exame de Português, 2008, prova 639/2ª fase, repetindo a sequência de análises cluster usada em **3.8.3.** para os dados de GESTÃO, obtêm-se os seguintes resultados usando o mesmo *software*: PAET e SPSS vs 17 (ver Anexo A).

Na **figura 3.8.4.1.** aparece o dendograma relativo às respostas dos estudantes, cortadas ao nível 27 de modo a evidenciar cinco classes de textos semelhantes (ao nível do corte).

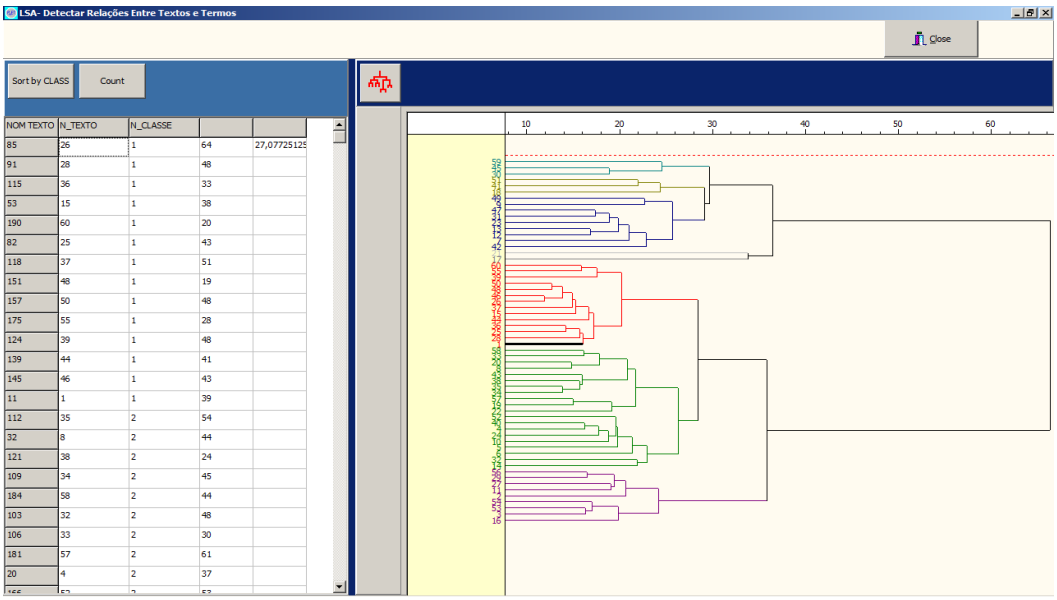


Figura 3.8.4.1. Resultado da análise de clusters dos textos da resposta à questão GI A) dos exames nacionais de Português 2008.

Observe-se que as distâncias envolvidas variam entre 10 e 50 – escala horizontal da parte superior direita da tabela anterior.

A **tabela 3.8.4.1.** mostra o resultado do cruzamento das classificações atribuídas pelos professores contratados pelo ministério com as classes ou grupos gerados automaticamente pela análise de clusters.

Esta árvore – dendograma – foi obtida calculando as distâncias ou dissemelhança entre textos, usando a distância euclidiana. Para agregar grupos de textos usou-se o critério de Ward. Para outras combinações (distância, critério) as árvores poderiam ser diferentes.

ClProf * Classe Crosstabulation

Count

		Classe								
		1	2	3	4	5	6	7	8	Total
ClProf	19	1	0	0	0	0	0	0	0	1
	20	1	0	0	0	0	0	0	0	1
	21	0	0	0	0	0	1	0	0	1
	24	0	1	0	0	0	0	0	0	1
	26	0	0	0	0	1	0	0	0	1
	28	1	0	0	0	0	0	0	0	1
	30	0	1	0	0	0	0	0	0	1
	33	1	0	0	0	1	0	0	0	2
	34	0	1	0	0	0	0	0	0	1
	35	0	2	0	0	0	1	0	0	3
	36	0	0	0	0	1	0	0	0	1
	37	0	1	0	2	0	0	0	0	3
	38	1	0	0	1	0	0	0	0	2
	39	1	0	0	0	0	0	0	0	1
	40	0	0	1	1	0	0	0	0	2
	41	1	0	0	0	0	0	0	0	1
	43	2	2	0	1	1	0	0	0	6
	44	0	3	0	0	1	0	0	0	4
	45	0	2	0	0	0	0	0	1	3
	46	0	0	1	0	0	0	0	0	1
	48	3	2	0	0	0	1	0	0	6
	49	0	1	0	1	1	0	0	0	3
	50	0	0	0	2	0	0	0	0	2
	51	1	0	0	0	0	0	0	0	1
	53	0	1	0	0	1	0	0	0	2
	54	0	1	0	0	1	0	0	0	2
	56	0	0	0	1	0	0	1	0	2
	57	0	1	0	0	0	0	0	0	1
	59	0	0	0	0	1	0	0	0	1
	60	0	0	1	0	0	0	0	0	1
	61	0	1	0	0	0	0	0	0	1
	64	1	0	0	0	0	0	0	0	1
Total		14	20	3	9	9	3	1	1	60

Tabela 3.8.4.1. Tabela de contingência cruzando as classificações atribuídas pelos professores e as classes obtidas por análise de clusters.

Constata-se agora que, face à distribuição das contagens do número de textos pelas células (classificação, classe), ao contrário do **exemplo 3.8.3.1.**, não se detetam quaisquer regularidades que indiciem evidência de dependência. Este resultado tem possivelmente a ver com o modo como essas classificações são atribuídas (de acordo com regras rígidas) e, sobretudo, pela pequena dimensão da amostra (n= 60) que, face ao número de classes = 8 e ao número de classificações distintas, torna esta experiência inconclusiva.

No contexto de uma avaliação formativa, contudo, considera-se que o exame dos conteúdos dos agrupamentos (clusters) poderia ser um *input* importante para a organização do trabalho de classificação do professor, dada a semelhança dos textos dentro de cada grupo. Isto é, antes de classificar, o professor poderia agrupar os testes segundo o critério sugerido pela árvore antes de proceder à classificação. Nesse caso, ao atribuir um texto a um estudante, sabe que está a fazê-lo no contexto de um grupo de estudantes considerado homogéneo. Antes de optar por uma classificação final proceder rotineiramente a um estudo estatístico como o sugerido aqui.

3.9. Avaliação Automática de textos usando o Método LSA/Biplots.

3.9.1. Introdução.

No que se segue usa-se a abreviatura LSA/Biplot para designar a representação dos textos baseada na decomposição em valores e vetores singulares da matriz X de frequências de um conjunto de n palavras sobre um conjunto de p textos. Já se viu no **Capítulo II** que a visualização desta decomposição em espaços até à dimensão d= 3 se designa habitualmente por biplot mas podemos imaginar biplots de qualquer dimensão d= 3, …, p, ainda que estes não se possam visualizar com os sistemas cartesianos habituais. As respetivas visualizações e conteúdos são possíveis usando biplots cilíndricos (Vairinhos & Galindo, 2012).

Neste número apresentam-se os resultados de diversas análises a que se submeteram os dados disponíveis, procurando-se validar não só o funcionamento do programa desenvolvido, como pôr em evidência os problemas que a implementação prática de um sistema de análise estatística de textos ao serviço da atividade de professores e estudantes levantar.

Apresentam-se, pois, em particular, os resultados da aplicação aos dois conjuntos de dados (EX-MIN, GESTÃO). Os resultados gerados pelo programa desenvolvido

(PAET) para este projeto, comparando esses resultados com os das classificações dos professores, na perspetiva do esquema de validação deste sistema recomendado na literatura.

O programa PAET contém uma implementação do esquema de classificação automática baseado na decomposição da matriz de frequências em valores e vetores próprios, sugerida no **Capítulo II**. Como se refere nesse capítulo a propósito da classificação automática baseada no método da ASL / Biplots – ver **figura 2.2.3.3.** – os dados são organizados do seguinte modo – ver **figura 3.9.1.1.**

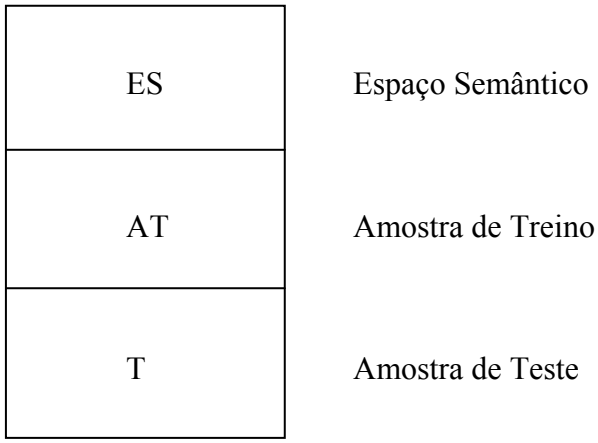


Figura 3.9.1.1. Organização dos dados para treinar um classificador automático.

Isto é, os dados envolvidos na análise são divididos em três partes: os textos usados para construir o Espaço Semântico (ES), os textos para construir a Amostra de Treino (AT) e os textos (T), que constituem a amostra de treino.

O ES resulta da decomposição em valores e vetores singulares da matriz de frequências, relativa ao primeiro conjunto de textos.

Nas experiências descritas no que se segue - em **3.9.2.** e **3.9.3.** - este primeiro conjunto é formado por 90 textos extraídos de manuais de ensino usados para explicar a obra de José Saramago "Memorial do Convento" aos estudantes do Ensino Secundário, no ano letivo 2007/2008.

Os segundo (AT) e terceiro (T) conjuntos são textos classificados por professores e usados, respetivamente, para **treinar** e **testar** o classificador, permitindo estimar a respetiva validade.

Nas experiências descritas em **3.9.2.** e **3.9.3.**, os textos a usar para este efeito são os 61 textos de resposta á questão GI A) dos exames oficiais de Português de 2008 (prova

639/2ª fase). Estes 61 textos são divididos em duas partes através da fixação do valor do parâmetro % da amostra de treino. Por exemplo, no programa PAET assume-se como valor por defeito, para este parâmetro, 50% o que corresponde a usar 31 testes para treinar o classificador e 30 testes para o respetivo teste.

Nas experiências descritas em **3.9.3.** e **3.9.4.**, o Espaço Semântico é o mesmo (90 textos usados no ensino do Memorial do Convento de José Saramago aos estudantes do secundário). Contudo, em cada uma dessas experiências usam-se as respostas às questões GI A) e B) que são posicionadas sobre o mesmo espaço semântico.

Finalmente, no número **3.9.6.** comparam-se as classificações atribuídas pelos professores contratados pelo Ministério com as classificações atribuídas por um professor do Ensino Secundário que aceitou participar nesta investigação, classificando esses mesmo 61 textos.

Em síntese, nas análises de dados abaixo descritas vão estar envolvidos: 90 textos relativos ao ES, 30 textos relativos à AT e 30 textos relativos à amostra T.

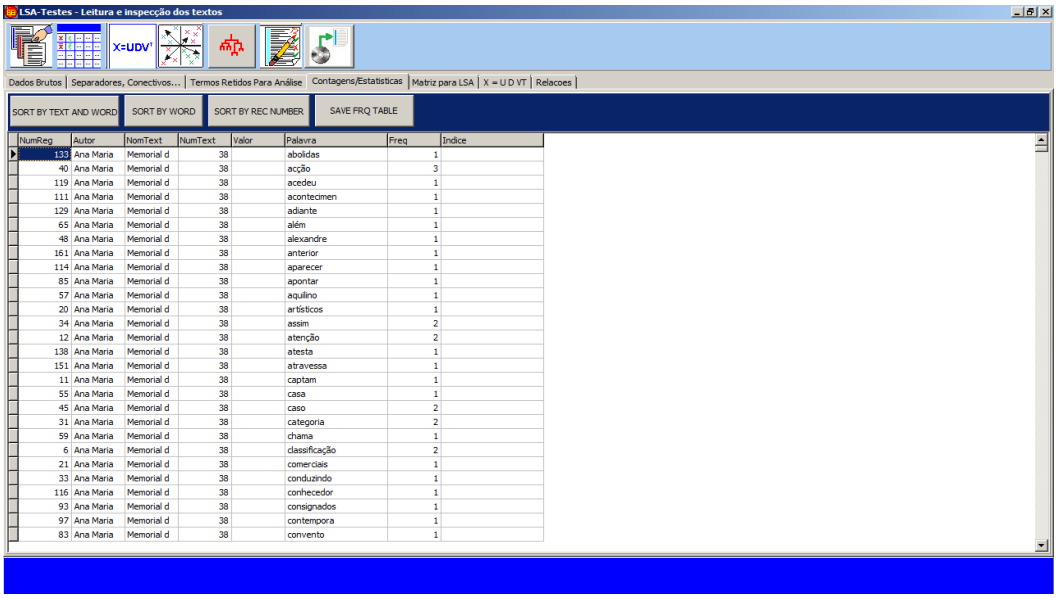
Embora o número total de textos disponíveis seja cerca de 600 (ver **3.7.**) o facto é que esses textos estão divididos pelas provas de várias disciplinas para as quais não foi possível obter recursos que permitissem a digitalização dos manuais de ensino e textos de resposta correspondentes e consequente construção dos respetivos espaços semânticos.

3.9.2. Construção do Espaço Semântico (ES).

Como se disse, o ES construído é relativo ao romance Memorial do Convento de José Saramago e é formado pelas coordenadas das palavras identificadas em 90 textos extraídos de manuais de ensino da obra deste escritor, destinada ao ensino secundário no ano dos exames referidos.

Na **figura 3.9.2.1.** está uma imagem construída pelo programa PAET no final da operação de extração de palavras destes 90 textos e respetivas contagens. Por exemplo, a palavra "santo" aparece 1 vez no texto 42, 3 vezes no texto cujo identificador é 688, 2 vezes no texto identificado por 701 – aparecendo ainda em outros 8 dos 90 textos usados.

A matriz $X(n, p)$ resultantes destas contagens tinha como dimensão finais n= 5678, p= 90. Isto é, o ES em questão abrangia 5678 palavras detetadas nos 90 textos envolvidos.



NumReg	Autor	NomText	NumText	Valor	Palavra	Freq	Indice
133	Ana Maria	Memorial d	38		abolidas	1	
40	Ana Maria	Memorial d	38		acção	3	
119	Ana Maria	Memorial d	38		acedeu	1	
111	Ana Maria	Memorial d	38		acontecimen	1	
129	Ana Maria	Memorial d	38		adiante	1	
65	Ana Maria	Memorial d	38		além	1	
48	Ana Maria	Memorial d	38		alexandre	1	
161	Ana Maria	Memorial d	38		anterior	1	
114	Ana Maria	Memorial d	38		aparecer	1	
85	Ana Maria	Memorial d	38		apontar	1	
57	Ana Maria	Memorial d	38		aquilino	1	
20	Ana Maria	Memorial d	38		artísticos	1	
34	Ana Maria	Memorial d	38		assim	2	
12	Ana Maria	Memorial d	38		atenção	2	
138	Ana Maria	Memorial d	38		atesta	1	
151	Ana Maria	Memorial d	38		atravessa	1	
11	Ana Maria	Memorial d	38		captam	1	
55	Ana Maria	Memorial d	38		casa	1	
45	Ana Maria	Memorial d	38		caso	2	
31	Ana Maria	Memorial d	38		categoria	2	
59	Ana Maria	Memorial d	38		chama	1	
6	Ana Maria	Memorial d	38		classificação	2	
21	Ana Maria	Memorial d	38		comerciais	1	
33	Ana Maria	Memorial d	38		conduzindo	1	
116	Ana Maria	Memorial d	38		conhecedor	1	
93	Ana Maria	Memorial d	38		consignados	1	
97	Ana Maria	Memorial d	38		contempora	1	
83	Ana Maria	Memorial d	38		convento	1	

Figura 3.9.2.1. Uso do PAET para contagem de palavras dos 90 textos do ES usados para construir o ES.

Realizada a operação de decomposição em valores e vetores singulares da matriz de frequências correspondente a este ES, verifica-se que 45 das p= 90 dimensões iniciais acumulam 85.4% de toda a variância/variabilidade/informação contida nos dados.

Isto significa que, em vez da dimensão inicial p= 90 podemos, neste caso, representar textos e palavras num espaço de dimensão d= 45, preservando quase toda a informação dos dados, obtendo-se uma notável compreensão dos mesmos (redução da dimensionalidade em cerca de 50%, com perda apenas de cerca de 15% da variabilidade).

Constata-se também que as duas primeiras dimensões (plano 1, 2) representam 27% da informação total. A **figura 3.9.2.2.** apresenta a imagem deste ES captada por um biplot construído apenas com os dois primeiros eixos.

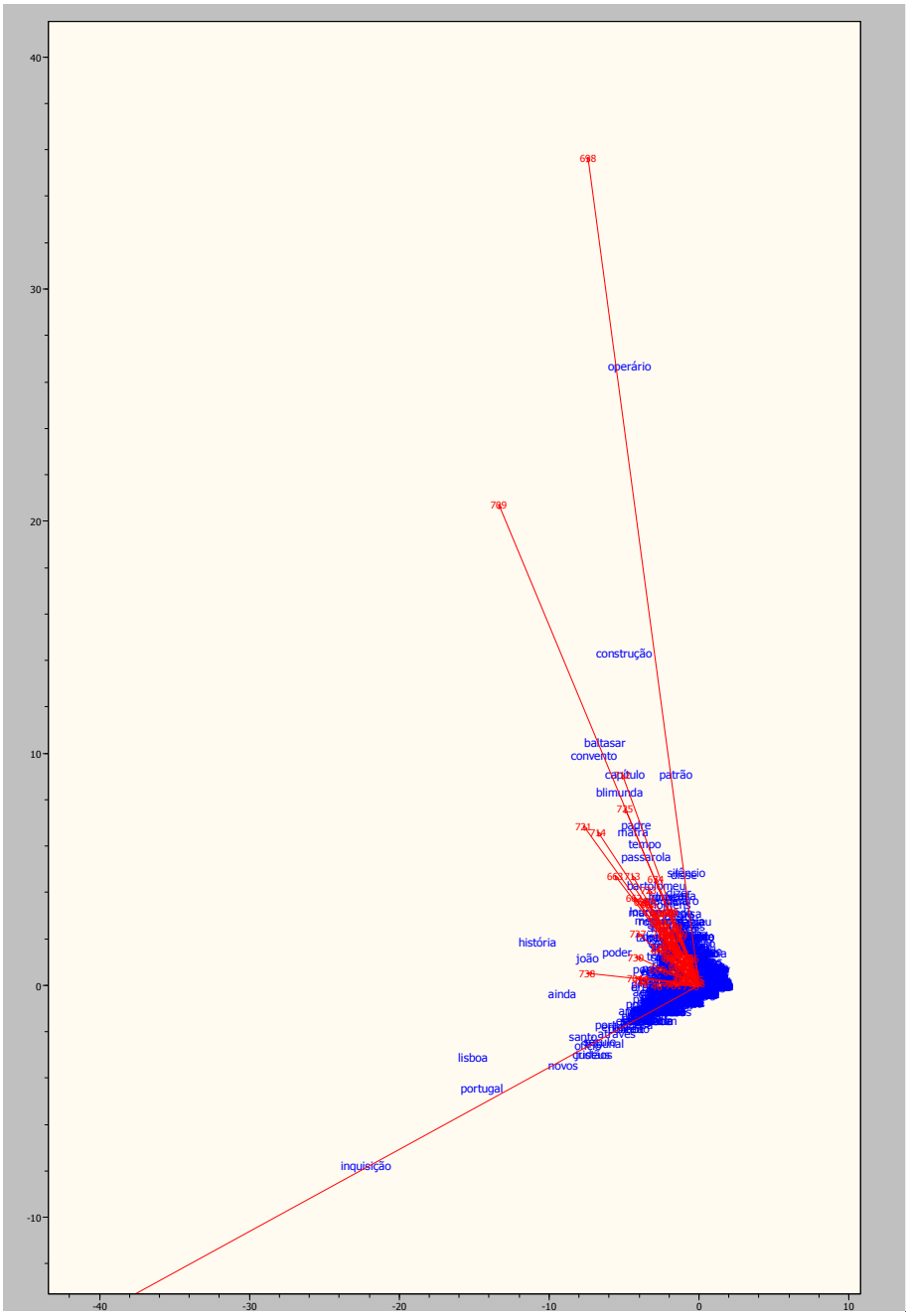


Figura 3.9.2.2. Biplot correspondente aos dois primeiros eixos do espaço semântico relativo a 90 textos e 5678 palavras envolvidos no ensino do Memorial do Convento de José Saramago, representando 27% da informação total.

Examinando esta primeira versão do ES constata-se que o texto número 471 – que é o de maior número de palavras do conjunto dos 90 – comporta-se como um "*outlier*" uma vez que condiciona só por si toda a geometria do ES. Também se vê que este texto aparece associado essencialmente à palavra "inquisição". Examinando esta palavra – que ocorre em 18 outros textos – a sua frequência no texto 741 é 24. Isto sugeriu que se construísse o ES eliminando o texto 741. Verificou-se também que o texto 698 estava excessivamente

dependente da palavra "operário" pelo que o ES foi refeito sem esses dois textos, retendo-se então o ES que está representado, usando os dois primeiros eixos, pelo biplot da **figura 3.9.2.3.** seguinte. Examinando essa figura, verifica-se um maior "equilíbrio" no sentido de que a respetiva estrutura não está excessivamente dependente de um só texto ou palavra.

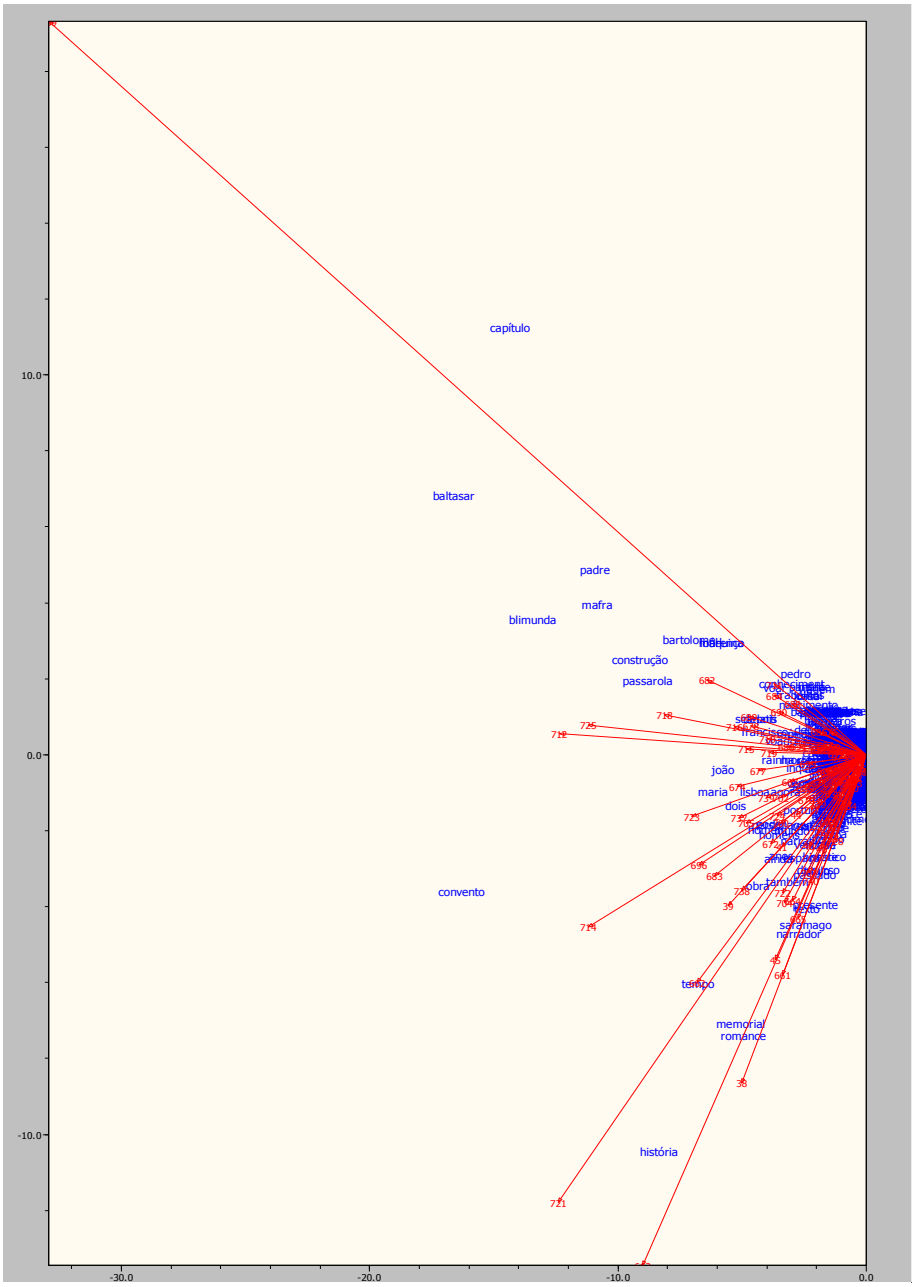


Figura 3.9.2.3. Espaço semântico retido para análise, baseado em 88 textos usados no ensino do Memorial do Convento de José Saramago.

A tabela de frequências $X(n, p)$ associada a este novo biplot – a usar nas experiências deste capítulo – tem dimensão n= 5015 palavras e p= 88. Com d= 45 dimensões garante-se 82% da informação ou variância de X.

Examinando este novo ES vê-se que nele destacam-se os textos 700, 682, associados a palavras como "capítulo", "Blimunda"; os textos 38, 45, 662, 663, 721 associados a palavras como "história", "memorial", "romance", "tempo", "Saramago", "narrador".

A posição central do espaço semântico é ocupada pela palavra "convento" – o que é muito apropriado.

Para obter uma percentagem de variância de 85% da informação total são necessárias d= 50 (das 88) dimensões para o espaço de representação do ES.

É este ES o usado nas experiências descritas no que se segue, em **3.9.3.**, **3.9.4.** e **3.9.6.**

3.9.3. Dados EX-MIN. Experiência nº 1.

Nesta primeira experiência usa-se o ES descrito em **3.9.2.** (88 textos extraídos de manuais de ensino do Memorial do Convento de José Saramago) e 60 textos de resposta dos estudantes á questão do Grupo I A), conforme descrito em **3.7.**

A proporção destes 60 textos a usar para treinar o classificador (AT) e para testar o mesmo (T) é um dos dois parâmetros a fixar pelo investigador.

O outro parâmetro é o número k dos vizinhos mais próximos, sobre o ES, de um texto que se pretenda classificar automaticamente.

O valor deste parâmetro é um geral pequeno. Nas experiências realizadas neste trabalho, k varia entre 1 e 8.

O sistema PAET usa as classificações (ClProf) atribuídas pelo Professor (neste caso, professores contratados pelo Ministério) aos textos da amostra de treino (AT) para calcular a nota (ClSys) a atribuir pelo sistema.

As classificações automáticas atribuídas pelo sistema a esta amostra de teste (T) são, em seguida, correlacionadas com as classificações atribuídas pelos professores contratados, validando-se assim o classificador.

Nas experiências aqui realizadas e na versão atual do programa PAET,

$$ClSys = \frac{1}{k}\sum_{i=1}^{k} Cl\,\mathrm{Pr}\,of_i$$

isto é, o sistema atribui ao texto a classificar uma classificação que é a média das classificações dos testes da amostra de treino (AT) que correspondam aos maiores cossenos (maiores proximidades) entre o texto a classificar e os textos da amostra de treino (AT).

Na presente experiência, fixou-se a proporção de textos a usar na AT em 0.75, o que significa $0.75 \times 60 = 45$ textos na AT e $0.15 \times 60 = 15$ textos na amostra T.

Trata-se obviamente de números insuficientes de acordo com a literatura consultada mas que, mesmo assim, são úteis para validar aspetos importantes dos sistema PAET.

Projetando sobre o ES os textos da AT e da amostra T (Bellegarda, 2007) o resultado é o posicionamento destes 60 textos adicionais (45 + 15) sobre um ES que, pelo que foi visto no número anterior, usa d= 50 dimensões para garantir uma percentagem acumulada de informação (variância) de cerca de 85% da matriz X de frequências.

Esta projeção não altera o ES que serve agora de referencial a esta e a outras experiências. A inspeção visual do biplot representado na **figura 3.9.3.1.** permite formar uma ideia da situação, usando apenas d= 2 (27% da informação) das 50 dimensões necessárias para garantir 85% da informação.

Na **figura 3.9.3.1.**, obtida pelo PAET, as setas a negro correspondem aos textos da AT e as setas pintadas de verde correspondem à amostra de teste T.

Usando a funcionalidade do programa PAET – ver Manual no Anexo A – pode examinar-se interactivamente a posição relativa desses textos e "tomar notas" úteis à interpretação do que está em causa – o que não pode ser apreciado numa imagem estática como a anterior.

Note-se, contudo, que este biplot representa apenas 22% do total da informação total de X. Ao examinar o biplot deve sempre estar presente a perceção de que as proximidades entre textos (ângulos pequenos) observadas no biplot podem não corresponder à realidade da dimensão d= 50, por exemplo, em que estão representados esses eixos (Vairinhos, 2003).

No caso da experiência aqui descrita, para cada um dos 15 textos da amostra T foram calculados os cossenos dos ângulos de cada um desses textos com os 45 da AT.

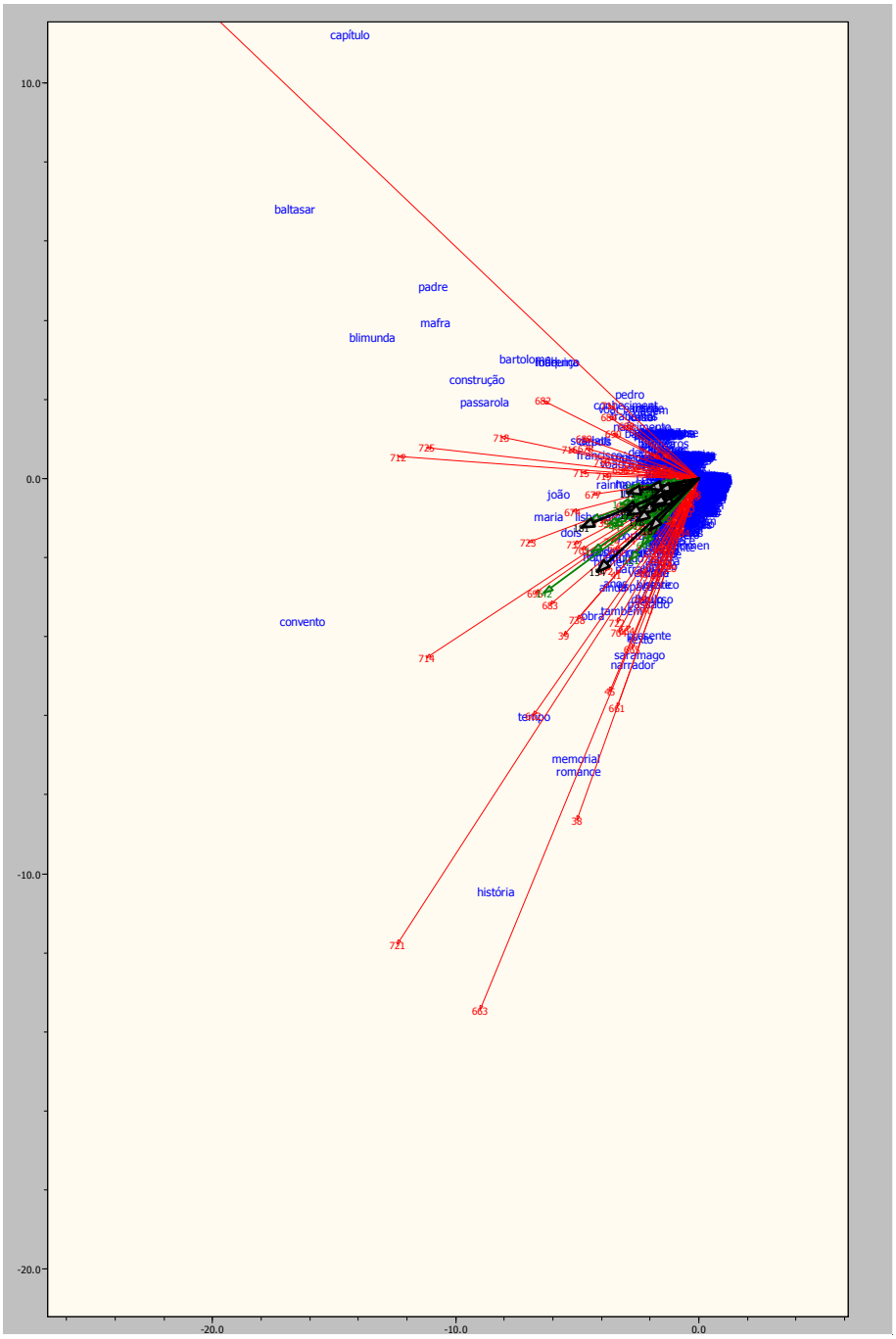
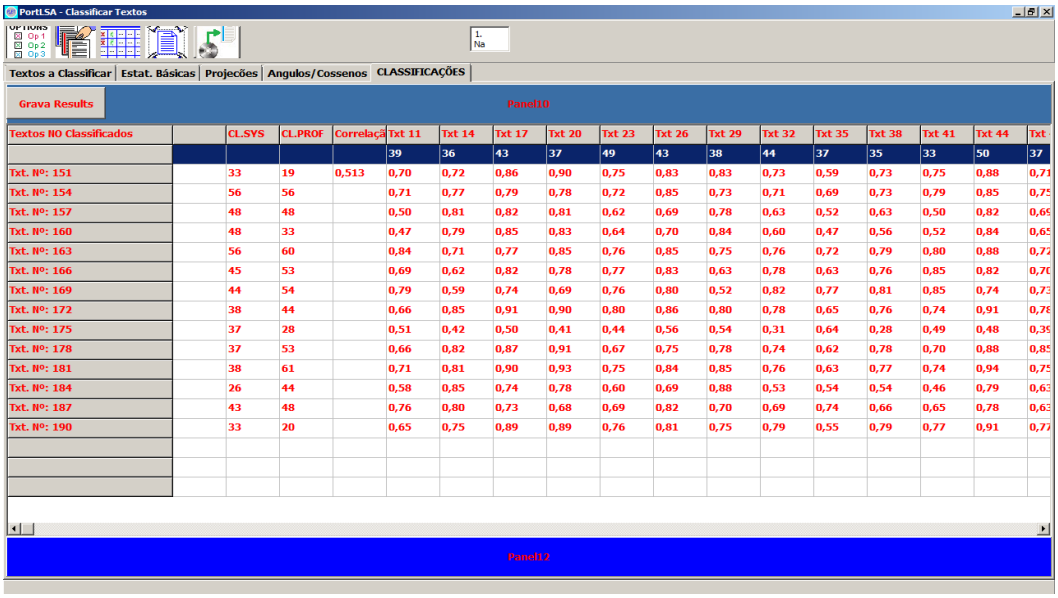


Figura 3.9.3.1. Biplot que permite visualizar nas duas primeiras dimensões, as projeções sobre o ES dos textos das amostras AT e T.

Em seguida, considerando apenas os *k*= 1 textos da AT mais próximos de cada um desses 15 textos a classificar (maiores cossenos), foi calculada a média das classificações atribuídas pelos professores a esses *k* testes da AT. O resultado é a classificação ClSys atribuída ao teste específico da amostra T.

A **figura 3.9.3.2.** apresenta o resultado desses cálculos, realizados pelo programa PAET.



Textos NO Classificados		CL.SYS	CL.PROF	Correlaçã	Txt 11	Txt 14	Txt 17	Txt 20	Txt 23	Txt 26	Txt 29	Txt 32	Txt 35	Txt 38	Txt 41	Txt 44	Txt
					39	36	43	37	49	43	38	44	37	35	33	50	37
Txt. Nº: 151		33	19	0,513	0,70	0,72	0,86	0,90	0,75	0,83	0,83	0,73	0,59	0,73	0,75	0,88	0,71
Txt. Nº: 154		56	56		0,71	0,77	0,79	0,78	0,72	0,85	0,73	0,71	0,69	0,73	0,79	0,85	0,75
Txt. Nº: 157		48	48		0,50	0,81	0,82	0,81	0,62	0,69	0,78	0,63	0,52	0,63	0,50	0,82	0,69
Txt. Nº: 160		48	33		0,47	0,79	0,85	0,83	0,64	0,70	0,84	0,60	0,47	0,56	0,52	0,84	0,65
Txt. Nº: 163		56	60		0,84	0,71	0,77	0,85	0,76	0,85	0,75	0,76	0,72	0,79	0,80	0,88	0,72
Txt. Nº: 166		45	53		0,69	0,62	0,82	0,78	0,77	0,83	0,63	0,78	0,63	0,76	0,85	0,82	0,76
Txt. Nº: 169		44	54		0,79	0,59	0,74	0,69	0,76	0,80	0,52	0,82	0,77	0,81	0,85	0,74	0,73
Txt. Nº: 172		38	44		0,66	0,85	0,91	0,90	0,80	0,86	0,80	0,78	0,65	0,76	0,74	0,91	0,78
Txt. Nº: 175		37	28		0,51	0,42	0,50	0,41	0,44	0,56	0,54	0,31	0,64	0,28	0,49	0,48	0,35
Txt. Nº: 178		37	53		0,66	0,82	0,87	0,91	0,67	0,75	0,78	0,74	0,62	0,78	0,70	0,88	0,85
Txt. Nº: 181		38	61		0,71	0,81	0,90	0,93	0,75	0,84	0,85	0,76	0,63	0,77	0,74	0,94	0,75
Txt. Nº: 184		26	44		0,58	0,85	0,74	0,78	0,60	0,69	0,88	0,53	0,54	0,54	0,46	0,79	0,63
Txt. Nº: 187		43	48		0,76	0,80	0,73	0,68	0,69	0,82	0,70	0,69	0,74	0,66	0,65	0,78	0,63
Txt. Nº: 190		33	20		0,65	0,75	0,89	0,89	0,76	0,81	0,75	0,79	0,55	0,79	0,77	0,91	0,77

Figura 3.9.3.2. Nas duas primeiras colunas, comparação das notas atribuídas pelo sistema PAET (à esquerda) e as notas atribuídas pelos professores.

Na **figura 3.9.3.2.** a linha do topo contém os identificadores de alguns dos 45 textos na amostra de treino e a primeira coluna do lado esquerdo os identificadores dos 15 textos da amostra T.

No corpo da tabela figuram os cossenos dos ângulos entre os testes da amostra T e os da AT, num espaço de d= 50 dimensões do ES.

Na coluna 5 da tabela da figura anterior é apresentada a correlação – 0.513 – entre as classificações atribuídas pelo sistema (ClSys) e as atribuídas pelos professores.

Exportando estes dados para o exterior e estudando-os com o EXCEL e o SPSS obtêm-se os resultados a seguir descritos na **tabela 3.9.3.1.**

A **tabela 3.9.3.1.** contém os resultados obtidos pelo sistema (ClSys) e os resultados atribuídos pelos professores (ClProf) do Ministério.

Recorde-se que os resultados da **tabela 3.9.3.1.** foram obtidos com uma proporção para a amostra de treino de 75% e com k= 1 (número de vizinhos mais próximos no ES, a usar para calcular ClSys) – o que equivale, neste caso, a escolher o resultado da amostra de treino correspondente ao cosseno máximo.

O coeficiente de correlação (ClSys, ClProf) para estes dados foi 0.513 (significância = 0.06), que fica próximo do valor necessário para ser considerado significativo ao nível 0.05.

Textos da Amostra T	ClSys	ClProf
151	33	19
154	56	56
157	48	48
160	48	33
163	56	60
166	45	53
169	44	54
172	38	44
175	37	28
178	37	53
181	38	61
184	26	44
187	43	48
190	33	20

Tabela 3.9.3.1. Resultados obtidos pelo classificador automático da PAET (ClSys) e atribuídos pelos professores (ClProf).

O gráfico da **figura 3.9.3.3.** (obtido com o EXCEL) é o gráfico de dispersão dos dados da **tabela 3.9.3.1.** tendo sobreposta a recta de regressão a que corresponde R^2= 0.263.

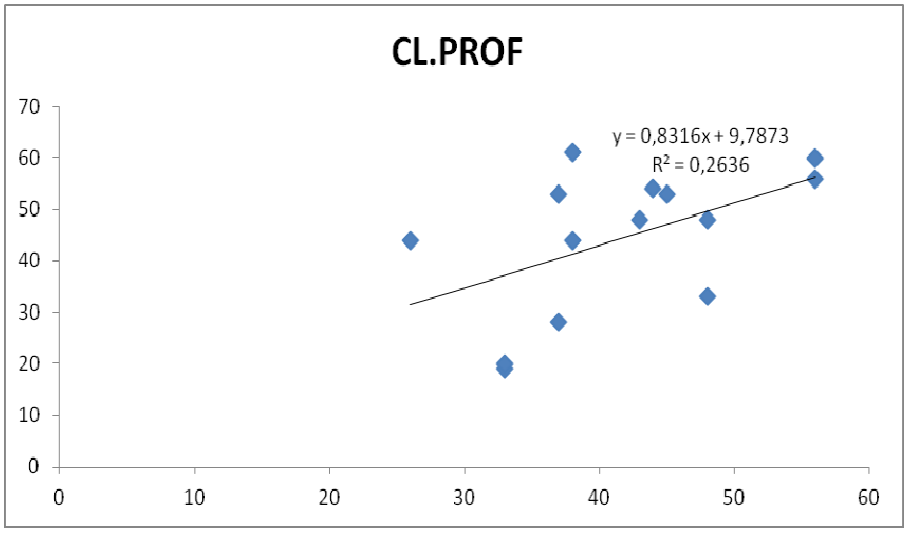


Figura 3.9.3.3. Regressão das classificações do professor em função das classificações do sistema.

Como se constata pela **figura 3.9.3.3.**, a resposta à questão "*são as classificações "*reais*" (dos professores) preditíveis em função das do sistema?*" é a de que sim, mas com fraca qualidade, no caso destes dados.

Face à exiguidade das amostras usadas (88 no ES, 45 na AT e 15na T) consideram-se estes valores, apesar de tudo, animadores, se bem que longe dos valores publicados na literatura e obtidos com sistemas industriais, produto de anos de aperfeiçoamento.

A razão pela qual se escolheu os parâmetros (75%, *d*= 1) aparece justificada na **tabela 3.9.3.2.** e também na **figura 3.9.3.4.** que contêm valores alternativos para a correlação *r*(ClSys, ClProf) obtidos quando se variam os valores dos parâmetros (Percentagem da AT, *k*) com Percentagem da AT= 50%, 60%, 75% e *k*= 1, 2, …, 5.

	Percentagem da Amostra de Treino (AT)		
k	**50%**	**60%**	**75%**
1	0.14	0.23	0.51
2	0.08	0.21	0.28
3	0.09	0.28	0.34
4	0.23	0.20	0.20
5	0.20	0.20	0.17

Tabela 3.9.3.2. Coeficientes de Correlação (ClSys, ClProf) quando se varia *k* e a % de textos na amostra de treino.

Representando estes dados graficamente obtém-se a **figura 3.9.3.4.** que permite constatar que o valor máximo desse coeficiente de correlação obtém-se quando (*Percent*= 75%, *k*= 1) – sendo esse valor 0.51.

Deve notar-se que o valor 75% para a amostra de treino é muito elevado e só tem justificação num contexto como o aqui descrito em que há tão poucas observações disponíveis que, para aumentar a sensibilidade do classificador é imperativo aumentar a informação necessária ao respetivo treino mesmo à custa da amostra de teste o que, por sua vez, prejudica o coeficiente de correlação *r*(ClSys, ClProf). Num estudo sem essas limitações esta proporção é substancialmente mais baixa, garantindo, mesmo assim, um número (centenas) de textos suficientes para realizar o treino do classificador e os estudos de validade.

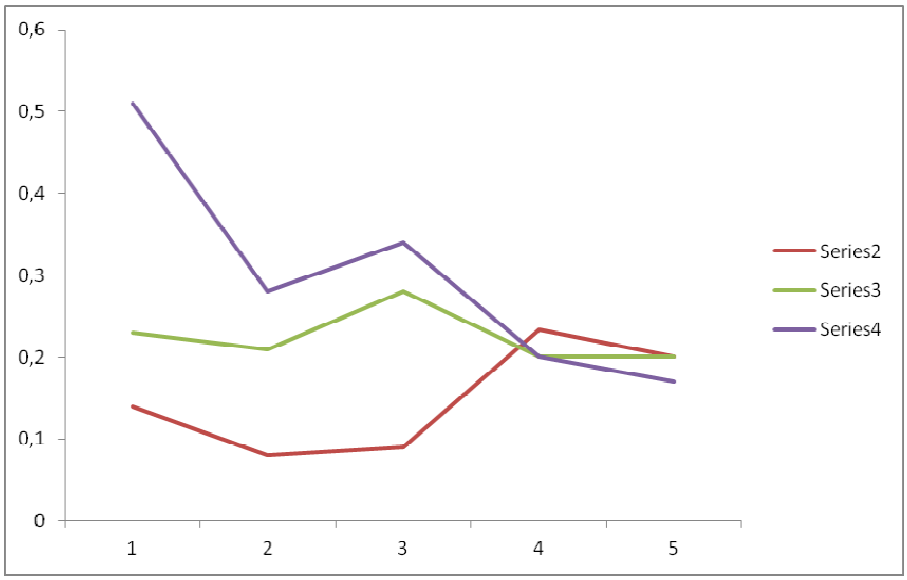


Figura 3.9.3.4. Gráfico de coeficiente de correlação (ClSys, ClProf) em função de *k* para vários valores da percentagem da AT.

3.9.4. Dados EX-MIN. Experiência nº 2.

Neste número apresentam-se resultados produzidos pelo programa PAET relativos à análise das respostas Grupo I A) e B) simultaneamente.

Embora a literatura não recomende a classificação simultânea de mais do que uma questão de cada vez – no decurso da análise serão evidenciados alguns dos problemas resultantes – por outro lado, isso permite, envolvendo muito mais textos de resposta, pôr em evidência e confirmar através de dados reais as propriedades da metodologia ASL e a validade da respetiva implementação no programa PAET.

Antes de realizar a experiência de classificação, usando o ES descrito, relativo ao Memorial do Convento, consideremos a **figura 3.9.4.1.** que apresenta a imagem em duas dimensões (biplot) do resultado da decomposição em valores e vetores singulares dos textos relativos às respostas à A) e também à B) do Grupo I dos dados EX-MIN. Ambas as questões se referiam ao Memorial do Convento de José Saramago mas a A) tinha a ver com a viagem da princesa de Montemor a Évora e a B) tinha a ver com o romance amoroso Baltazar/Blimunda tendo como pano de fundo a construção do Convento de Mafra.

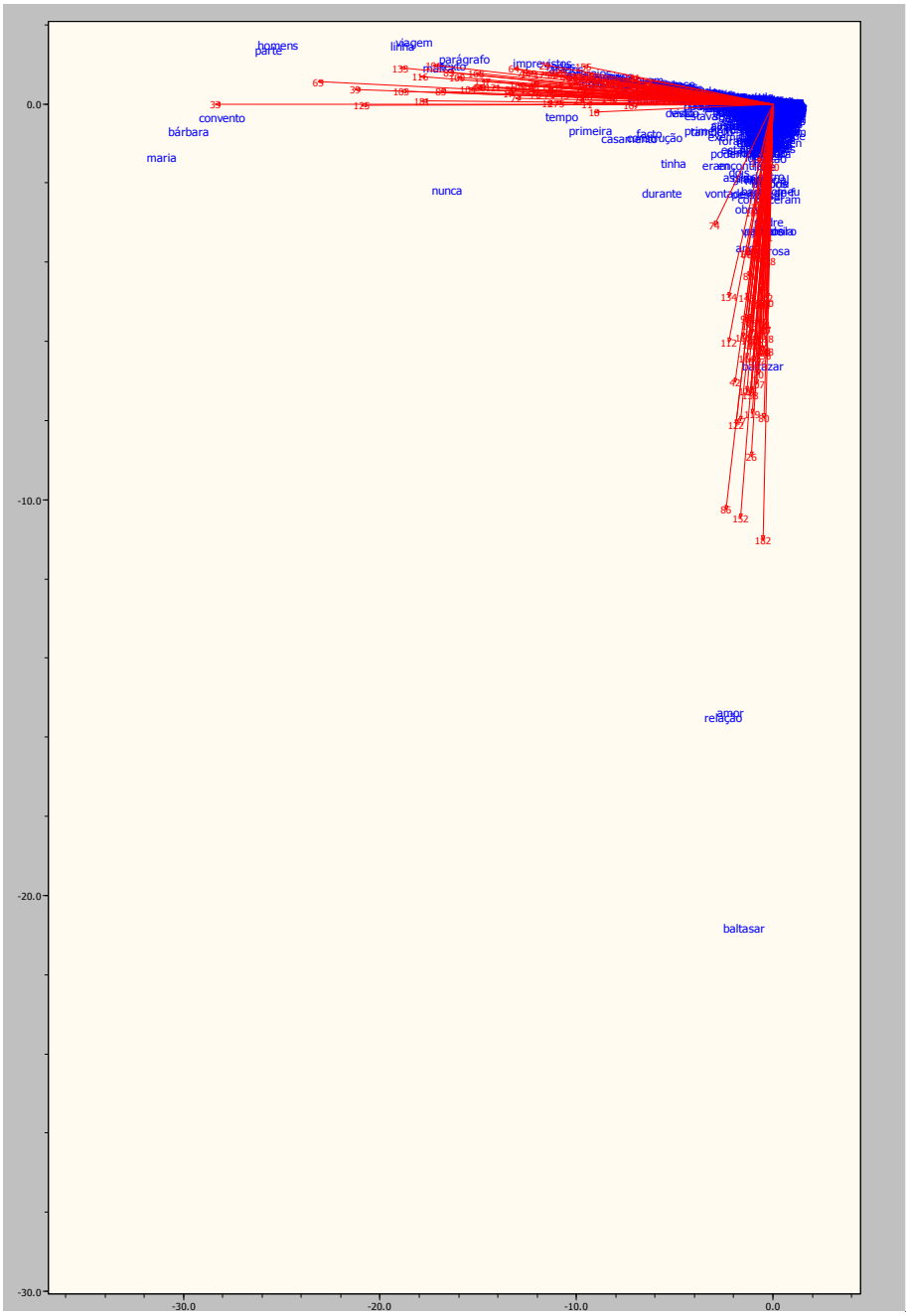


Figura 3.9.4.1. Biplot correspondente aos textos das respostas às questões A) e B), relativas ao Memorial do Convento.

A **figura 3.9.4.1.** foi construída usando a informação de 60 respostas à A) e 60 respostas à B), com as quais foi construída uma matriz de frequências X (n, p) com n= 2497 palavras e p= 120 textos. Procedendo à decomposição em valores e vetores singulares dessa matriz, verifica-se que é necessária uma dimensionalidade d= 60 para captar 90% de toda a informação dos dados.

A **figura 3.9.4.1.** foi construída com as duas primeiras dimensões, correspondendo a uma variância/informação de 41%, o que significa que este biplot é altamente representativo da informação contida nos 120 textos.

Considera-se a **figura 3.9.4.1.** notável e traduzindo muito bem o poder deste tipo de análise.

Com efeito, a figura envolve dois grupos de textos (setas a vermelho) que, pela sua quase ortogonalidade, definem a topologia do gráfico. Os textos contidos em cada um desses grupos formam ângulos muito pequenos, significando que são textos cujo significado é semelhante.

Se examinarmos agora os textos que fazem parte de cada um dos grupos (ver as elipses sobrepostas aos grupos de textos), verifica-se que o grupo superior (horizontal) agrupa os textos de resposta à A) – relativos à viagem Montemor/Évora – e o grupo vertical, à direita, é relativo aos textos da resposta à B).

As palavras mais representativas (escritas a azul no gráfico) abrangem {*bárbara, convento, homens, linha, Mafra, Maria, nunca, parágrafo, parte, princesa, texto, viagem*}.

Os textos mais associados ao grupo vertical são todos textos de respostas á B), tais como {26, 86, 152, …} e palavras como {*amor, Baltazar, relação, amorosa, juntos, passarola, verdadeiro*} fortemente evocativos do romance Baltazar/Blimunda, como devem.

Em síntese, a **figura 3.9.4.1.** capta, apesar de ser construída apenas com dois eixos (dos 60 necessários para obter 90% da informação), quase na perfeição, o conhecimento prévio de que os textos se dividem em dois grupos correspondentes a temas semelhantes mas distintos, separando esses textos em dois significados bem definidos em perfeita concordância com o significado dos mesmos.

Acresce ainda que em cada grupo coexistem textos de qualidade muito diversa (tanto no conteúdo como na natureza e número das palavras usadas). Essa variabilidade é captada pela diversidade dos pequenos ângulos entre textos do mesmo grupo.

Mesmo não sabendo, pela análise representada anteriormente, que os dois grupos têm significados diferentes (A e B), o mero exame desta figura informa-nos que subjacente (latente) a estes 120 textos se encontram dois significados que explicam as proximidades entre os textos do mesmo grupo e a presença dos dois grupos que formam um ângulo praticamente de 90%, significando ausência de correlação entre esses grupos.

Se repetíssemos esta análise mas escolhendo ao acaso um certo número de textos abrangendo agora respostas a A), B) e Grupo III, obter-se-iam biplots como o ilustrado na **figura 3.9.4.2.**

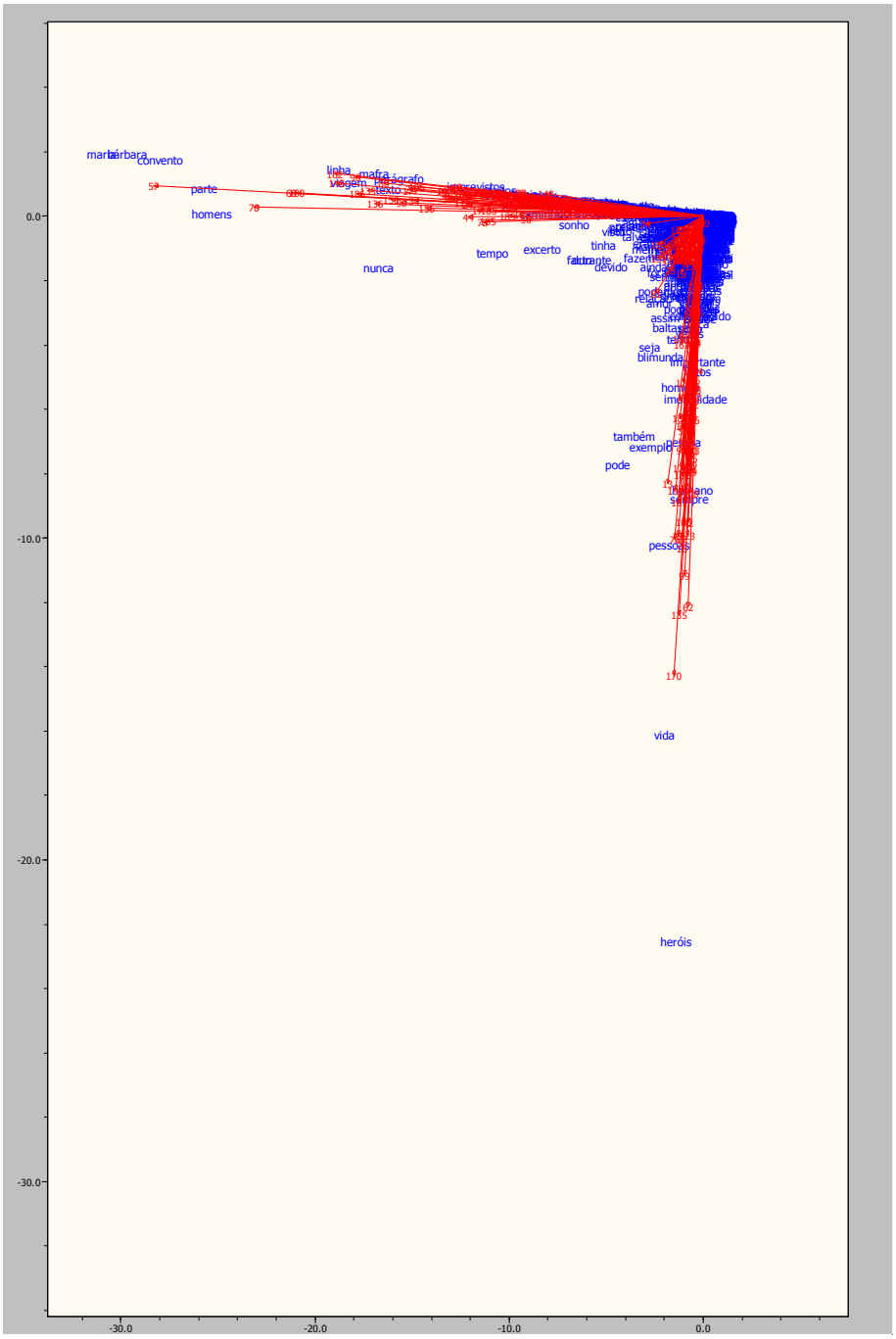


Figura 3.9.4.2. Biplot construído com os dois primeiros eixos da decomposição de uma matriz abrangendo as respostas A), B) e GIII.

Esta figura corresponde a uma matriz X (n, p) com n= 3993 palavras, p= 181 formada a partir de 181 textos de resposta correspondentes às A) e B) do Grupo I e Grupo III.

A variância correspondente a este biplot (dois primeiros eixos) é 37% do total. Pode constatar-se agora que os três grupos de textos correspondem, respetivamente, aos Grupo III, B) e A).

Isso pode ser visto na **figura 3.9.4.3.** em que é apresentado um biplot de d= 3 dimensões (representando 43% da variância total) e em que se vê, claramente, os três grupos de textos, cada um correspondente aos textos de resposta a uma das questões GI A), B) e GII.

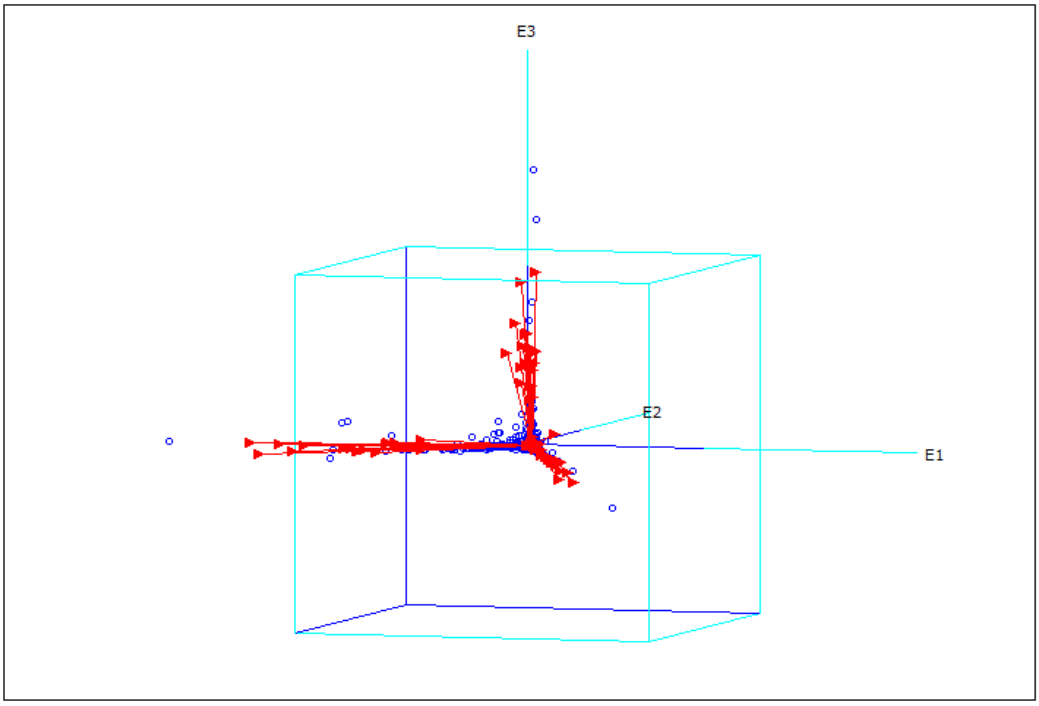


Figura 3.9.4.3. Biplot correspondente aos três primeiros eixos da análise envolvente GI A), B) e GII.

Esta análise confirmou a reconhecida capacidade das análises LSA/Biplot para descobrirem a estrutura dos significados subjacentes a um certo conjunto de textos considerandos apenas as frequências de ocorrência das palavras nos textos como a validade da respetiva implementação no programa PAET desenvolvido para operar as ideias desta tese.

Voltando agora ao ES, construído com os textos usados no ensino do Memorial do Convento de José Saramago, vai-se usar como AT (Amostra de Testes) as 120 respostas A) e B) às duas questões do exame de Português (12º ano), de 2008, relativas a esse tema.

A **figura 3.9.4.4.** apresenta agora não só o ES como as projeções sobre este ES das 120 respostas, metade das quais relativas à AT e as restantes relativas à amostra de treino (T).

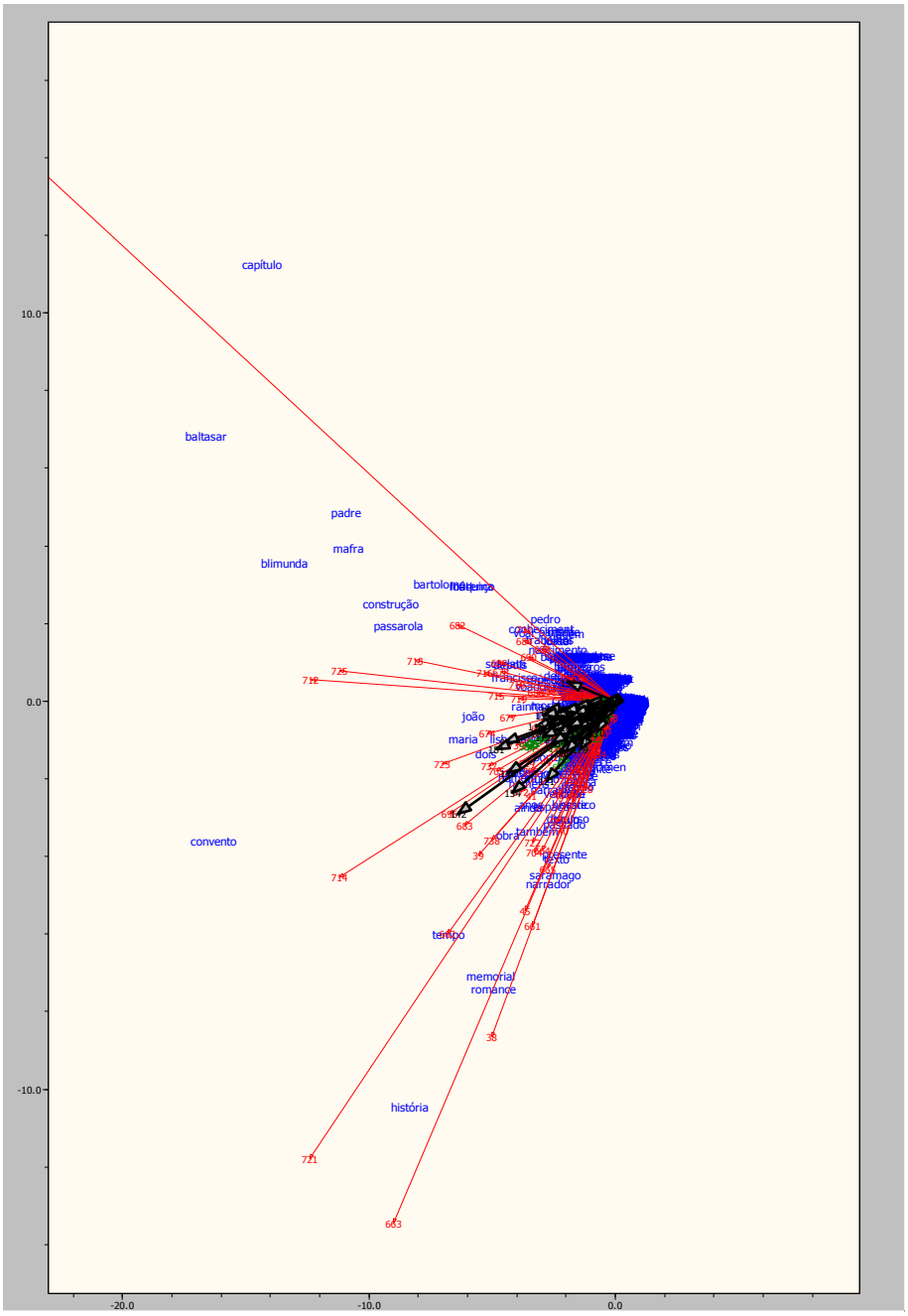


Figura 3.9.4.4. Espaço semântico relativo ao Memorial do Convento de José Saramago e as projeções dos textos das 120 respostas à A) e à B).

Recorde-se que a matriz de frequências analisada – $X\,(n, p)$ – tem n= 5012 linhas (palavras) e p= 88. Ver **3.9.1.**

Na **figura 3.9.4.5.** vê-se a mesma informação mas relativa apenas a uma zona limitada do gráfico anterior, obtida fazendo uso do programa PAET. Nela se podem ver, com maior nitidez, as posições relativas da AT (vetores a negro) e da amostra de teste T (a verde) – para lá das palavras e textos do ES (a azul e vermelho)..

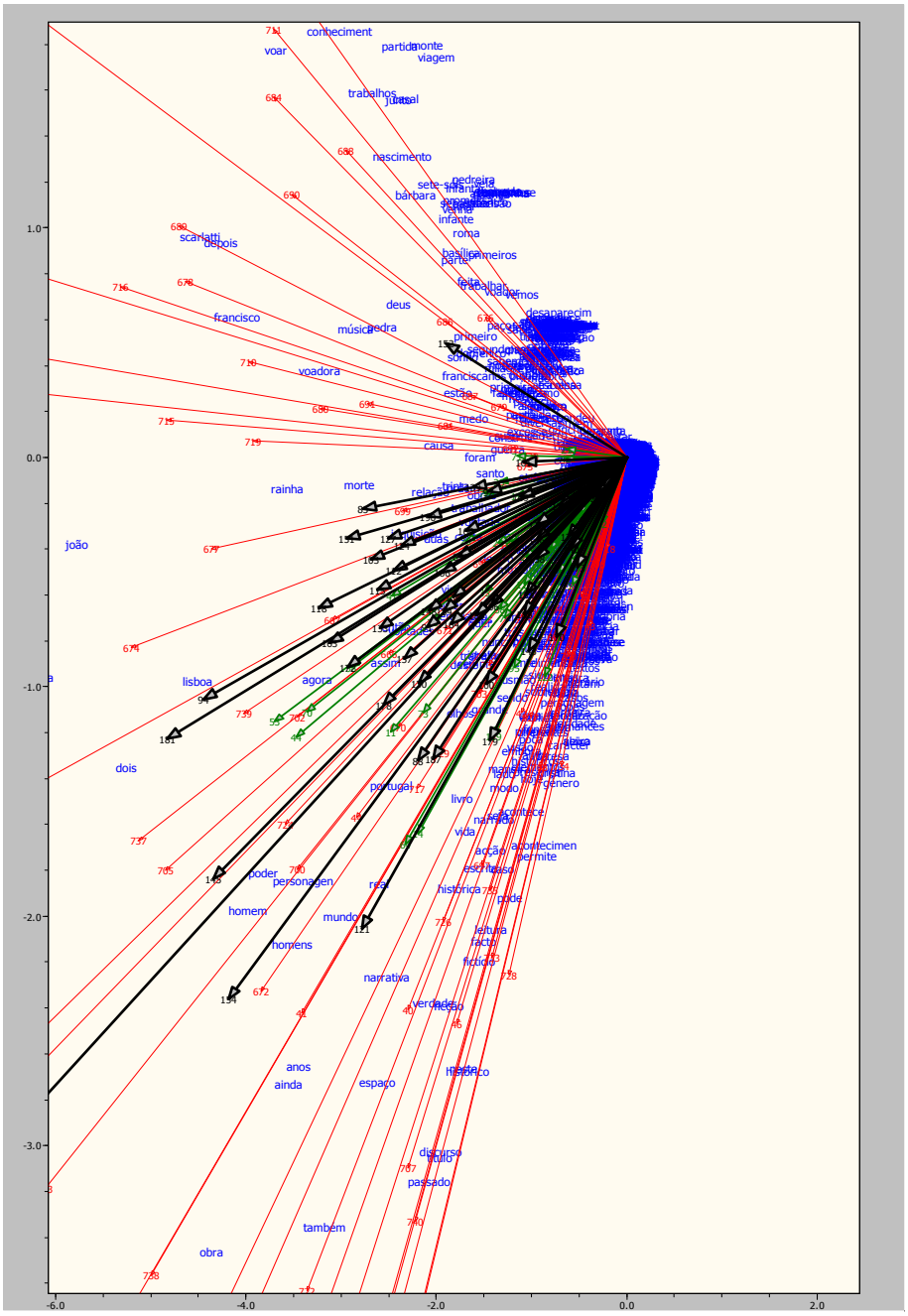
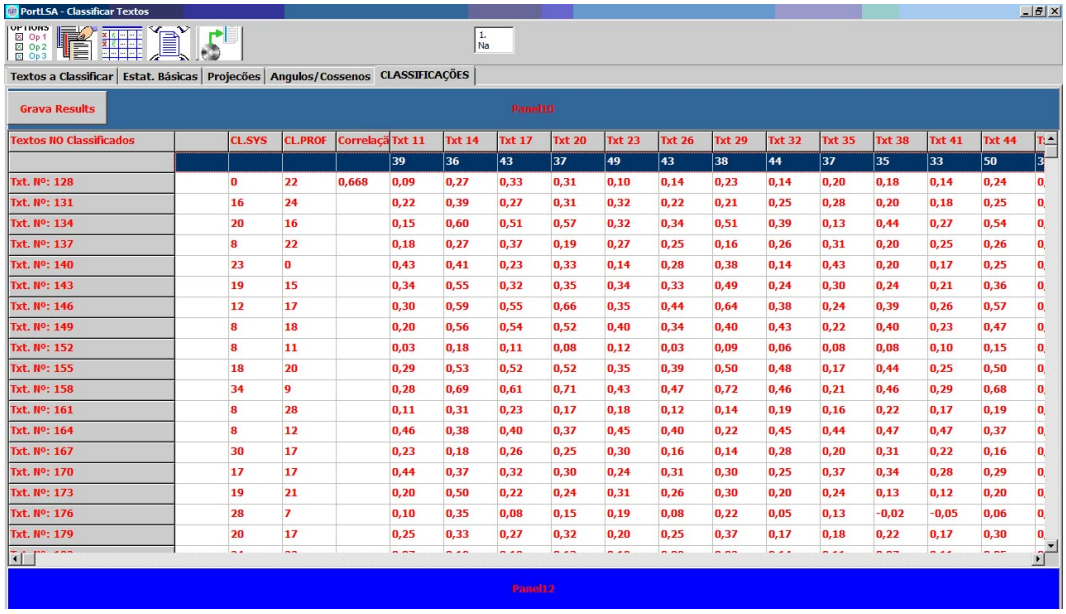


Figura 3.9.4.5. Amplificação de uma zona da **figura 3.9.4.4.** permitindo ver com maior nitidez as amostras de treino (vetores a negro) e amostra de teste (vetores a verde).

A **figura 3.9.4.6.** apresenta os resultados ao classificar simultaneamente textos das resposta A) e B) e a que corresponde um coeficiente de correlação 0.668 (para uma proporção da AT de 50%) e *d*= 1 no cálculo dessas classificações.


PortLSA - Classificar Textos

Textos a Classificar | Estat. Básicas | Projecções | Angulos/Cossenos | CLASSIFICAÇÕES

Grava Results

Textos NO Classificados		CL.SYS	CL.PROF	Correlaçã	Txt 11	Txt 14	Txt 17	Txt 20	Txt 23	Txt 26	Txt 29	Txt 32	Txt 35	Txt 38	Txt 41	Txt 44	T
					39	36	43	37	49	43	38	44	37	35	33	50	3
Txt. Nº: 128		0	22	0,668	0,09	0,27	0,33	0,31	0,10	0,14	0,23	0,14	0,20	0,18	0,14	0,24	0
Txt. Nº: 131		16	24		0,22	0,39	0,27	0,31	0,32	0,22	0,21	0,25	0,28	0,20	0,18	0,25	0
Txt. Nº: 134		20	16		0,15	0,60	0,51	0,57	0,32	0,34	0,51	0,39	0,13	0,44	0,27	0,54	0
Txt. Nº: 137		8	22		0,18	0,27	0,37	0,19	0,27	0,25	0,16	0,26	0,31	0,20	0,25	0,26	0
Txt. Nº: 140		23	0		0,43	0,41	0,23	0,33	0,14	0,28	0,38	0,14	0,43	0,20	0,17	0,25	0
Txt. Nº: 143		19	15		0,34	0,55	0,32	0,35	0,34	0,33	0,49	0,24	0,30	0,24	0,21	0,36	0
Txt. Nº: 146		12	17		0,30	0,59	0,55	0,66	0,35	0,44	0,64	0,38	0,24	0,39	0,26	0,57	0
Txt. Nº: 149		8	18		0,20	0,56	0,54	0,52	0,40	0,34	0,40	0,43	0,22	0,40	0,23	0,47	0
Txt. Nº: 152		8	11		0,03	0,18	0,11	0,08	0,12	0,03	0,09	0,06	0,08	0,08	0,10	0,15	0
Txt. Nº: 155		18	20		0,29	0,53	0,52	0,52	0,35	0,39	0,50	0,48	0,17	0,44	0,25	0,50	0
Txt. Nº: 158		34	9		0,28	0,69	0,61	0,71	0,43	0,47	0,72	0,46	0,21	0,46	0,29	0,68	0
Txt. Nº: 161		8	28		0,11	0,31	0,23	0,17	0,18	0,12	0,14	0,19	0,16	0,22	0,17	0,19	0
Txt. Nº: 164		8	12		0,46	0,38	0,40	0,37	0,45	0,40	0,22	0,45	0,44	0,47	0,47	0,37	0
Txt. Nº: 167		30	17		0,23	0,18	0,26	0,25	0,30	0,16	0,14	0,28	0,20	0,31	0,22	0,16	0
Txt. Nº: 170		17	17		0,44	0,37	0,32	0,30	0,24	0,31	0,30	0,25	0,37	0,34	0,28	0,29	0
Txt. Nº: 173		19	21		0,20	0,50	0,22	0,24	0,31	0,26	0,30	0,20	0,24	0,13	0,12	0,20	0
Txt. Nº: 176		28	7		0,10	0,35	0,08	0,15	0,19	0,08	0,22	0,05	0,13	-0,02	-0,05	0,06	0
Txt. Nº: 179		20	17		0,25	0,33	0,27	0,32	0,20	0,25	0,37	0,17	0,18	0,22	0,17	0,30	0

Figura 3.9.4.6. Resultados da classificação (usando o espaço semântico relativo ao Memorial do Convento) de 60 textos de resposta (relativos a A) e a B)).

Analisando os dados desta tabela com a folha de cálculo EXCEL, obtém-se a **figura 3.9.4.7.**, em que se "verifica" que uma reta de regressão permitiria predizer com relativa segurança (qualidade R^2= 0.445) correspondente ao coeficiente de correlação 0.67, os valores de ClProf (Ministério) em função dos valores atribuídos pelo sistema (ClSys).

A **figura 3.9.4.7.** – obtida com o EXCEL – mostra que esta reta de regressão é significativa e que os coeficientes respetivos são significativos ao nível 0.05.

Estes resultados devem ser considerados com prudência visto que, ao examinar novamente essa figura, nota-se a presença de dois grupos imputáveis à diferença de escalas de classificação das respostas A) e B) pelo que a correlação elevada pode estar associada a este efeito e não à efetiva preditibilidade das classificações dos professores em função das classificações dentro dos grupos.

Trata-se de um dos efeitos que justifica a recomendação encontrada na literatura no sentido de evitar classificar mais do que um item de cada vez.

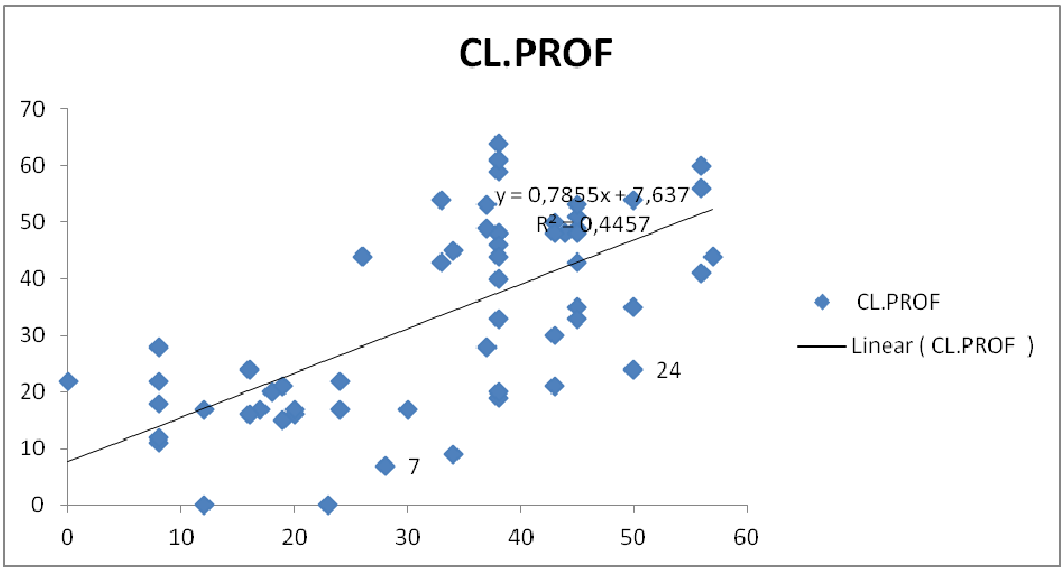


Figura 3.9.4.7. Significado da reta de regressão da **figura 3.9.4.6.**

3.9.5. Dados de GESTÃO. Experiência nº 1.

Neste número apresentam-se os resultados de uma experiência realizada usando os dados GESTÃO relativos a uma avaliação formativa de estudantes da disciplina de gestão de um Instituto Politécnico.

Esta experiência destina-se a investigar a possibilidade de usar as respostas dos próprios estudantes em exames anteriores, as respostas dos professores ou uma mistura dos dois para classificar, depois, novos testes.

Este esquema tem importância prática uma vez que se este tipo de instrumento se destina a apoiar o trabalho do professor individual, não se deve assumir mais informação do que aquela a que o professor tem acesso - manuais de ensino, a cultura do professor e os textos produzidos por estudantes com que tenha ensinado.

Usando toda a informação relativa a estes dados resultantes dos textos de resposta a uma única questão relativa ao conceito de Qualidade Total, obtém-se uma matriz de frequências $X(n, p)$ em que n= 1937 linhas (palavras) e p= 187 textos.

Realizando a decomposição em valores e vetores singulares verifica-se que para obter uma variância de 90% basta chegar à dimensão d= 78. Usando apenas a informação correspondente às duas primeiras dimensões para construir um biplot, a informação correspondente é 48%. Isto significa que essas duas primeiras dimensões absorvem uma percentagem muito considerável da informação total.

Na **figura 3.9.5.1.** apresenta-se esse biplot.

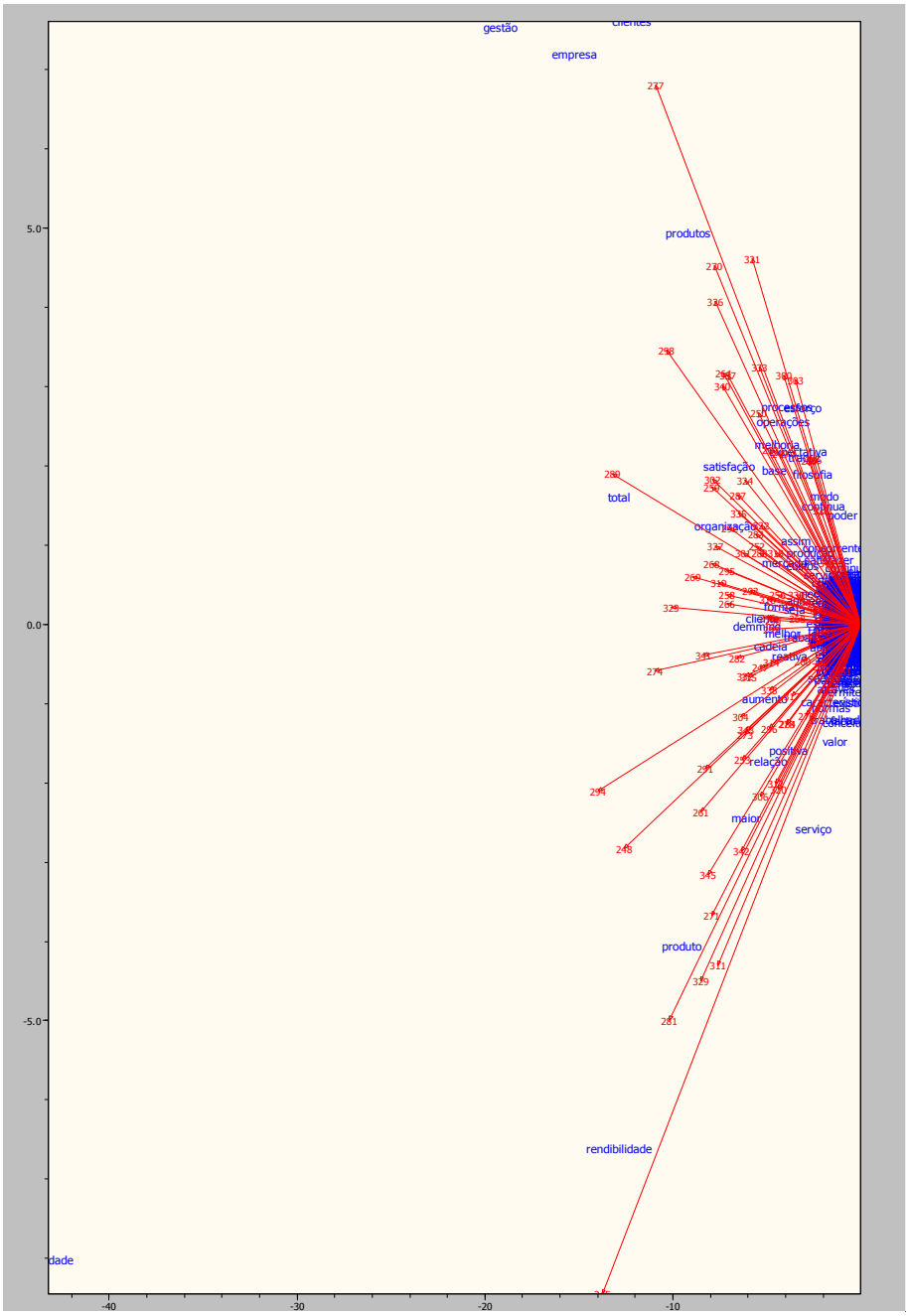


Figura 3.9.5.1. Biplot correspondente às duas primeiras dimensões do espaço semântico construído com a totalidade das respostas de 187 estudantes a uma questão relativa à qualidade total.

Examinando a figura anterior parece que a estrutura global das respostas se pode descrever em função de três grupos associados a questões conceptuais (textos do grupo que aponta para o canto superior esquerdo), textos associados a questões de eficiência (grupo do centro) e testes ligados a questões operacionais e resultados (grupo que aponta para o canto inferior esquerdo). Isto é, o biplot anterior transmite a estrutura subjacente à totalidade da informação associada às 187 respostas.

Usando agora apenas 85 destes textos, escolhidos ao acaso, para construir um espaço semântico que retrate a linguagem dos estudantes – e nos quais poderiam também ser incluídas algumas respostas elaboradas pelo professor – obtém-se uma matriz de frequências $X(n, p)$ com n= 1303 palavras e p= 85 textos.

Procedendo à decomposição em valores e vetores singulares, verifica-se que a dimensão d necessária a obter 90% da informação é d= 41 e que um biplot correspondente às primeiras duas dimensões absorve 46% dessa informação.

Na **figura 3.9.5.2.** apresenta-se esse biplot, constatando-se que tem uma estrutura muito semelhante à do biplot construído com a totalidade (187) dos textos de resposta.

Examinando o biplot do ES, constata-se que não se notam distorções estruturais ou enviesamentos associados a textos específicos, podendo assumir-se que este biplot é representativo da estrutura do significado da totalidade da informação.

Os restantes 187 – 85= 102 textos de resposta são em seguida usados para construir (em função do valor do parâmetro proporção) as amostras de treino (AT) e de teste (T) a validar.

Na **figura 3.9.5.3.** apresenta-se agora o biplot correspondente ao ES retido e sobre o qual foram projetadas as amostras AT e T.

Examinando o biplot da **figura 3.9.5.3.** – usando as facilidades incluídas no programa PAET – verifica-se que as respostas estão na maior parte dos casos relacionadas com a parte do espaço semântico (ES) ligada às questões operacionais e dos resultados da empresa.

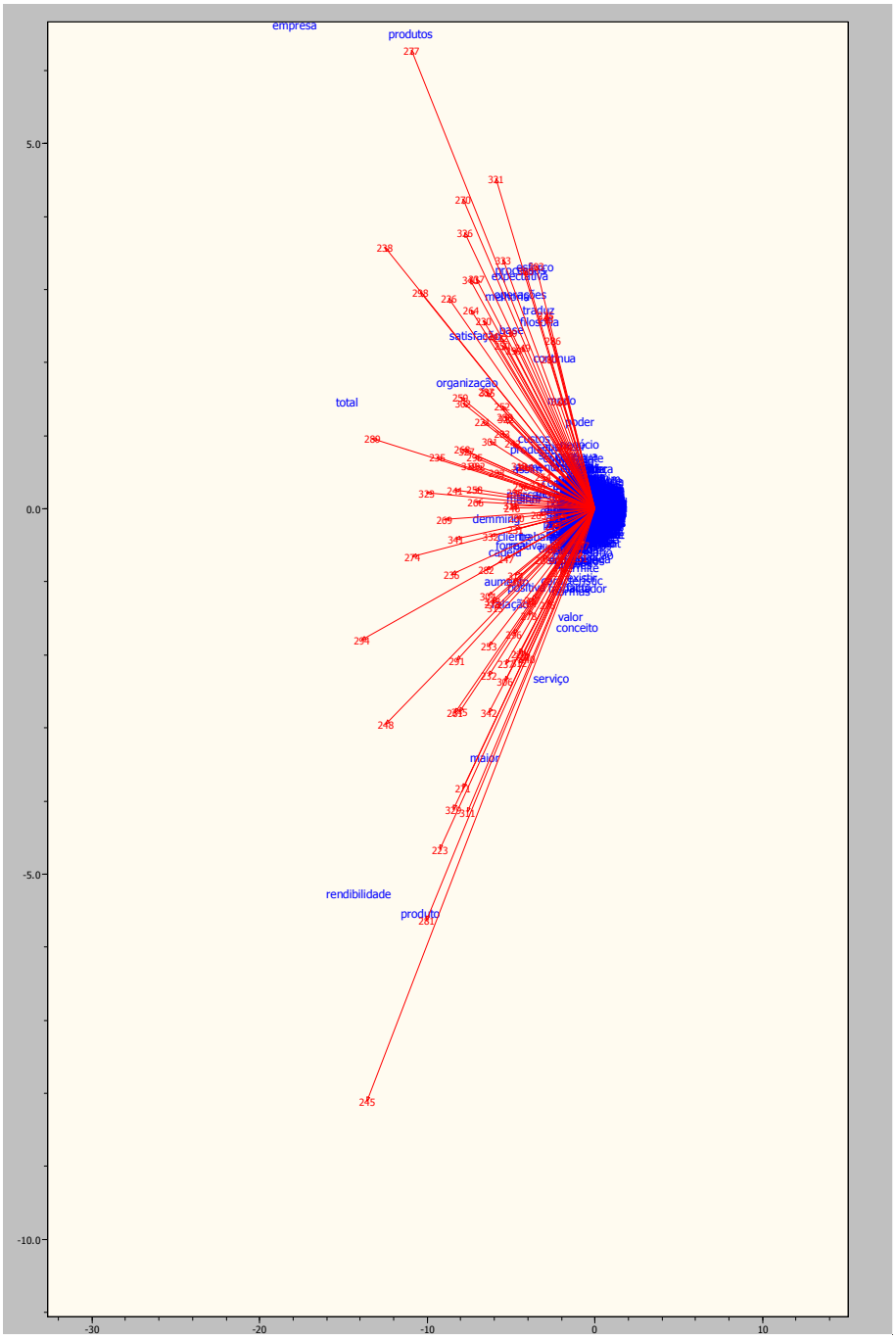


Figura 3.9.5.2. Biplot do Espaço Semântico correspondente a 85 dos textos dos dados GESTÃO.

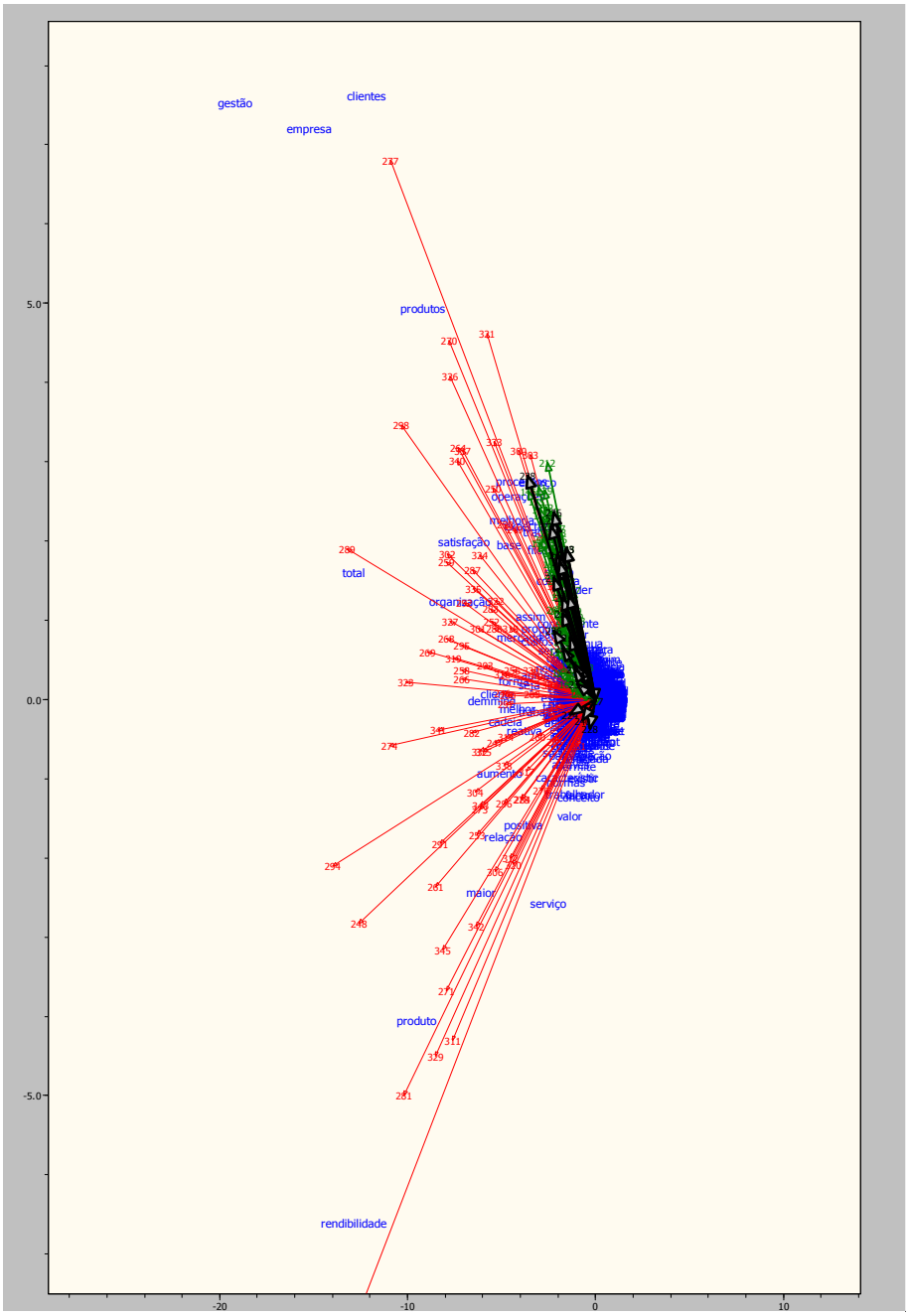


Figura 3.9.5.3. Espaço Semântico de 77 textos de Gestão sobre o qual foram projetados os restantes (AT e T).

A tabela seguinte – **tabela 3.9.5.1.** – apresenta os dados de validação dos resultados correspondentes aos parâmetros 0.75 para a proporção da amostra de testes (AT) e *k*= 5.

CL.SYS	CL.PROF
54	50
38	20
55	50
57	50
54	50
35	50
35	25
58	60
52	60
28	0
34	15
25	0
57	50
41	40
33	40
54	60
43	40
31	40
50	60
50	50
32	30
66	75
60	40
58	40
59	50

Tabela 3.9.5.1. Validação do classificador para 0.75 da AT e *k*= 5 vizinhos a considerar no cálculo ClSys. Valores de ClSys comparados com os valores reais ClProf.

Constata-se que, para estes valores experimentais – ver à frente a razão desta escolha – obtém-se um coeficiente de correlação (ClSys, ClProf) de 0.781 – largamente significativo.

A **figura 3.9.5.4.** contém uma reta de regressão construída para este valor usando os valores dos pares (ClSys, ClProf), verificando-se uma qualidade (R^2= 0.61) razoável para predizer ClProf em função de ClSys no contexto de uma avaliação contínua na sala de aula (não para atribuir uma nota com consequências administrativas).

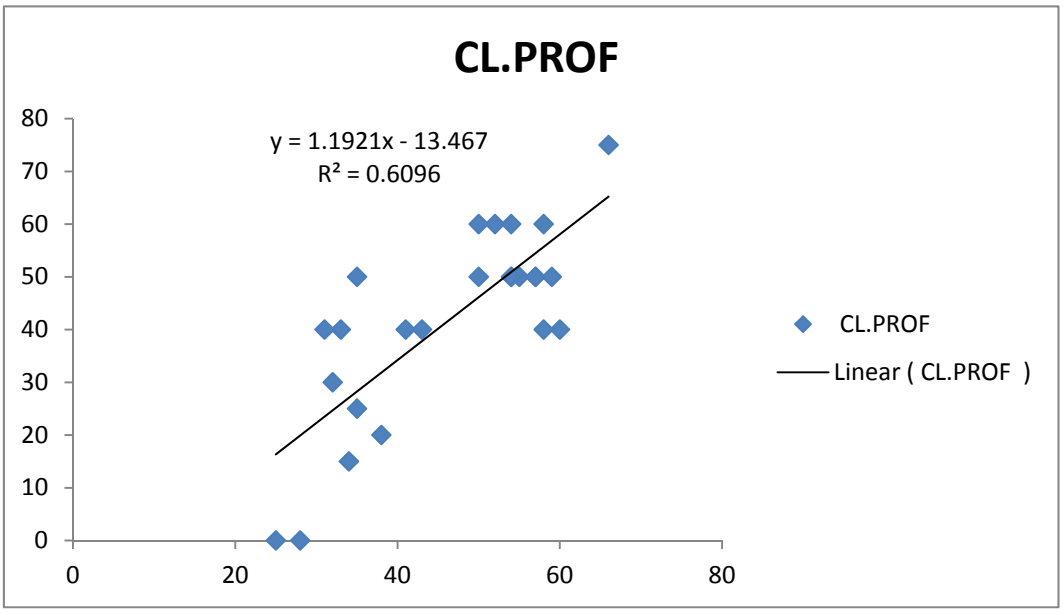


Figura 3.9.5.4. Reta de regressão de ClProf em função de ClSys para os parâmetros (0.75, 5).

Na tabela seguinte podem ver-se uma ANOVA destes valores, obtida com o SPSS.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4934	1	4934	35,912	,000[a]
	Residual	3160	23	137,391		
	Total	8094	24			

a. Predictors: (Constant), Clsys

b. Dependent Variable: ClProf

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-13,467	9,516		-1,415	0,17
	Clsys	1,192	0,199	0,781	5,993	0

a. Dependent Variable: ClProf

Tabela 3.9.5.2. Árvore dos valores.

Face a estes resultados, convém ainda testar as hipóteses de que as duas séries de valores (ClSys e ClProf) vêm da mesma distribuição e se vêm de populações com a mesma média.

A **tabela 3.9.5.3.** apresenta o resultado do teste de Kolmogorov-Smirnov para comparar duas amostras, concluindo-se que não há diferenças significativas (nível 0.01) entre as distribuições de origem de ClSys e ClProf.

Test Statistics[a]

		Notas
Most Extreme Differences	Absolute	0,253
	Positive	0,131
	Negative	-0,253
Kolmogorov-Smirnov Z		0,894
Asymp. Sig. (2-tailed)		0,4

a. Grouping Variable: Sys_Prof

Tabela 3.9.5.3. Clsys e ClProf têm a mesma distribuição.

Usando o método da ANOVA implementado no SPSS, constata-se que não há diferença significativa entre as médias das suas populações ClSys e ClProf ao nível de significância 0.01 – ver **tabela 3.9.5.4.**

ANOVA

Notas

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	317,222	1	317,222	1,323	0,256
Within Groups	11508,458	48	239,76		
Total	11825,68	49			

Tabela 3.9.5.4. Comparação de médias das duas amostras de valores ClSys e ClProf.

Finalmente, a figura seguinte, obtida com EXCEL – **figura 3.9.5.5.** – parece sugerir que um modelo mais adequado para predizer os valores de ClProf conhecendo os de ClSys fornecidos pelo sistema implementado (PAET) é um modelo quadrático a que corresponde R^2= 0.645 um pouco melhor.

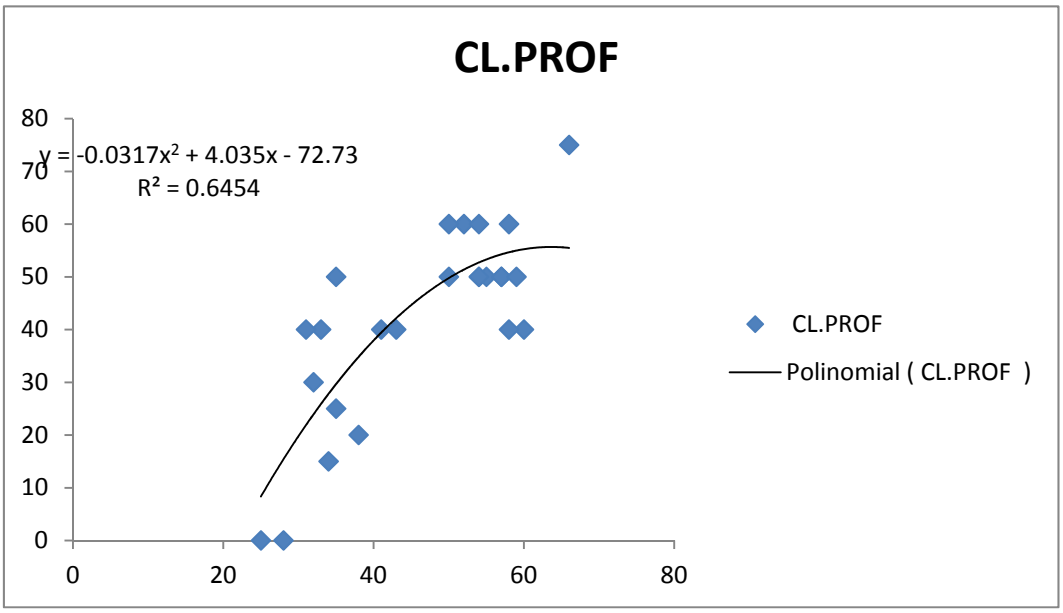


Figura 3.9.5.5. Ajustamento dos dados (ClSys e ClProf) usando um modelo quadrático.

A **figura 3.9.5.6.** contém os resultados das experiências conduzidas com estes dados variando o valor dos parâmetros (Percentagem para AT, Número de Vizinhos). Para cada par de valores é apresentado o coeficiente de correlação (ClSys, ClProf) para *k*= 1, 2, …, 8 e percentagem da AT= 50%, 60% e 70% constata-se que a melhor combinação (a que conduz ao máximo do coeficiente de correlação) é o par (75%, 8) com o coeficiente de correlação 0.78 objeto da análise anterior.

Nessa mesma figura pode ver-se o gráfico dos valores mencionados, constatando-se que uma percentagm de 75% para a amostra de treino conduz, consequentemente, a valores mais elevados do coeficiente de correlação (ClSys, ClProf).

k	50%	60%	75%
1	0,211	0,301	0,571
2	0,381	0,572	0,729
3	0,433	0,438	0,657
4	0,529	0,511	0,696
5	0,585	0,536	0,711
6	0,563	0,550	0,700
7	0,624	0,592	0,700
8	0,662	0,567	0,781

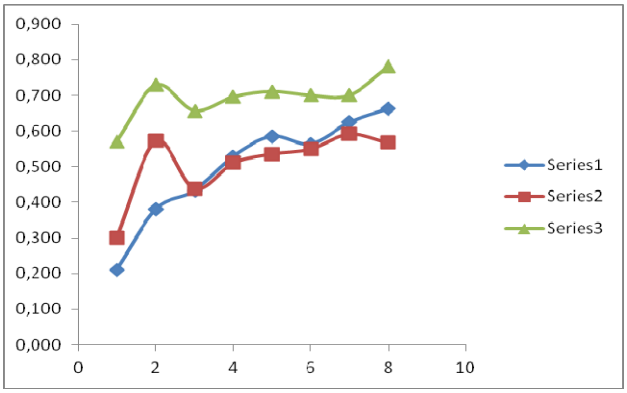


Figura 3.9.5.6. Obtém-se a correlação máxima r(Clsys,ClProf) para 75% e *k*= 8.

Em síntese, esta experiência mostra que é possível obter classificadores utilizáveis no âmbito da avaliação contínua, usando a metodologia da LSA/Biplot e dados

exclusivamente ao alcance dos professores na sua atividade letiva – o que constitui evidência a favor de uma das questões fundamentais em investigação.

3.9.6. Estudo Comparativo das Classificações dos Professores.

Os testes do conjunto EXMIN foram classificados segundo regras fixadas num protocolo publicado para cada um dos exames. Nesses protocolos são comuns regras a respeito de correção linguística (Sintaxe, Morfologia, Pontuação, Ortografia, Léxico), número mínimo de palavras, estruturação do texto, qualidade da fundamentação da resposta e outras.

Os professores contratados são instruídos na aplicação destes protocolos sendo o objetivo garantir homogeneidade de critérios e variabilidade mínima imputável a erros ou desvios do corretor.

Em última análise, o objetivo é o de que, para um exame específico, dirigido a uma população específica, os resultados, ainda que obtidos por corretores diferentes, tenham a mesma distribuição de resultados. Isto é, os resultados específicos atribuídos por um corretor específico devem ter características tais que, estatisticamente, não se distingam das distribuições dos outros corretores.

Quando não é este o caso (distribuições diferentes) pode ser necessário proceder a operações como "*test equating*", isto é, uma operação de homogeneização dos resultados que passe por converter certos resultados nos resultados equivalentes de uma distribuição comum.

No caso presente, o resultado usado é o que aparece imputado pelo corretor – inscrito na prova produzida pelo estudante e, portanto, antes de qualquer eventual homogeneização.

Uma vez que se ignora se esses resultados a que se teve acesso foram produzidos por um único corretor ou por vários, assume-se que foram produzidos por um único corretor fictício aqui designado por Ministério.

De acordo com a literatura consultada (Landauer, Foltz, & Laham, 1998), em particular as classificações atribuídas pelos classificadores automáticos construídos segundo a metodologia da ASL assumem uma perspetiva holística para a classificação. De resto, inicialmente assumia-se que estes classificadores funcionavam com base no chamado modelo "*bag of words*" para vincar bem que o que contava eram simplesmente as

frequências de ocorrência das palavras nos textos, não importando a ordem de ocorrência nem os aspetos gramaticais.

Contudo, a tendência atual é, sem alterar a natureza da ASL, a consideração de certos aspetos gramaticais relacionados com a ordem das palavras e a estrutura das frases, "inferidas" de modo experimental, guiados pelos dados (Kintisch, 2001; Dennis, 2005).

Sendo o nosso trabalho baseado no modelo "*bag of words*" sentiu-se a necessidade de verificar em que medida os resultados de uma classificação holística realizada por professores que classificassem os textos já classificados segundo os protocolos de correção do Ministério diferiam dessas classificações.

Estando, por razões económicas e de tempo, fora de causa contratar professores que reclassificassem todos os testes disponíveis, obteve-se a colaboração de um professor de Português do Ensino Secundário que, de forma totalmente graciosa, se dispôs a classificar, segundo este critério holístico, os testes do Ministério relativos a 60 respostas relativos à prova 639 de 2008.

A **tabela 3.9.6.1.** a seguir apresentada contém não só as classificações do Ministério relativas às duas questões envolvendo o Memorial do Convento, como as classificações atribuídas, segundo o critério holístico, a esses mesmos textos pelo professor que connosco colaborou.

Nº	Curso	Num Conv	Fase	Cod Exam	Data	A Min	B Min	Total Min	A HOL	B HOL	Total HOL
1	Ciências Sociais e Humanas	845	2.ª	639	22-07-2008	3,9	2	5,9	3	4	7
2	Ciências Sociais e Humanas	846	2.ª	639	22-07-2008	3,6	1,7	5,3	3,5	4	7,5
3	Ciências Sociais e Humanas	847	2.ª	639	22-07-2008	4,3	2,5	6,8	3	4	7
4	Ciências Sociais e Humanas	848	2.ª	639	22-07-2008	3,7	2,8	6,5	2,5	4,5	7
5	Ciências Sociais e Humanas	849	2.ª	639	22-07-2008	4,9	1,5	6,4	3,5	3,9	7,4
6	Ciências Sociais e Humanas	850	2.ª	639	22-07-2008	4,3	1,9	6,2	4	5	9
7	Ciências Sociais e Humanas	851	2.ª	639	22-07-2008	3,8	1,8	5,6	3	5	8
8	Ciências Sociais e Humanas	852	2.ª	639	22-07-2008	4,4	1,8	6,2	2,5	4	6,5
9	Ciências Sociais e Humanas	853	2.ª	639	22-07-2008	3,7	2,3	6	2,5	4	6,5
10	Ciências Sociais e Humanas	854	2.ª	639	22-07-2008	3,5	2,1	5,6	2	4	6
11	Ciências Sociais e Humanas	855	2.ª	639	22-07-2008	3,3	1,7	5	3	4	7
12	Ciências Sociais e Humanas	856	2.ª	639	22-07-2008	5	1,2	6,2	4	4,7	8,7
13	Ciências Sociais e Humanas	857	2.ª	639	22-07-2008	3,7	1,5	5,2	3	4	7
14	Ciências Sociais e Humanas	858	2.ª	639	22-07-2008	3,4	1,6	5	2	3,7	5,7
15	Ciências Sociais e Humanas	859	2.ª	639	22-07-2008	3,8	1,5	5,3	2,5	3,5	6
16	Comunicação	856	2.ª	639	22-07-2008	2,6	0	2,6	3	0	3
17	Ciências e Tecnologia	861	2.ª	639	22-07-2008	5,6	0,8	6,4	4,5	3,5	8
18	Ciências e Tecnologia	862	2.ª	639	22-07-2008	4	1,7	5,7	3	3	6
19	Ciências e Tecnologia	863	2.ª	639	22-07-2008	4,8	2	6,8	3	4	7
20	Ciências Sócio-Económicas	864	2.ª	639	22-07-2008	4,5	0	4,5	3	0	3
21	Ciências Sociais e Humanas	865	2.ª	639	22-07-2008	4,5	0,8	5,3	4	3	7
22	Ciências e Tecnologia	866	2.ª	639	22-07-2008	5,7	1,9	7,6	3,5	4	7,5
23	Ciências e Tecnologia	867	2.ª	639	22-07-2008	4,3	1	5,3	3,5	3	6,5

(continuação)

Nº	Curso	Num Conv	Fase	Cod Exam	Data	A Min	B Min	Total Min	A HOL	B HOL	Total HOL
24	Administração	868	2.ª	639	22-07-2008	4,3	1,9	6,2	4	4	8
25	Ciências e Tecnologia	869	2.ª	639	22-07-2008	4,3	0,8	5,1	2,5	4	6,5
26	Ensino Recorrente	870	2.ª	639	22-07-2008	6,4	2,8	9,2	3	4	7
27	Ciências e Tecnologia	871	2.ª	639	22-07-2008	5,9	2	7,9	4		4
28	Ciências e Tecnologia	872	2.ª	639	22-07-2008	4,9	2,1	7	3,5	4	7,5
29	Administração	873	2.ª	639	22-07-2008	4,9	1,7	6,6	2,5	4	6,5
30	Ciências e Tecnologia	874	2.ª	639	22-07-2008	2,1	0,6	2,7	2	3	5
31	Ciências Sócio-Económicas	875	2.ª	639	23-07-2008	5	2,4	7,4	4,5	4,5	9
32	Ciências e Tecnologia	876	2.ª	639	23-07-2008	4,8	2,3	7,1	4	4	8
33	Ciências e Tecnologia	877	2.ª	639	23-07-2008	3	2	5	3	3	6
34	Artes Visuais	878	2.ª	639	23-07-2008	4,5	1,7	6,2	3,5	4	7,5
35	----------	879	2.ª	639	23-07-2008	5,4	3	8,4	5	4,9	9,9
36	Ciências Sócio e Humanas	880	2.ª	639	23-07-2008	3,3	1,6	4,9	2	3,5	5,5
37	Ciências Prof. Comunicação	881	2.ª	639	23-07-2008	5,1	2,4	7,5	3,5	3,7	7,2
38	Ciências e Tecnologia	882	2.ª	639	23-07-2008	2,4	2,1	4,5	2	3	5
39	I - Informática	883	2.ª	639	23-07-2008	4,8	0,7	5,5	4	3	7
40	Ciências e Tecnologia	884	2.ª	639	23-07-2008	3,5	2,2	5,7	3	4	7
41	Tecn. de Administração	885	2.ª	639	23-07-2008	4,6	2,4	7	3,5	4	7,5
42	Ciências e Tecnologia	886	2.ª	639	23-07-2008	4	1,6	5,6	2,5	3	5,5
43	1 - Científico-Natural	887	2.ª	639	23-07-2008	4,4	2,2	6,6	3	3	6
44	Agrupamento 1 / Geral	888	2.ª	639	23-07-2008	4,1	0	4,1	3	1	4
45	Ciências Sociais e Humanas	889	2.ª	639	23-07-2008	3,5	1,5	5	3,5	4	7,5
46	Ciências Sociais e Humanas	890	2.ª	639	23-07-2008	4,3	1,7	6	3,5	4	7,5
47	Tecn. de Administração	891	2.ª	639	23-07-2008	4,9	1,8	6,7	2,5	2,5	5
48	----------	892	2.ª	639	23-07-2008	1,9	1,1	3	1,5	2,5	4
49	Ciências e Tecnologia	893	2.ª	639	23-07-2008	5,6	2	7,6	4,5	4,5	9
50	Ciências Sociais e Humanas	894	2.ª	639	23-07-2008	4,8	0,9	5,7	3,5	3,5	7
51	Ciências e Tecnologia	895	2.ª	639	23-07-2008	3,3	2,8	6,1	3,5	4,5	8
52	Ciências e Tecnologia	896	2.ª	639	23-07-2008	6	1,2	7,2	2,5	4,5	7
53	----------	897	2.ª	639	23-07-2008	5,3	1,7	7	2	3	5
54	Ciências Sociais e Humanas	898	2.ª	639	23-07-2008	5,4	1,7	7,1	3	4	7
55	Ciências e Tecnologia	899	2.ª	639	23-07-2008	4,4	2,1	6,5	3	4	7
56	Ciências Socioeconómicas	900	2.ª	639	23-07-2008	2,8	0,7	3,5	3,5	4	7,5
57	Ciências e Tecnologia	901	2.ª	639	23-07-2008	5,3	1,7	7	4	3	7
58	Ciências e Tecnologia	902	2.ª	639	23-07-2008	6,1	2,2	8,3	4	4	8
59	Ciências e Tecnologia	903	2.ª	639	23-07-2008	4,4	1,7	6,1	3	3,5	6,5
60	Tecn. de Administração	904	2.ª	639	23-07-2008	4,8	0	4,8	3,5	0	3,5
61	Ciências e Tecnologia	905	2.ª	639	23-07-2008	2	1,6	3,6	1	4	5

Tabela 3.9.6.1. Classificações do Ministério e classificações atribuídas segundo o critério holístico.

Ao reclassificar os textos segundo um critério holístico, o professor referido ignorava a nota oficial do Ministério.

Na **tabela 3.9.6.2.** apresenta-se o resumo estatístico dos resultados oficiais (MIN) e das classificações holísticas (HOL), obtido com o *software* SPSS v17.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
A_Min	61	1,9	6,4	4,292	1,0159
B_Min	61	0	3	1,66	0,7
Total_Min	61	2,6	9,2	5,948	1,3422
A_HOL	61	1	5	3,14	0,781
B_HOL	60	0	5	3,57	1,067
Total_Hol	61	3	10	6,65	1,459
Valid N (listwise)	60				

Tabela 3.9.6.2. Resumos estatísticos das classificações holísticas (HOL) e do Ministério (MIN).

Como se pode verificar, as maiores diferenças a nível de médias observam-se em relação à B) que no caso das classificações tem um valor substancialmente inferior à nota holística. Vê-se que, neste caso, as duas médias A) e B) são da mesma ordem de grandeza ao passo que no caso do Ministério são substancialmente diferentes.

Precisando estas comparações através de uma análise de variância para comparação de médias (em que os grupos MIN ou HOL são definidos por uma variável com dois valores 1 e 2) obtém-se a **tabela 3.9.6.3.**

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
A	**Between Groups**	40,509	1	40,509	49,355	0
	Within Groups	98,491	120	0,821		
	Total	139	121			
B	**Between Groups**	110,262	1	110,262	135,905	0
	Within Groups	96,547	119	0,811		
	Total	206,809	120			
Total	**Between Groups**	11,588	1	11,588	5,412	0,022
	Within Groups	256,933	120	2,141		
	Total	268,521	121			

Tabela 3.9.6.3. Comparação de médias entre grupos MIN e HOL para as respostas à A), à B) e Total.

Nesta tabela confirma-se a perceção captada pelas estatísticas descritivas da **tabela 3.9.6.1.**

Isto é, a nível de médias obtidas com os dois métodos (MIN e HOL) verifica-se que as médias A), B) e Total) são significativamente distintas ao nível de significância 0.05,

sendo as médias obtidas com o segundo método (HOL) significativamente superiores no caso A) (4.3 > 3.1) e o contrário no caso B) (1.7 < 3.6).

Uma outra questão é a de saber se seria possível predizer os resultados do MIN conhecendo os resultados da classificação holística.

A **tabela 3.9.6.4.** contém as correlações entre as três variáveis relativas ao Ministério e as variáveis correspondentes, relativas à classificação holística.

Verifica-se que existem correlações significativas entre os resultados obtidos pelo ministério e os obtidos pelo método holístico. Com efeito,

r (A-MIN, A-HOL)= 0.577 (Significativo ao nível 0.01)

r (B-MIN, B-HOL)= 0.696 (Significativo ao nível 0.01)

r (Total-MIN, Total-HOL)= 0.562 (Significativo ao nível 0.01)

Correlations

		A_Min	B_Min	A_HOL	B_HOL	Total_Min	Total_Hol
A_Min	Pearson Correlation	1	0,196	,577**	0,18	,859**	,372**
	Sig. (2-tailed)		0,129	0	0,168	0	0,003
	N	61	61	61	60	61	61
B_Min	Pearson Correlation	0,196	1	0,1	,696**	,670**	,537**
	Sig. (2-tailed)	0,129		0,444	0	0	0
	N	61	61	61	60	61	61
A_HOL	Pearson Correlation	,577**	0,1	1	0,181	,489**	,620**
	Sig. (2-tailed)	0	0,444		0,167	0	0
	N	61	61	61	60	61	61
B_HOL	Pearson Correlation	0,18	,696**	0,181	1	,505**	,844**
	Sig. (2-tailed)	0,168	0	0,167		0	0
	N	60	60	60	60	60	60
Total_Min	Pearson Correlation	,859**	,670**	,489**	,505**	1	,562**
	Sig. (2-tailed)	0	0	0	0		0
	N	61	61	61	60	61	61
Total_Hol	Pearson Correlation	,372**	,537**	,620**	,844**	,562**	1
	Sig. (2-tailed)	0,003	0	0	0	0	
	N	61	61	61	60	61	61

**. Correlation is significant at the 0.01 level (2-tailed).

Tabela 3.9.6.4. Correlações entre as classificações do ministério e holístico.

No caso dos totais, a figura seguinte – **figura 3.9.6.1.** – mostra essa informação graficamente e responde à questão de saber se seria possível predizer os resultados do Ministério conhecendo os holísticos. A resposta é afirmativa mas com grandes erros, como mostra abaixa qualidade da regressão (R^2= 0.32) da ordem de grandeza dos valores obtidos

ao tentar predizer as classificações do ministério usando as classificações do sistema implementado.

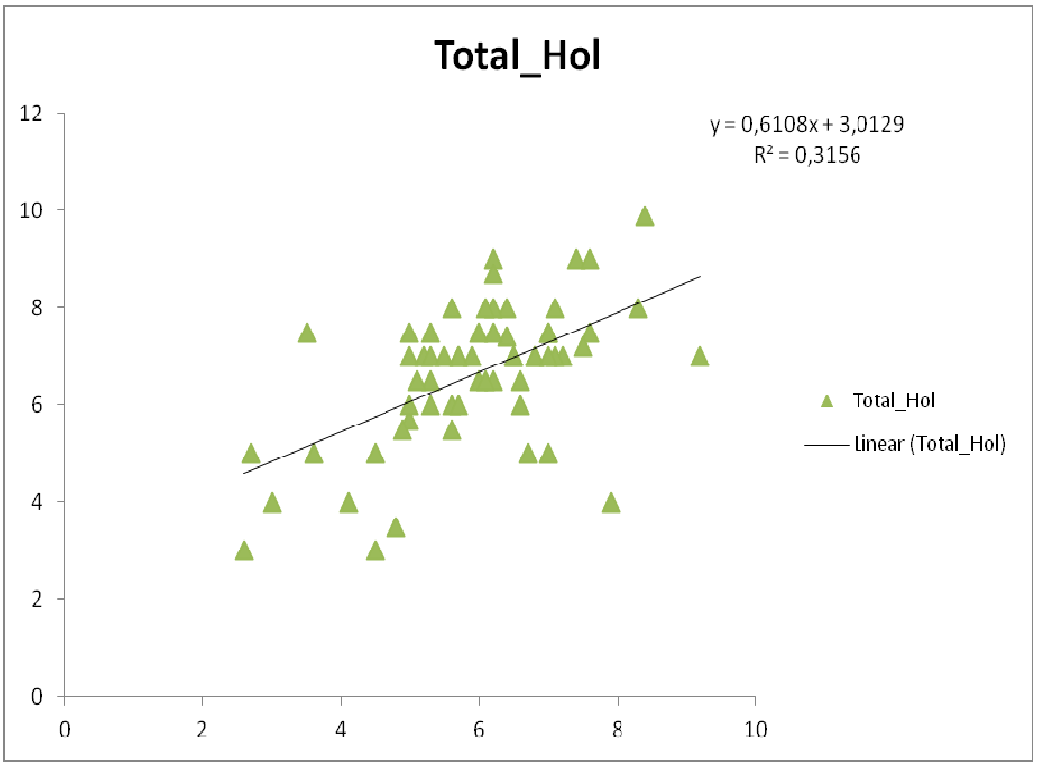


Figura 3.9.6.1. A figura mostra que a reta de previsão dos resultados do ministério conhecendo os do método holístico tem fraca qualidade.

Os gráficos para A) e B) são semelhantes.

Finalmente, interessa responder à questão de saber se – sim ou não – se pode considerar que os resultados holísticos e do Ministério vêm da mesma distribuição populacional. A resposta a esta questão obtém-se realizando testes (não paramétricos) de Mann-Whitney e Kolmogorov-Smirnov, cujos resultados aparecem na **tabela 3.9.6.5.**

Test Statistics[a]

	A	B	Total
Mann-Whitney U	683	241	1323
Wilcoxon W	2574	2132	3214
Z	-6,052	-8,288	-2,759
Asymp. Sig. (2-tailed)	0	0	0,006

a. Grouping Variable: Classif

Test Statistics[a]

		A	B	Total
Most Extreme Differences	Absolute	0,557	0,884	0,328
	Positive	0	0,884	0,328
	Negative	-0,557	0	-0,033
Kolmogorov-Smirnov Z		3,078	4,86	1,811
Asymp. Sig. (2-tailed)		0	0	0,003

a. Grouping Variable: Classif

Tabela 3.9.6.5. Os resultados dos testes mostram que as distribuições (HOL) são significativamente diferentes das distribuições (MIN).

Como pode ver-se, há diferenças significativas nas distribuições populacionais dos resultados obtidos pelos dois métodos de classificação (holístico e segundo os critérios do Ministério).

Em síntese, embora na experiência se usem os resultados de um só professor, pode concluir-se que os resultados obtidos pelo método holístico aplicado são substancialmente distintos dos obtidos pelo método do Ministério. Além disso, constata-se que, do ponto de vista da validade (dada pelas correlações) as correlações (ClSys, ClProf) e (ClHol, ClProf) são da mesma ordem de grandeza, especialmente quando consideramos os resultados obtidos pelos dados GESTÃO – ver **3.9.5.** acima.

Isto significa que os resultados obtidos com o método de classificação automática de textos de resposta aberta integrado no sistema PAET são encorajadores e apoiam a expetativa de futuros desenvolvimentos interessantes nesta direção.

Conclusão

Na introdução a este trabalho foram estabelecidos os grandes objetivos, motivações e resultados a atingir.

Foram aí também formuladas as grandes questões a investigar (**I**, **II** e **III**). Neste capítulo sintetizam-se os principais resultados e as respostas às questões formuladas.

Em relação à Questão I (estado da investigação relativa à avaliação automática com base nos textos de resposta a questões abertas), constatou-se que os Estados Unidos, depois de uma experiência acumulada de décadas, preparam-se para generalizar a todo o território, o uso da avaliação automática até mesmo dos exames finais, com implicações para a vida dos estudantes. Estes desenvolvimentos e a investigação aplicada em que se baseou assentam em organismos independentes como o ETS (*Educational Testing Service*). Depois de naturais resistências iniciais por parte dos professores esta tecnologia tem agora uma aceitação generalizada pelo facto de esses instrumentos serem percebidos não como substitutos do professor mas como garantia de uma melhor qualidade do trabalho dos mesmos.

Em relação à Questão II (identificação de problemas ligados à construção de sistemas automáticos de avaliação de conhecimentos), a nossa própria experiência mostrou que há que distinguir claramente entre a criação de sistemas que sejam capazes de produzir resultados indistinguíveis dos obtidos pelos classificadores humanos de exames sumativos nacionais (como sucede atualmente nos Estados Unidos) e sistemas capazes de produzir resultados úteis à atividade diária dos professores na sua atividade letiva diária, implicando conceber e avaliar resultados de testes formativos associados à avaliação contínua.

Pela nossa própria experiência, ao conceber e implementar um sistema experimental deste tipo (ver **Capítulos II** e **III**, análise de dados) e depois ao analisar estatisticamente os seus resultados, concluiu-se que um sistema para os exames sumativos – especialmente os pensados para exames nacionais – têm de ter caraterísticas que incorporem critérios de avaliação que têm que ver com questões ligadas aos aspetos linguísticos (Sintaxe, Morfologia, Pontuação, Ortografia, Léxico), estrutura de texto, presença ou ausência de certas características que só podem ser incorporadas através da utilização de conceitos de inteligência artificial – para lá de exigirem depois a realização de

experiências com grande quantidade de dados, controladas segundo metodologias psicométricas. Isto significa que, embora a esmagadora maioria das técnicas e metodologias seja relativamente bem conhecidas, a sua implementação exige uma organização com largos recursos humanos, científicos e financeiros a nível nacional.

Uma vez que a implementação dessa metodologia permite economizar depois recursos consideráveis ligados à administração do processo de exames, crê-se que é um problema a considerar na estruturação futura do Ministério de Educação.

Já quanto à implementação de sistemas que apoiem o professor na elaboração e avaliação de questões de resposta aberta ligadas à avaliação contínua, constatou-se pela nossa própria experiência de elaboração de um protótipo de um sistema desse tipo (o PAET – Programa de Análise Estatística de Textos) que não só é realizável como é possível com recursos modestos, como os resultados experimentais apresentados em **3.9.6.** sugerem.

Isto é, é possível pensar em desenvolver sistemas deste tipo que não exijam mais recursos para funcionar do que aqueles que estão ao alcance dos professores na sua atividade letiva. É admissível que estes sistemas possam ser produzidos e testados por empresas privadas e que, para serem usados pelos professores dependam apenas de recursos ao seu alcance.

Em relação à Questão III, os factos têm ultrapassado as expetativas, verificando-se a emergência recente, durante o tempo que durou a atividade investigadora que conduziu a esta tese, do desenvolvimento da atividade designada por EDM (*Educational Data Mining*) que poderíamos designar um Português por MDE – Mineração de Dados Educacionais – abrangendo o uso intenso das técnicas de análise estatística de textos e outros dados na extração de conhecimentos e tomada de decisão (por estudantes e professores) a partir do funcionamento das plataformas eletrónicas formas de ensino, sistemas administrativos e sistemas de investigação ligadas ao ensino. Considera-se, pois, que a elaboração desta tese e a correspondente investigação se encontra em completa sintonia com a realidade atual e tendências percetíveis.

A atividade investigadora teve de ultrapassar algumas dificuldades. A principal das consistiu na ausência de apoios financeiros, obrigando o signatário a financiar do seu próprio bolso atividades como a recolha e o registo em suportes magnéticos de todos os textos envolvidos no projeto.

Com efeito, esta metodologia só é aplicável na prática se incidir em documentos facilmente acessíveis em suportes magnéticos – como é, felizmente, a tendência atual. Apesar dos progressos no OCR (*Optical Character Recognition*), está para já fora de questão usá-lo para as tarefas de transcrição de documentos manuscritos.

As resistências encontradas conduziram à impossibilidade prática de aceder aos conteúdos e experiências baseadas em plataformas de *e-learning* e a ter de encarar a necessidade de pedir autorização ao Ministério de Educação para aceder a exames nacionais em papel e respetivas classificações, arquivadas nas escolas onde foram realizados.

Houve, enseguida, a necessidade de fotocopiar todo esse material e manuais de ensino selecionados que tiveram de ser transcritos para suportes digitais. Todo este processo levou muitos meses e consumiu consideráveis recursos – com grandes inconvenientes e prejuízo do calendário inicial.

Apesar de todo o esforço, a amostra com que foi possível trabalhar não permitiu conclusões isentas de alguma ambiguidade. Ver experiências descritas em **3.9.1.** e **3.9.2.**

Em contraste, os textos relativos à formação contínua no contexto de um instituto politécnico disponibilizados pela Professora Florinda Matos, já tinham as imagens digitalizadas pelo que foi apenas necessário realizar a respetiva transcrição mediante um processador de texto para textos utilizáveis pelo programa desenvolvido.

Considera-se como um dos principais resultados obtidos, o desenvolvimento do programa PAET que, servindo de protótipo para futuros desenvolvimentos, serviu também de banco de ensaios sobre o qual foram testadas ideias a ter em conta no futuro e com o qual se realizaram as experiências de análise multivariadas descritas no **Capítulo III**.

Um outro resultado que se considera importante é o do estabelecimento da relação entre ASL e os biplots, sendo a exploração desta relação um instrumento importante da exploração dos dados experimentais usados nesta tese e em futuros sistemas a desenvolver.

Com efeito, embora a literatura mostre que nas análises de textos das respostas a questões abertas seja necessário considerar e usar dimensões muito superiores às dimensões 1, 2, ou 3 onde é habitualmente possível visualizar a informação, o uso dos biplots para representar esta informação em baixa dimensão é um poderoso auxiliar de intuição e do raciocínio geométrico, suporte da intuição. Concluiu-se que o uso sistemático dessa dualidade é a chave no desenvolvimento futuro de sistemas operacionais "amigos do

utilizador" tanto mais que desenvolvimentos recentes relativos aos chamados biplots cilíndricos reforçam esta perspetiva (Vairinhos & Galindo, 2012).

Tendo em conta o que precede, as conclusões desta investigação são as seguintes:

1. É possível e aconselhável o uso crescente de técnicas de mineração de textos em tarefas ligadas à atividade letiva tanto de professores como de estudantes no contexto, de uma avaliação formativa contínua.
2. É possível – com meios pouco dispendiosos, como foi demonstrado pelo protótipo desenvolvido – desenvolver instrumentos de apoio à atividade docente que torne possível o uso em grande escala de questões de resposta aberta, sem grande sobrecarga para os possíveis utilizadores.
3. É possível desenvolver sistemas desse tipo que façam uso apenas de recursos acessíveis à generalidade dos docentes.
4. O funcionamento de sistemas deste tipo só é possível quando os textos envolvidos (manuais e respostas) estão em suporte magnético e sob a forma de texto.
5. O desenvolvimento de sistemas do tipo SAAT para exames sumativos a nível nacional é uma tendência internacional mas exige infraestruturas de investigação e produção a nível do Ministério da Educação.
6. Foi desenvolvido um protótipo que permite servir de banco de ensaios às metodologias envolvidas no desenvolvimento de sistemas futuros de apoio à atividade de elaboração de testes baseados em questões abertas.
7. As análises de dados realizadas no sistema protótipo desenvolvido puseram em evidência questões teóricas e práticas a abordar em projetos futuros nesta direção.

Referências Bibliográficas

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: a four process architecture. *Journal of Technology, Learning, and Assessment, 1*(5). Retrieved from: http://www.itla.org.

Alvares, R. V. (2005). *Investigação do processo de Stemming na língua portuguesa*. Dissertação de Mestrado. Niterói: Universidade Federal Fluminense.

Amershi, S., & Conati, C. (2009). Combining Unsupervised and Supervised Machine Learning to Build User Models for Exploratory Learning Environments. *Journal of Educational Data Mining*, *1*(1), 18-71.

Anaya, A. R., & Boticário, J. G. (2011). Application of machine learning techniques to analyse student interactions and improve the collaboration process. *Expert Systems with Applications*, *38*, 1171–1181. doi: 10.1016/j.eswa.2010.05.010

Antunes, C. (2010) Anticipating student's failure as soon as possible. *In Romero, Ventura, Pechenizkiy, & Baker*, 353-363.

APA – American Psychological Association (2003). Guidelines on multicultural education, training, research, practice, and organizational change for psychologists. *American Psychologist,* May 2003, 377-402. Retrieved from http://www.indigenouspsych.org//resources/guidelines(2003)%20apa.pdf.

Arends, R. (2008). *Aprender a ensinar*. Lisboa: McGraw-Hill (7.ª ed.).

Arias, R. M. (1996). *Psicometría: teoria de los testes psicológicos y educativos.* Sintesís Psicologia. Madrid: Síntesis.

Arnold, K.E. (2010). Signals: Applying Academic Analytics. *Educause Quarterly, 33*(1).

Arroyo, I., & Woolf, B. P. (2005). Inferring learning and attitudes from a Bayesian Network of log file data. *Proceedings of the 12th International Conference on Artificial Intelligence and Education*, 33-40.

Attali, Y., & Bustein, J (2004). Automated essay scoring with E-rates© v.2.0. *Conference of International Association for Educational Assessment (IAEA). Philadelphia, PA.,* June 13 -18.

Baker, F. B. (1992). *Item response theory, parameters estimation techniques*. New York: Marcel Dekker, Inc.

Baker, F. B., Ryan, S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: a review and future visions. *2nd International Conference on Educational Data Mining. Cordoba, Spain, July 1-3, 2009*.

Bejar, I. I. (2011). A validity-based approach to quality control and assurance of automated scoring, assessment in education. *Principles, Policy & Practice, 18*(3), 319-341.

Bellegarda, J. R. (2007). *Latent semantic mapping: principles & applications*. Georgia, Atlanta: Morgan & Claypool Publishers.

Benoit, G. (2002). Data Mining. *Annual Review of Information Science and Technology*, *6*, 265-310.

Benzécri, J. (1973). *L'analyse dês données*. Paris: Dunod.

Bergner, Y., Dröschler, S., Kortmeyers, G., Rayyan, S., Seaton, D., & Pritchard, D. E. (2012). Model – Based collaborative filtering analysis of student response data: machine-learning item response theory. *Proceeding of the 5th International Conference on Educational Data Mining*, 87-94.

Biletska, O., Biletskiy, Y., Li, H., & Vovk, R. (2010). A semantic approach to expert system for e-Assessment of credentials and competencies. *Expert Systems with Applications,* 37, 7003–7014. doi:10.1016/j.eswa.2010.03.018

Bloom, P. (2000). *How children learn the meaning of words*. Cambridge, MA: MIT.

Bond, T. G., & Fox, C. M. (2007). *Applying the rasch model. Fundamental measurement in the human sciences*. Mahwah, New Jersey: Lawrence Erlebaum Associates (LEA).

Botana, G. J., Léon, J. A., Olmos, R., & Escudero, I. (2010). Latent semantic analysis parameters for essay evaluation using small-scale corpora. *Journal of Quantitative Linguistic*, *17*(1), 1-29.

Bridgeman, B., Trapant, C., & Attalli, Y. (2012). Comparasion of human and machine scoring of essay's: differences by gender, ethnicity and country. *Applied Measurement in Education, 25*(1), 27-40.

Bruner, J. (1996). *The culture of education*. Cambridge, MA: Harvard University Press.

Burnstein, J. (2003). The E-rater® scoring engine: automated essay scoring with natural language processing. *In Shermis & J. Burnstein*, 113-122.

Cheng, S. L. (2008). Intelligent cognition-based systems approach to multiple criteria computerized essay assessment. *System Research*, *27*(6), 680-696.

Chodorow, M., & Burnstein, J. (2004). Beyond essay length: evaluating e-rater©'s performance or TOEFL® essays. *ETS – Educational Testing Service Research Reports*. Report Number: RR-04-04, TOEFL-RR-73.

Chomsky, N. (1968). *Language and mind*. New York: Harcourt, Brace & World. Retrieved from http://www.chomsky.info/books.htm.

Cizek, G. J., & Page, B. A. (2003). The concept of reliability in the context of automated essay scoring. *In Shermis, & Burnstein*, 125-146.

Dennis, Simon (2005). A memory-based theory of verbal cognition. *Cognitive Science*, *29*, 145-193.

Dewey, J. (1916). *Democracy and education*. New York: Macmillan.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment, 5*(1). Retrieved from: //www.jtla.org

D'Mello, S., & Graesses, A. (2010) Mining bodily patterns of affective experience during learning. *Proceedings of the 3rd Educational Data Mining*.

Dolezal, S. E., Welsh, L. M., Pressley, M., & Vincent, M. M. (2003). How nine third-grade teachers motivate student academic engagement. *Elementary School Journal*, *103*(3), 239-269.

Duwairi, R. M. (2006). A framework for the computerized assessment of university student essay. *Computers in Human Behavior*, *22*(3), 381-388.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement*. New Delhi: Prentice-Hall.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika, 1*(3), 212-218.

Ercan, T. (2011). Benefits of semantic approach in the learning environment. *Procedia - Social and Behavioral Sciences*, *28*, 963 – 967. doi: 10.1016/j.sbspro.2011.11.177

ETS – Educational Testing Service (2013). *Automated scoring of natural language*. Retrieved from: http://www.ets.org/research/topics/as-nlp/

Executive Office of the President – President's Council of Advisors on Science and Technology (2010). *Report to the president and congress designing a digital future: Federally funded research and development in networking and information technology*. Executive Office of the President of the United States.

Fayyad, U. M. , Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy R. (1996). *Advances in knowledge discovery and data mining*. California: MIT Press.

Feldman, R., & Sanger, J. (2007). *The text Mining handbook*: *advanced approaches in analyzing unstructured data*. Cambridge: Cambridge University Press.

Friedman, J. H., & Jacqueline, J. M. (2004). Clustering objects on subsets of attributes. *Journal of the Royal Statistical Society*. *Series B (Statistical Methodology)*, *66*(4), 815-849. Retrieved from: http://www.jstor.org/stable/3647651

Gabriel, K. R. (1971). The biplot graphie of matrices with application to principal component analysis. *Biometrika*, *58*(3), 453-467.

Galindo, M. P. (1985). *Constribuiciones a la representación simultanea de datos multidimensionales*. Tesis Doctoral. Universidad de Salamanca.

He, Y., Hui, S. C., & Quan, T. T. (2009). Automatic summary assessment for intelligent tutoring systems. *Computers and Education*, *53*, 890-899.

Hill, S. L., Wang, Y., Riachi, I., Schurmann, F., & Markram, H. (2012). *Statistical connectivity provides a sufficient foundation for specify functional connectivity in neo cortical neural microcircuits*. PNAS, E2885-E2894.

Islam, M. M., & Hoque, A. S. M. L. (2010). Automated essay scoring using generalized latent semantic analysis. *Proceedings of the 15th International Conference of Computer and Information Technology (I CCIT 2010)*.

Jorge-Botana, G., León, J. A., Olmos, R., & Escudero, I. (2010). Latent semantic analysis parameters for essay evaluation using small-scale corpora. *Journal of Quantitattive Linguistics, 17*(1), 1-29.

Keeves, J. P. (1988). *Educational research, methodology and measurement*: *An international handbook*. Oxford, England: Pergamon Press.

Keith, T. Z. (2003). Validity of automated essay scoring systems. *In Shermis, & Burmstein,* 147-167.

Kintsch, Walter (2001). Predication. *Cognitive Science*, *25*, 173-202.

Lam, Dullon, & Chang (2010). Towards the use of semi-structured annotators for automated essay grading. 4th IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2010).

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representations of knowledge. *Psychological Review*, *104*, 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*, 259-284.

Landauer, T. K., Foltz, P. W., & Laham, D. (2003). Automatic essay assessment. *Assessment in Education, 10*(3), 295-308.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis*. Mahwah, New Jersey: Lawrence Erlbaum Associates (LEA).

Landauer, T. K. (2011). Pearson's text complexity measure. White paper. Pearson. Retrieved from http://www.pearsonassessments.com/research.

Lebart, T., Salem, A. Y., & Benny, L. (1998). *Exploring textual data*. Dordrecht: Kluwer Academic Publishers.

Linden, W. J. (1997). *Handbook of modern item response theory*. New York. Springer-Verlag.

Liu, O. L., Lee, H.-S., & Linn, M. C. (2011). An investigation of explanation multiple-choice items in science assessment. *Educational Assessment*, *16*(3), 164-184.

Longford, N. T. (1995). *Models for uncertainty in educational testing*. NY: Springer-Verlag.

Louwerse, M. M., & Ventura, M. (2005). How children learn the meaning of words and how LSA does it (too). *Journal of the Learning Sciences, 14*(2), 301-309.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big Data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute. Retrieved from:

http://www.mckinsey.com/insights/business-technology/big-data-the-next-frontier-for-innovation.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching*. NJ: Pearson.

Mislevy, R. J. (1992). *Linking educational assessments: concepts, issues, methods, and prospects.* NY: Princeton.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence – centered design.* ETS – Educational Testing Service. NJ: Princeton.

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: evidence – centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining, Article 2, 4*(1).

Moretti, F. (2005). *Graphs, Maps, Trees*. London: Verso.

Mu, J., Stegmann, K., Mayfiled, E., Rosé, C., & Fisher, F. (2012). The ACODEA framework: Developing segmentation and classification schemes for fully automatic analysis of online discussion. *International Journal of Computer - Supported Collaborative Learning*, *7*(2), 285-305. Retrieved from: doi: 10.1007/s 11412 – 012 – 9147 – y.

Muñiz, H. (1996). *Psicometría*. Madrid: Editorial Universita, S. A.

Newman, D. A., & Vermunt, J. K. (2011). Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement equivalence. *Organizational Research Methods*, *14*(1), 147-176.

Oliveri, M. E., & Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, *53*(3), 315-333.

Orengo, V. M., & Huyck, C. (2001). *A stemming algorithm for the portuguese language*. IEEE String Processing and Information Retrieval. SPIRE 2001. IEEE Xplore Digital Library. Retrieved from: http://ieeexplore.ieee.org/xpe/.

Osuna Marin, Z. (2006). *Contribuciones al análisis de datos textuales*. Tesis Doctoral. Universidad de Salamanca.

Page, E. B. (1994). Computer grading prose, using modern concepts and software. *Journal of Educational Education, 62*(2), 127-133.

Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research, and applications* (2nd ed.). New York: Prentice-Hall.

Porter, M. F. (1980). *The Porter stemming algorithm*. Retrieved from http://tartarus.org/martin/PorterStemmer/index-????.htlm.

Powers, D. E., Burnstein, J. C., Chodorov, M., Fowles, M. E., & Kukich, K. (2001). Stumping E-rater: Challenging the validity of automated essay scoring. *ETS – Educational Texting Service*.

Ranadivé, & Maney (2011). *The two-second advantage*. New York: Crown Business.

Rauch, D. P., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling*, *52*(4), 354-379.

Ribeiro, L. C. (1989). *Avaliação de Aprendizagens*. Lisboa: Texto Editora.

Rocha, G. G., & Coelho, J. M. A. (2009). *Incorporação de um lematizador da língua portuguesa a um agente de recuperação de informação baseado em algoritmos genéticos*. Campinas: Anais do XIV Encontro de Iniciação Científica da PVC.

Romero, C., Ventura, S., Espejo, P. G., & Hevais, C. (2008). Data mining to classify students. *Processings of the 1st International Conference on Educational Data Mining, Montreal, Quebec, June 20-21, 2008.*

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. D. (2011). *Handbook of educational data mining*. New York: CRC Press.

Rudner, L. M., & Gagne, P. (2001). An overview of three approaches to scoring written essays by computer. *Practical Assessment Research & Evaluation*.

Rudner, L. M., & Liang, T (2002). Automated essay scoring using Bayes's theorem. *Journal of Technology, Learning and Assessment, 1*(2). Retrieved from: http://www.jtla.org.

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of the Intellimetric TM essay scoring systems. *The Journal of Technology, Learning and Assessment, 4*(4). Retrieved from: http://www.jtla.org.

Rupp, A. A., Nugent, R., & Nelson, B. (2012). Evidence centered design for diagnostic assessment within digital educational data mining. *Journal of Educational Data Mining, 4*(1), 1-10.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613-620.

SPSS (2007). *SPSS statistics 17.0 algorithms*. Chicago: SPAA Inc.

Schaefer, W. D. (1991). Essencial assessment skills in professional education for teachers. *Educational Measurement: Issues and Paractices, 10*(1), 3-6.

Shermis M. D., & Burstein, J. C (ed.) (2003). *Automated essay scoring - a cross - disciplinary perspective*. Mahwah: Lawrence Erlbaum.

Shmueli, G., & Koppins, O. (2011). Predictive analytics in information systems research. *MIS Quartely, 35*(3), 553-572.

Stiggins, R. J. (2004). *Student-involved assessment for learning* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Têng, Ssu-yu (1943). Chinese influence on the western examination system. *Harvard Journal of Asiatic Studies*, *7*(4), 267-312. Retrieved from http://www.jstor.org/stable/2717830

Vairinhos, V. M. (2003). *Desarollo de un sistema de numería de datos basado en los metodos biplot*. Tesis Doctoral. Universidad de Salamanca.

Vairinhos, V. M., & Galindo, M. P. (2012). Biplots cilíndricos. *XIX Jornadas de Classificação e Análise de Dados, Instituto Politécnico de Tomar*. Livro de Resumos, 132-137.

Vygotsky, L. S. (1978). *Mind in society. The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Vygotsky, L. S. (1994). The problem of environment. In Rene van der Veer & J. Valsiner (eds.), *The Vygotsky reader*. Cambridge, England: Blackwell.

Vojac, Kline, Cope, McCarthy, & Kalantzis (2011). New spaces and old places: Na analysis of writing assessment software. *Computers and Compositions, 28*, 97-111.

Wainer, H. (1990). *Computerized adaptive testing*. Mahwah, NJ: Lawrence Erlbaum.

Wanters, K. , Desment, P., & Noortgate, W. Van (2011). Acquaring item difficulty estimated: a collaborative effort of data and judgement. *Proceddings os the 4th International Conference on Educational Data Mining*, July 6-8, 2011.

Williamson, D. M., Benmett, R. E., Bernstein, J., Foltz, P. W., Landauer, T. K., Rusin, D. P., Way, W. D., & Sweeney, K. (2010). Automated scoring for the assessment of common core standards. *Educational Testing Service*, *Person Education, The College Board*.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and uue of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2-13.

Winne, P.H., & Hadwin, A.F. (1998). *Studying as self-regulated learning*. In D.J. Hacker, J.E. Dunlosky, and A.C. Graesser (Eds.). Metacognition in Educational Theory and Practice, 277-304. Mahwah: Lawrence Erlbaum Associates.

Winne, P.H. (2001). *Self-regulated learning viewed from models of information processing*. In B.J. Zimmerman and D.H. Schunk (Eds.). Self-regulated learning and academic achievement: Theoretical perspectives (2nd Ed.), 153-189. Mahwah: Lawrence Erlbaum Associates.

Winne, P.H. (2011). *A cognitive and metacognitive analysis of self-regulated learning*. In B.J. Zimmerman and D.H. Schunk (Eds.). Handbook of Self-Regulation of Learning and Performance, 15-32. New York: Routledge.

Yang, Y., Buckendahl, W., Juszkiewicz, P. J., & Bhola, D. (2012). A review of strategies for validating computer – automated scoring. *Applied Measurement in Education, 15*(4), 391 – 412.

Yin, J. (2012). Using tree diagrams as an assessment tool in statistics education. *Educational Assessment, 17*(1), 22-49.

Young, G., & Householder, A. S. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika, 3*(1), 19-22.

Zhang, M. (2013). Contrasting automated and human scoring of essays. *Educational Testing Service*. Retrieved from: www.ets.org.

Zeller, R. A. (1998). *Validity*. Keeves, John P., 322-330.

Zengina, K., Esgia, N., Erginerb, E., & Aksoya, M. E. (2011). A sample study on applying data mining research techniques in educational science: developing a more meaning of data. *Procedia Social and Behavioral Sciences*, 15, 4028–4032. doi: 10.1016/j.sbspro.2011.04.408

Índice das Figuras

Índice das Tabelas

ANEXO A

MANUAL DO PROGRAMA DE

ANÁLISE ESTATÍSTICA DE TEXTOS (PAET)

ÍNDICE

1. Introdução e Funções Principais do PAET.

Este programa foi desenvolvido para apoiar, ilustrar e experimentar as ideias principais da tese de Luís Agonia Pereira (2013) no departamento de Ciências da Educação da Universidade Nova de Lisboa.

O programa foi concebido para implementar as ideias desenvolvidas no **Capítulo II** (Metodologias) da tese em questão.

As funções principais são:

1. Análise estatística de conjuntos de textos previamente escolhidos pelo utilizador;

2. Classificação automática de textos que sejam respostas a questões abertas relativas a temas de ensino caracterizados por um certo espaço semântico (ES) usando a metodologia do AS.

Em relação às análises estatísticas, as funções básicas dessa análise são:

1.1. Construção das tabelas de frequências dos textos selecionados (Representação Vetorial dos Textos);

1.2. Construção de biplots usando a SVD da tabela de frequências;

1.3. Realização de análises cluster sobre os resultados da SVD da tabela de frequências.

O programa assume que os textos (manuais de ensino, textos de respostas a questões abertas) estão organizados em tabelas do sistema de gestão de bases de dados ACESS integrante do *Microsoft Office*, segundo a estrutura que está definida no número 7 deste manual.

Além das tabelas com os textos deve existir ainda uma tabela – exterior ao ACESS – contendo a lista de palavras funcionais.

O programa é invocado a partir do sistema *Windows* através do ícone que se apresenta na **figura 1.1.** e que invoca a janela principal da **figura 1.2.**

Figura 1.1. Ícone de invocação do PAET.

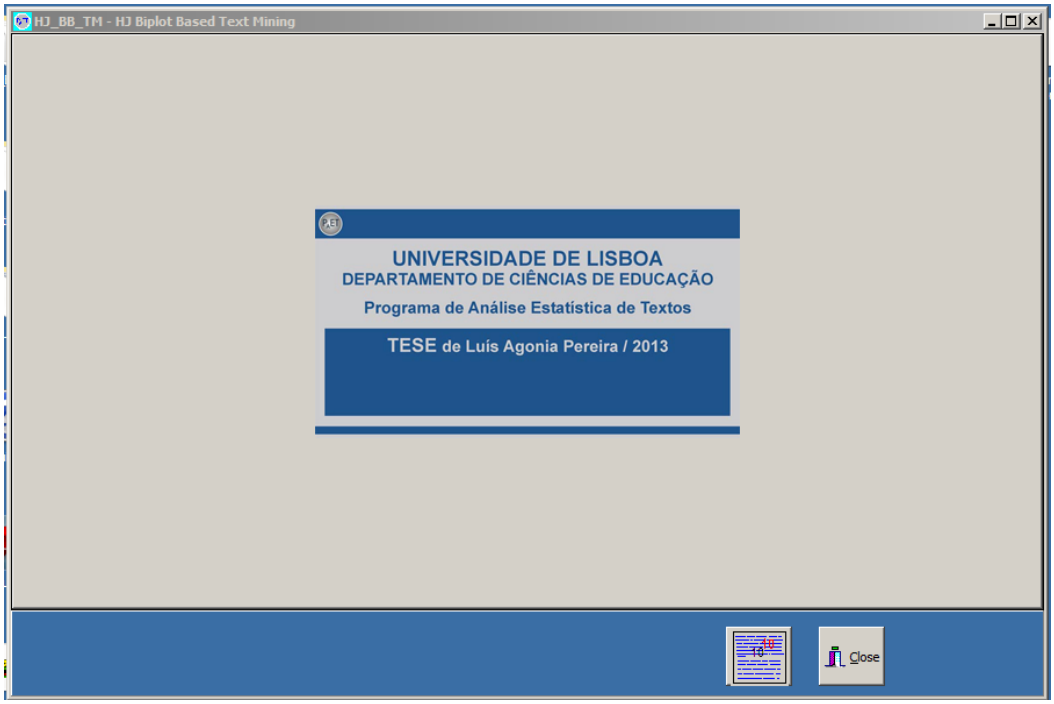


Figura 1.2. Janela principal do programa PAET.

Pressionando a parte central desta janela é apresentada uma janela de leitura do nome do ficheiro "gramática" onde estão armazenadas as palavras funcionais que não devem ser consideradas na construção da tabela de frequências, sendo apresentada a janela seguinte – ver **figura 1.3.**

2. Janela de Leitura e Análise de Textos.

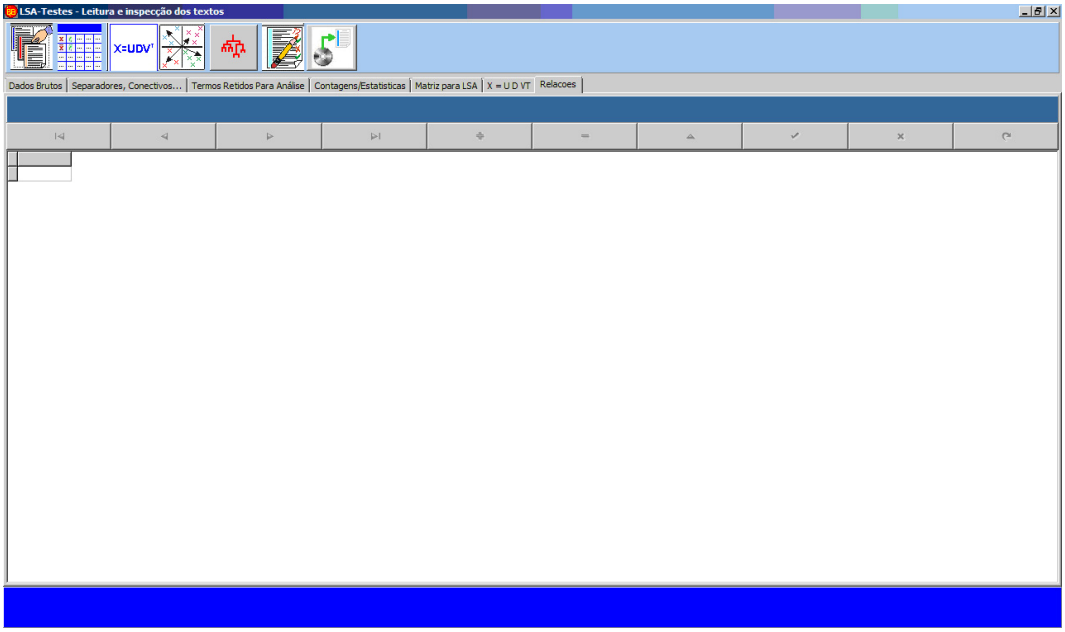


Figura 2.1. Janela se seleção e análise de textos.

Chega-se a esta janela depois de ter pressionado o ecrã de entrada (**figura 1.2.**) e indicada a posição do ficheiro designado "gramática" que guarda as palavras funcionais.

FUNÇÕES – Escolher os textos a analisar, identificar todas as palavras em todos os textos escolhidos e suas frequências (excluindo as palavras funcionais contidas no ficheiro "gramática"), construir a representação vetorial dos textos (palavras × frequências), realizar a decomposição em valores e vetores singulares e construir biplots com as duas primeiras componentes da decomposição anterior.

FUNÇÃO DOS BOTÕES

Botão de escolha da base de dados (entensão ".mdb") e, dentro da base de dados, escolha da tabela contendo os textos que interessa analisar. Veja **figuras 2.2.**, **2.3.**, **2.4.**

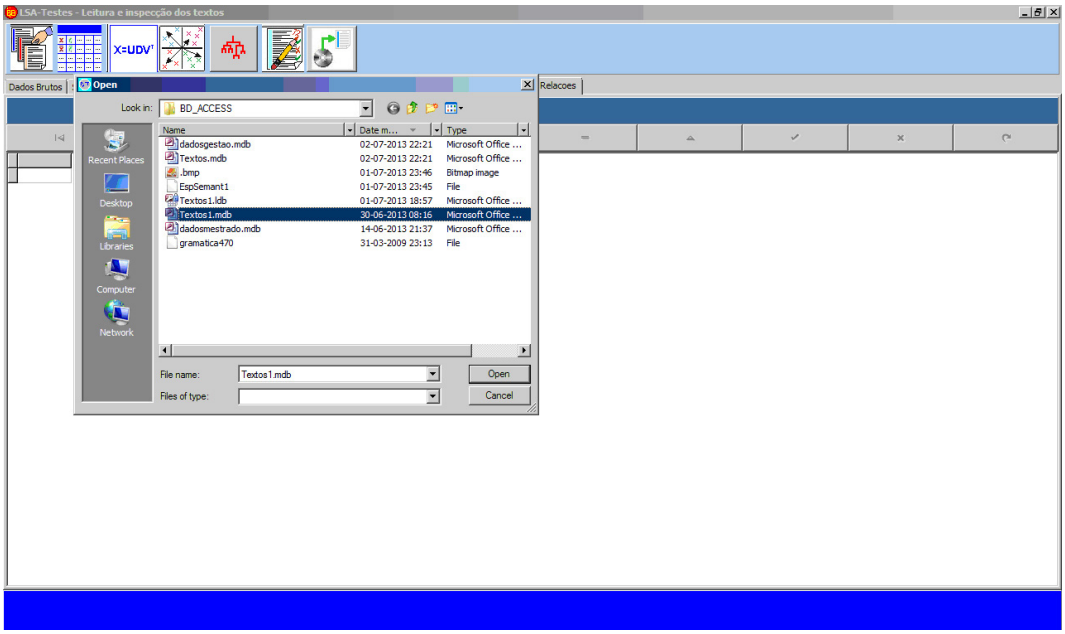


Figura 2.2. Escolha da base de dados contendo os textos a escolher (extensão ".mdb").

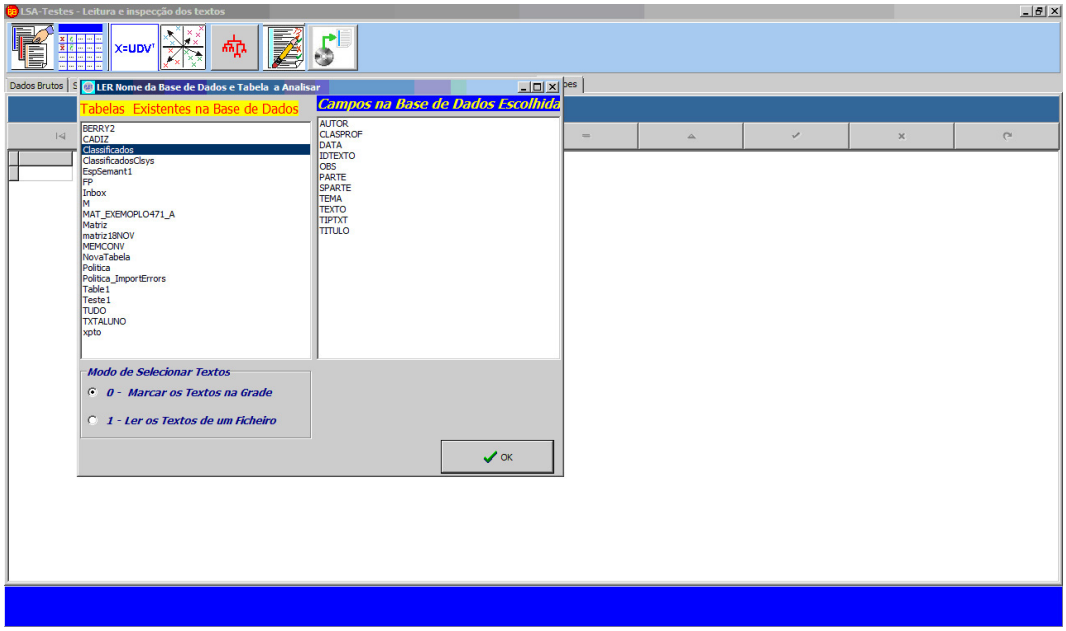


Figura 2.3. Dentro da base de dados escolha da tabela contendo os textos (classificador, neste caso).

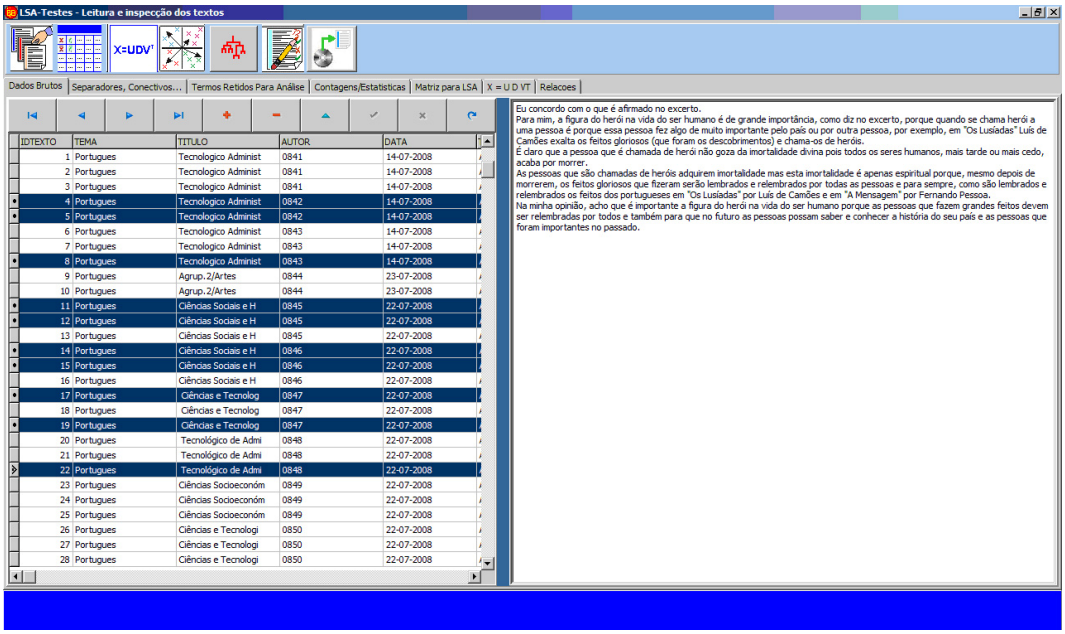
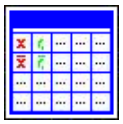
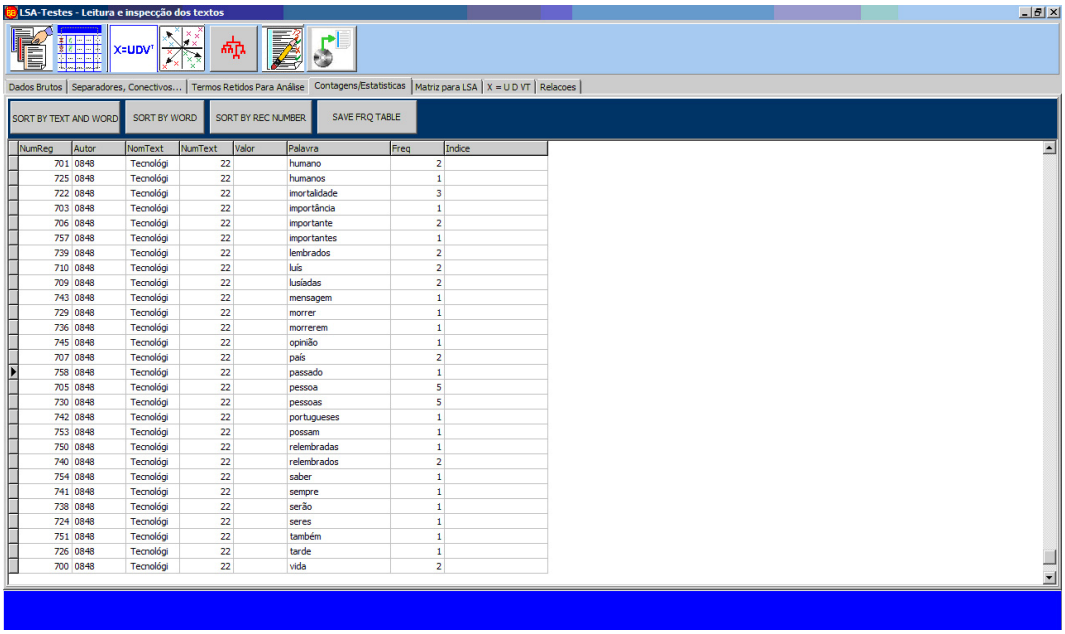


Figura 2.4. Escolhida a tabela com os textos, estes aparecem identificados à esquerda e cada um dos textos apontados aparece na janela da direita.



Botão de seleção de textos a analisar, dentro da tabela escolhida. Os textos podem ser escolhidos usando as teclas Shift ↓ , ↑ ou usando a tecla Ctrl e apontando os textos desejados. Caso não seja selecionado nenhum texto, quando se prime este botão aparece uma janela de diálogo de leitura do nome do ficheiro de texto onde estão os identificadores dos registos com os ficheiros selecionados. Com os textos selecionados calcula a tabela de frequências e apresenta-as. Veja a **figura 2.5.**

Botão que prepara a análise: obtém a representação vetorial dos textos (palavras × textos) e calcula a decomposição em valores e vetores singulares, apresentando o resultado dessa decomposição. Apresenta também a matriz que é submetida a análise (Matriz para LSA) e que pode ser exportada para um ficheiro de texto e lida por ACESS, SPSS ou EXCEL. Veja **figuras 2.6.**, **2.7.**, **2.8.**



NumReg	Autor	NomText	NumText	Valor	Palavra	Freq	Indice
701	0848	Tecnológi	22		humano	2	
725	0848	Tecnológi	22		humanos	1	
722	0848	Tecnológi	22		imortalidade	3	
703	0848	Tecnológi	22		importância	1	
706	0848	Tecnológi	22		importante	2	
757	0848	Tecnológi	22		importantes	1	
739	0848	Tecnológi	22		lembrados	2	
710	0848	Tecnológi	22		luís	2	
709	0848	Tecnológi	22		lusíadas	2	
743	0848	Tecnológi	22		mensagem	1	
729	0848	Tecnológi	22		morrer	1	
736	0848	Tecnológi	22		morrerem	1	
745	0848	Tecnológi	22		opinião	1	
707	0848	Tecnológi	22		país	2	
758	0848	Tecnológi	22		passado	1	
705	0848	Tecnológi	22		pessoa	5	
730	0848	Tecnológi	22		pessoas	5	
742	0848	Tecnológi	22		portugueses	1	
753	0848	Tecnológi	22		possam	1	
750	0848	Tecnológi	22		relembradas	1	
740	0848	Tecnológi	22		relembrados	2	
754	0848	Tecnológi	22		saber	1	
741	0848	Tecnológi	22		sempre	1	
738	0848	Tecnológi	22		serão	1	
724	0848	Tecnológi	22		seres	1	
751	0848	Tecnológi	22		também	1	
726	0848	Tecnológi	22		tarde	1	
700	0848	Tecnológi	22		vida	2	

Figura 2.5. Frequências das palavras nos textos selecionados.

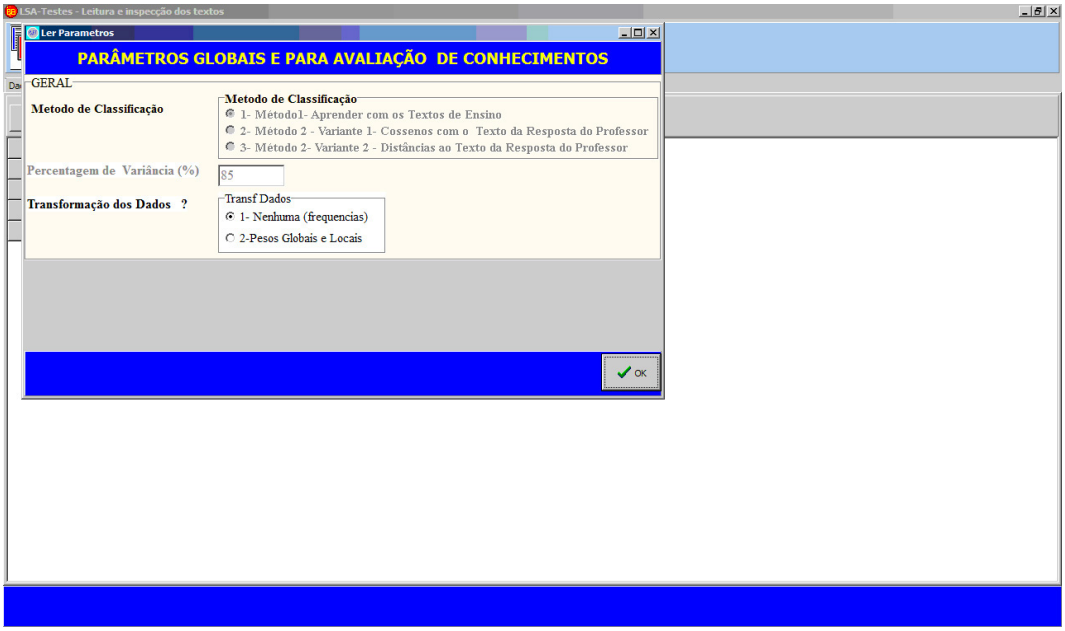
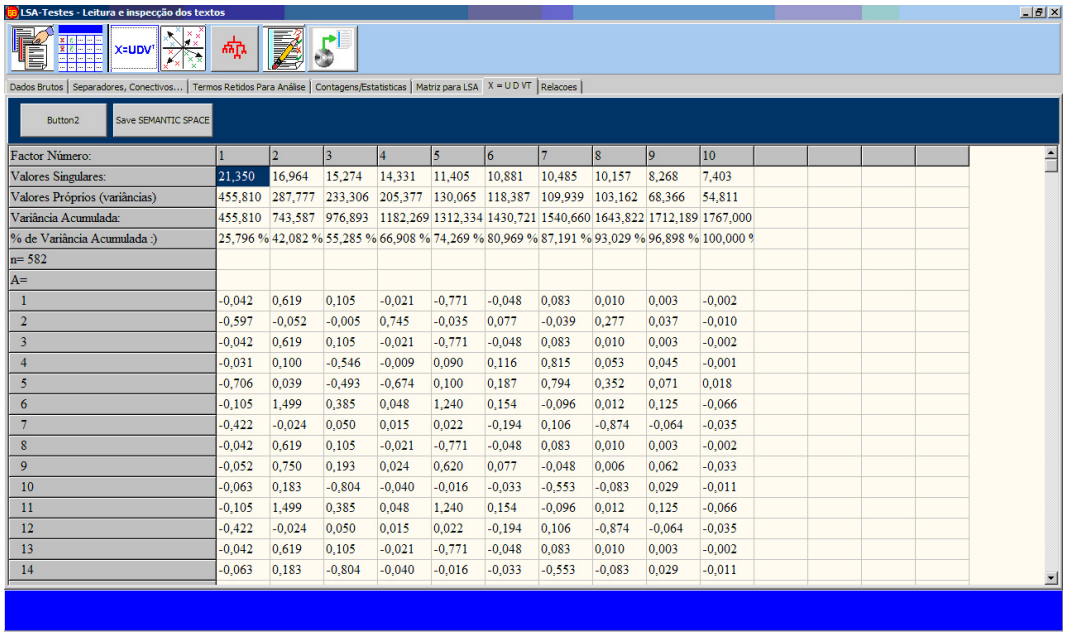
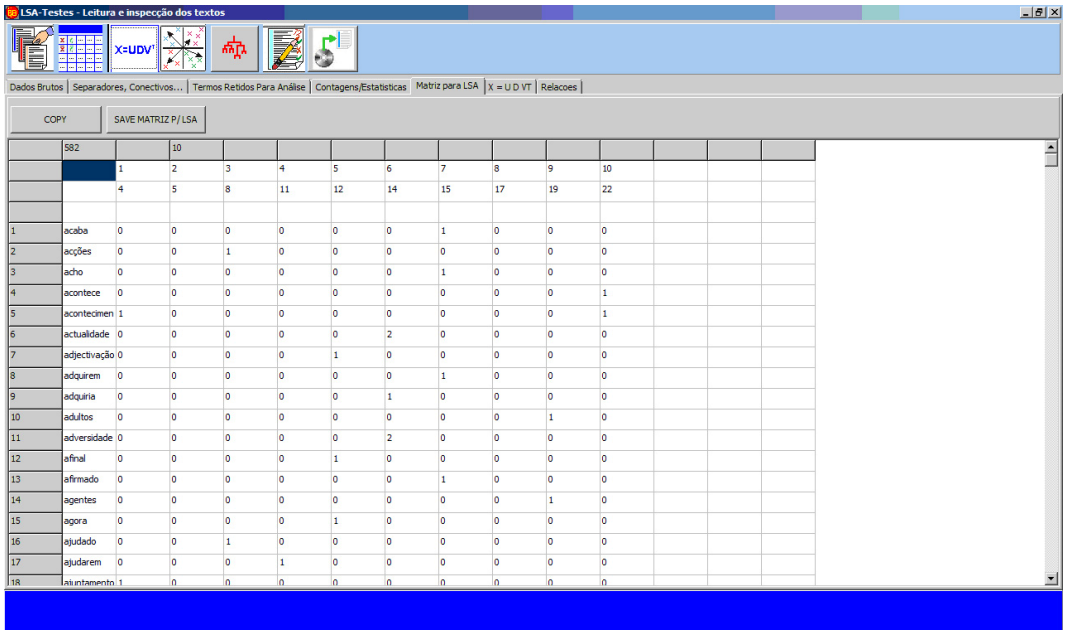


Figura 2.6. Escolha da opção de transformação da matriz de frequências antes da decomposição SVD.



Factor Número:	1	2	3	4	5	6	7	8	9	10
Valores Singulares:	21,350	16,964	15,274	14,331	11,405	10,881	10,485	10,157	8,268	7,403
Valores Próprios (variâncias)	455,810	287,777	233,306	205,377	130,065	118,387	109,939	103,162	68,366	54,811
Variância Acumulada:	455,810	743,587	976,893	1182,269	1312,334	1430,721	1540,660	1643,822	1712,189	1767,000
% de Variância Acumulada :)	25,796 %	42,082 %	55,285 %	66,908 %	74,269 %	80,969 %	87,191 %	93,029 %	96,898 %	100,000 %
n= 582										
A=										
1	-0,042	0,619	0,105	-0,021	-0,771	-0,048	0,083	0,010	0,003	-0,002
2	-0,597	-0,052	-0,005	0,745	-0,035	0,077	-0,039	0,277	0,037	-0,010
3	-0,042	0,619	0,105	-0,021	-0,771	-0,048	0,083	0,010	0,003	-0,002
4	-0,031	0,100	-0,546	-0,009	0,090	0,116	0,815	0,053	0,045	-0,001
5	-0,706	0,039	-0,493	-0,674	0,100	0,187	0,794	0,352	0,071	0,018
6	-0,105	1,499	0,385	0,048	1,240	0,154	-0,096	0,012	0,125	-0,066
7	-0,422	-0,024	0,050	0,015	0,022	-0,194	0,106	-0,874	-0,064	-0,035
8	-0,042	0,619	0,105	-0,021	-0,771	-0,048	0,083	0,010	0,003	-0,002
9	-0,052	0,750	0,193	0,024	0,620	0,077	-0,048	0,006	0,062	-0,033
10	-0,063	0,183	-0,804	-0,040	-0,016	-0,033	-0,553	-0,083	0,029	-0,011
11	-0,105	1,499	0,385	0,048	1,240	0,154	-0,096	0,012	0,125	-0,066
12	-0,422	-0,024	0,050	0,015	0,022	-0,194	0,106	-0,874	-0,064	-0,035
13	-0,042	0,619	0,105	-0,021	-0,771	-0,048	0,083	0,010	0,003	-0,002
14	-0,063	0,183	-0,804	-0,040	-0,016	-0,033	-0,553	-0,083	0,029	-0,011

Figura 2.7. Resultado da SVD da matriz de frequências (valores singulares, variância acumulada 4 vetores singulares à esquerda e à direita).

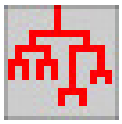


	582		10											
		1	2	3	4	5	6	7	8	9	10			
		4	5	8	11	12	14	15	17	19	22			
1	acaba	0	0	0	0	0	0	1	0	0	0			
2	acções	0	0	1	0	0	0	0	0	0	0			
3	acho	0	0	0	0	0	0	1	0	0	0			
4	acontece	0	0	0	0	0	0	0	0	0	1			
5	acontecimen	1	0	0	0	0	0	0	0	0	1			
6	actualidade	0	0	0	0	0	2	0	0	0	0			
7	adjectivação	0	0	0	0	1	0	0	0	0	0			
8	adquirem	0	0	0	0	0	0	1	0	0	0			
9	adquiria	0	0	0	0	0	1	0	0	0	0			
10	adultos	0	0	0	0	0	0	0	0	1	0			
11	adversidade	0	0	0	0	0	2	0	0	0	0			
12	afinal	0	0	0	0	1	0	0	0	0	0			
13	afirmado	0	0	0	0	0	0	1	0	0	0			
14	agentes	0	0	0	0	0	0	0	0	1	0			
15	agora	0	0	0	0	1	0	0	0	0	0			
16	ajudado	0	0	1	0	0	0	0	0	0	0			
17	ajudarem	0	0	0	1	0	0	0	0	0	0			
18	ajuntamento	1	0	0	0	0	0	0	0	0	0			

Figura 2.8. Matriz com as frequências × textos antes da decomposição SVD.

FUNÇÃO DOS BOTÕES

Apresenta o biplot construído com os eixos horizontal (eixo 1) e vertical (eixo 2) que forem indicados. Ver **figura 4.1.** com a janela de apresentação e estudo de biplots. Ver à frente **janela 3**.



Análise de clusters. Este botão realiza uma análise de cluster dos textos usando os resultados da decomposição em valores e vetores singulares da matriz de frequências. Ver **janela 4**.

Invoca o SAAT integrado neste sistema. Classificar automaticamente textos de resposta a questões abertas que tenham sido selecionados usando como referência o espaço semântico corrente. Este espaço semântico resulta da decomposição em valores e vetores singulares corrente, com a qual são produzidos os biplots e as análises de clusters correntes. Veja **janela 5**.

Guarda num ficheiro de texto os identificadores, na tabela em uso, dos textos que foram selecionados. Embora não seja obrigatório sugere-se que quando estes textos foram usados para construir um espaço semântico a usar numa futura sessão, se comece o nome do ficheiro por "ES", como no exemplo seguinte: ES_ xxx.txt.

Isto facilita muito a busca dos ficheiros deste tipo. Veja **janela 5**.

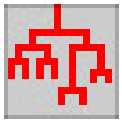
Em síntese, a sequência habitual para utilização da **janela 2** é:

Escolher a base de dados e a tabela, selecionando textos.

Examinar os textos selecionados ou lidos de um ficheiro quando nada foi selecionado.

Decompor a matriz em valores e vetores singulares.

Criar e estudar biplots com os resultados da decomposição.



Análise de clusters usando as coordenadas da SVD.

Realizar uma classificação automática, ascendente hierárquica (aglomerativa) de textos a classificar, usando o ES corrente.

Guardar num ficheiro de texto, os identificadores dos textos que foram selecionados para análise e que podem vir a ser usados numa futura sessão.

3. Construção e Estudo de Biplots.

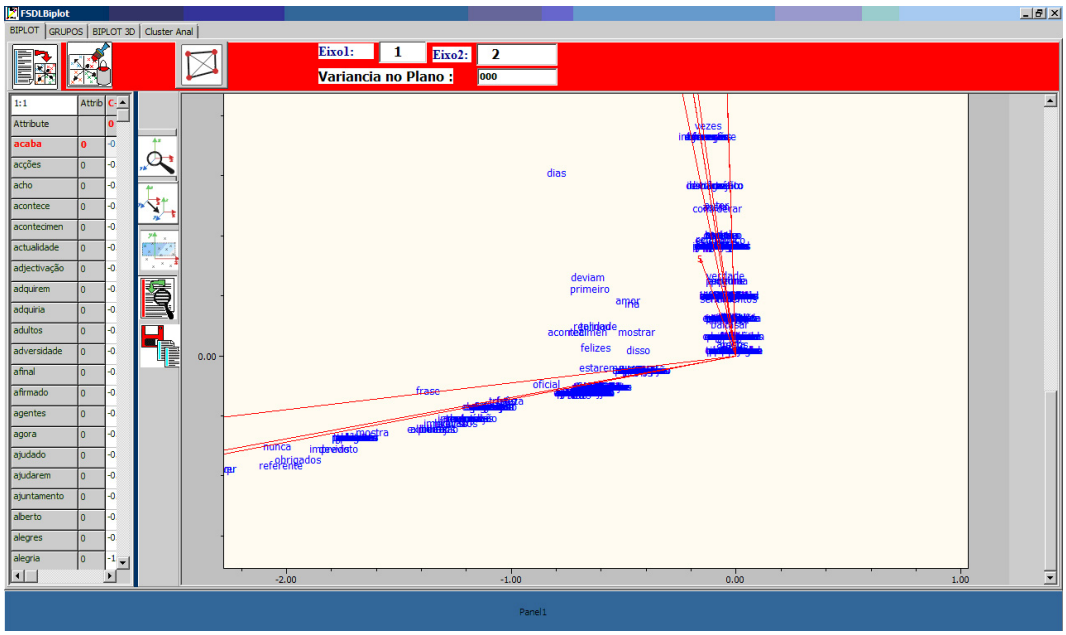


Figura 3.1. Janela para construção e estudo de biplots.

Chega-se a esta janela pressionando na **janela 1** o botão  .

FUNÇÕES – As funções principais a realizar, a partir desta janela, são as de preparar a estrutura de dados usada para a representação gráfica dos biplots e apresentar o

biplot do plano fatorial que for escolhido (escrevendo nas janelas do topo o número do eixo horizontal (eixo 1) e vertical (eixo 2) a usar nesse gráfico.

As funções dos botões que integram esta janela (veja **figura 3.1.**, segunda linha da esquerda para a direita):

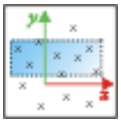
Este botão organiza os dados necessários à construção dos gráficos dos biplots. O resultado aparece na grade do lado esquerdo. Para cada palavra aparece o identificador da palavra seguida de um vetor de números que são as coordenadas no espaço semântico dessa palavra.

Terminadas as palavras seguem-se os textos tendo cada texto o vetor de coordenada correspondente ao seu lado direito. O valor da coluna "Atributo" é "0" para as palavras e "1" para os textos.

Botão de pintura do biplot escolhido. Depois de ter acionado o botão anterior, este apresenta o gráfico do biplot cujos eixos estão em eixo 1 e eixo 2. Ver **figura 3.1.**, um exemplo.

Este botão permite amplificar a escala. Depois de ter pressionado este botão, marcar com o rato um retângulo sobre o gráfico: "click, arrastar, soltar" o botão esquerdo do rato. O resultado é que a parte marcada do gráfico passa a ocupar agora todo o espaço visível da janela gráfica.

Este botão permite realizar translações do gráfico. Depois de o pressionar, fazer "click" com o apontador e, mantendo o botão esquerdo pressionado, arrastar todo o gráfico.



Este botão permite identificar todos os objetos (textos e palavras) que se encontram dentro de uma janela que foi marcada, passando a lista respetiva para uma grade (designada **Grupos**). Veja exemplo na **figura 3.2.** A sequência é: premir o botão esquerdo do rato, fazer "click" no canto superior esquerdo da janela que interessa, arrastar, largar. O resultado é uma lista como a que se apresenta na **figura 3.2.**

Apresenta a última imagem, antes da última transformação.

Guarda no ficheiro que for designado – com formato ".BMP" – e passa para o *clipboard* a imagem do biplot, de onde pode ser copiada por "paste" para outra aplicação qualquer como o Word, Excel, etc.

Estudar relações entre palavras e textos.

Sequência típica.

Carregar os dados do biplot.

Eixo 1 e Eixo 2 – escolher os eixos do biplot a representar. Exemplo: (15, 18).

Apresentar o gráfico do biplot definido pelos respetivos eixos.

Usar os botões de expansão, translação, etc., para estudar o gráfico.

Inspecionar uma certa zona do biplot e ver que textos e palavras a integram.

Estudar relações entre palavras e textos.

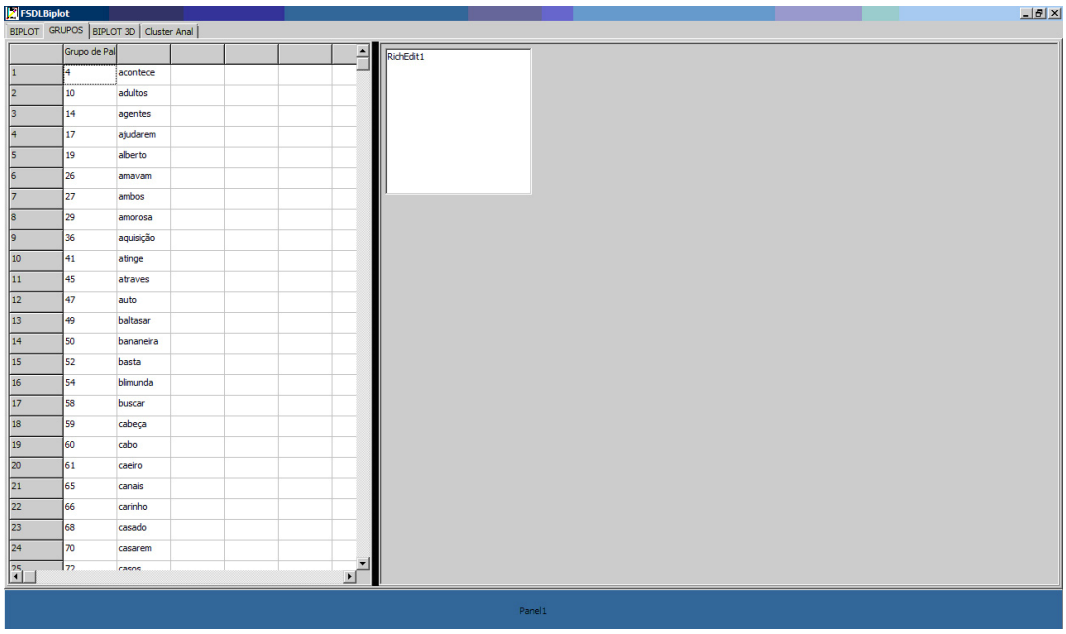


Figura 3.2. Lista de palavras e textos que estão numa janela marcada sobre o gráfico.

4. Janela para Relacionar Palavras e Textos.

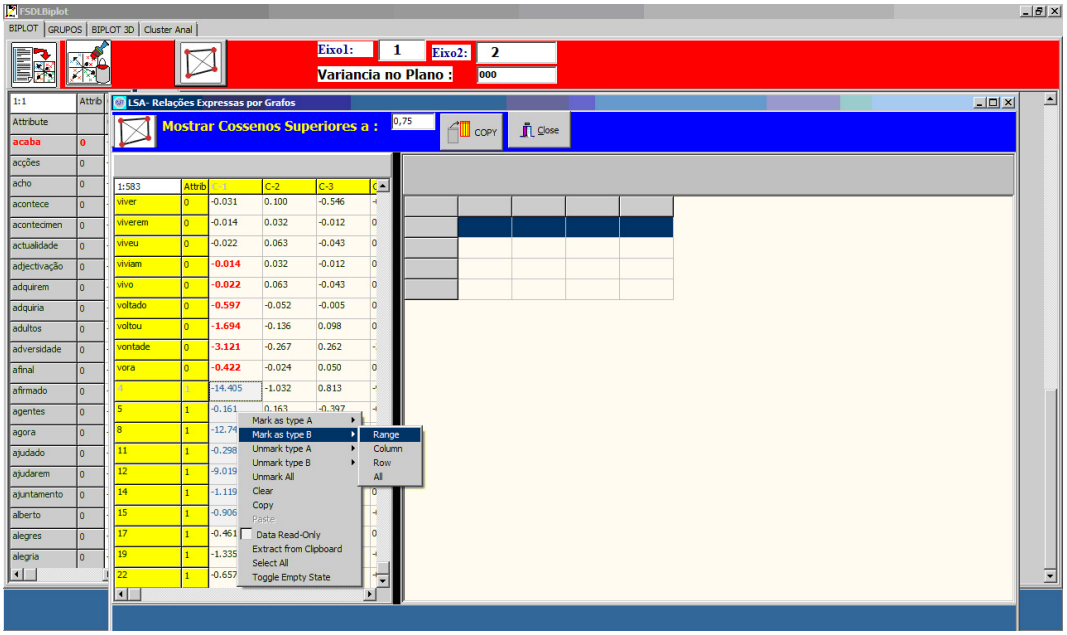


Figura 4.1. Janela para estudo de relações entre palavras e textos.

Usando a coluna 3 da grade do lado esquerdo (1ª coluna das coordenadas dos textos e das palavras), podem marcar-se duas sequências quaisquer: (palavras, palavras), (palavras, textos), (textos, textos), usando um menu flutuante a que se tem acesso pressionando o botão direito do rato. Ver **figura 4.1.** Depois de marcar uma sequência, indicar se é A ou B.

Feito isto pressionar o botão do canto superior esquerdo  . Na grade do lado direito aparecem os cossenos de valor acima do limiar que estiver inscrito na caixa à direita da frase: "*Mostrar Cossenos Superiores a*:". O valor por omissão é 0.75, mas pode ser indicado qualquer valor entre "-1" e "+1". Deste modo, é possível saber que objetos estão mais relacionados entre si a um nível de semelhança superior ao limiar indicado.

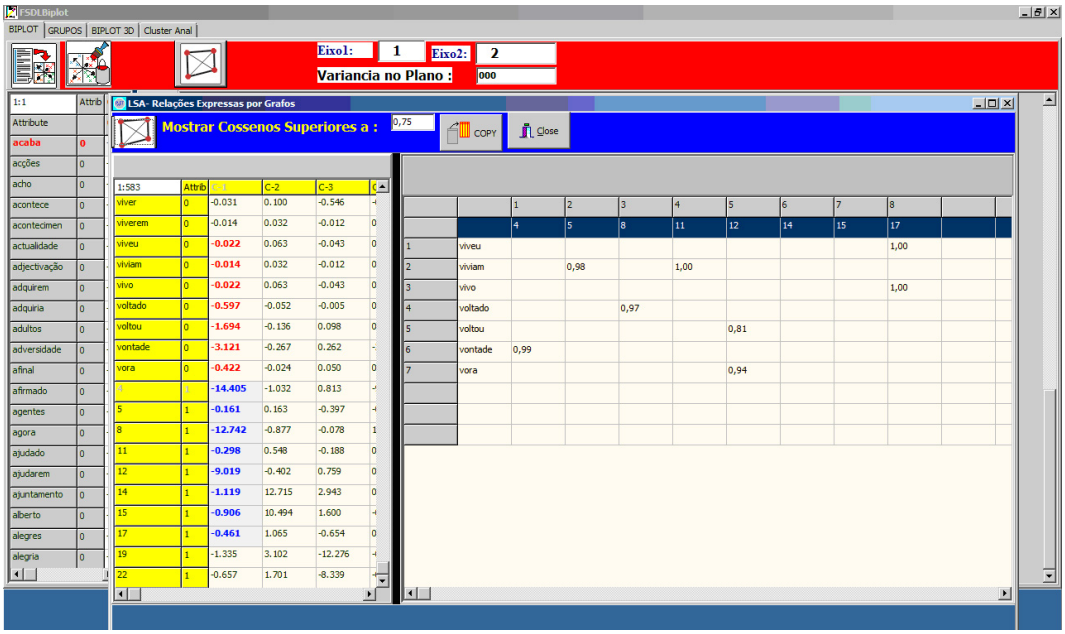


Figura 4.2. Tendo marcado na grade do lado esquerdo (coluna 1) os textos e as palavras a relacionar, a grade da direita apresenta os cossenos dos respetivos ângulos.

5. Janela para Análises de Clusters.

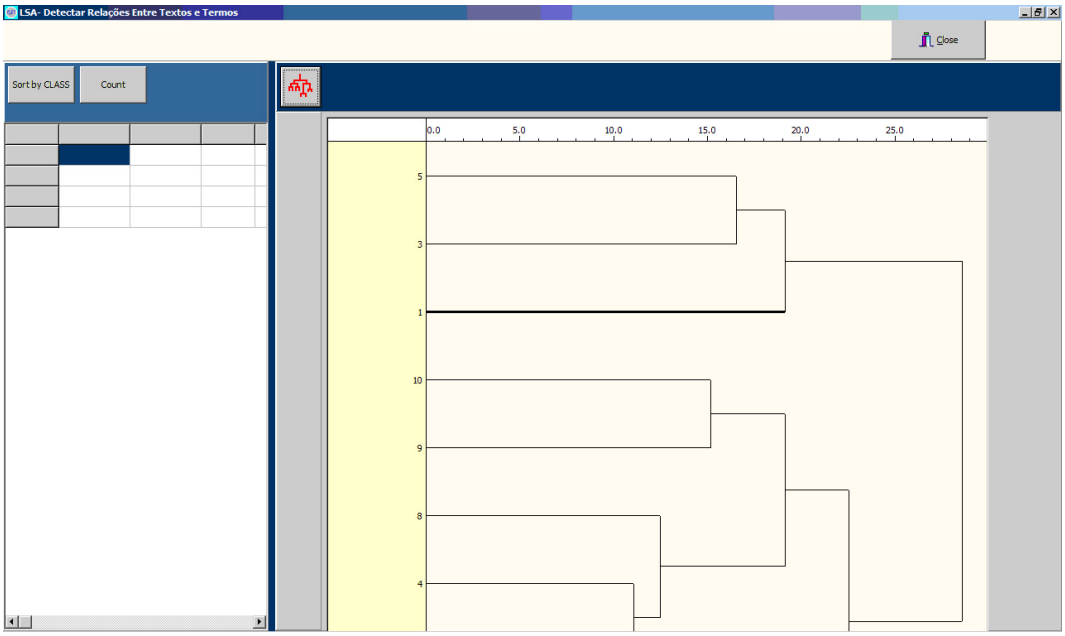
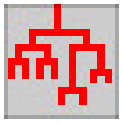


Figura 5.1. Janela de Análises Clusters.

Chega-se a esta janela através do botão  .

FUNÇÕES – Permite realizar classificação automática aglomerativa (ascendente hierárquica) de textos representados por coordenadas obtidas em resultado da decomposição em valores e vetores singulares da matriz de frequências.

A função dos botões associada a esta janela é:



Ao premir este botão – caso já tenha sido realizada a decomposição em valores singulares da matriz de frequências – aparece na janela gráfica do lado direito (veja **figura 5.1.**) uma árvore – deitada – que representa uma classificação ascendente hierárquica dos textos que estiverem selecionados.

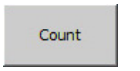
A escala indicada é o índice de dissemelhança. Os textos – identificados por números inteiros sequenciais dentro da seleção de textos – estão ao nível "0" e vão sendo agregados em classes cada vez menos homogéneas até que todos estejam agregados – neste caso, ao nível 28 numa única classe.

Se clicar no botão direito do rato aparece um menu flutuante que permite, entre outras funções – refazer a árvore, translação (*pam*), *zoom*, copiar para o clipboard e guardar a imagem – e cortar (*cut*) a árvore ao nível em que posicionarmos o rato. Veja a **figura 5.2.**

Quando se corta a árvore ao nível correspondente à posição do rato – veja **figura 5.3.** – o conteúdo de cada um dos ramos que existem a esse nível é mostrado na grade do lado esquerdo. Nessa grade aparece não só a identificação sequencial dos textos na seleção efetuada (1, 2, 3, …) como o identificador do texto na tabela da base de dados, a classe da arvore a que pertence o texto e a classificação atribuída pelo professor a esse texto – no caso de respostas a questões de resposta aberta.

Este botão – canto superior esquerdo da figura – não só ordena os textos pela classe resultante da análise de clusters, juntando os textos da mesma classe – como exporta estes resultados para um ficheiro de texto a designar pelo utilizador. Veja **figura 5.4.**. este ficheiro de texto pode ser importado por

EXCEL, SPSS, ACCESS para ser estudado e usado por qualquer desses sistemas.



Este botão não está ainda implementado na versão atual.

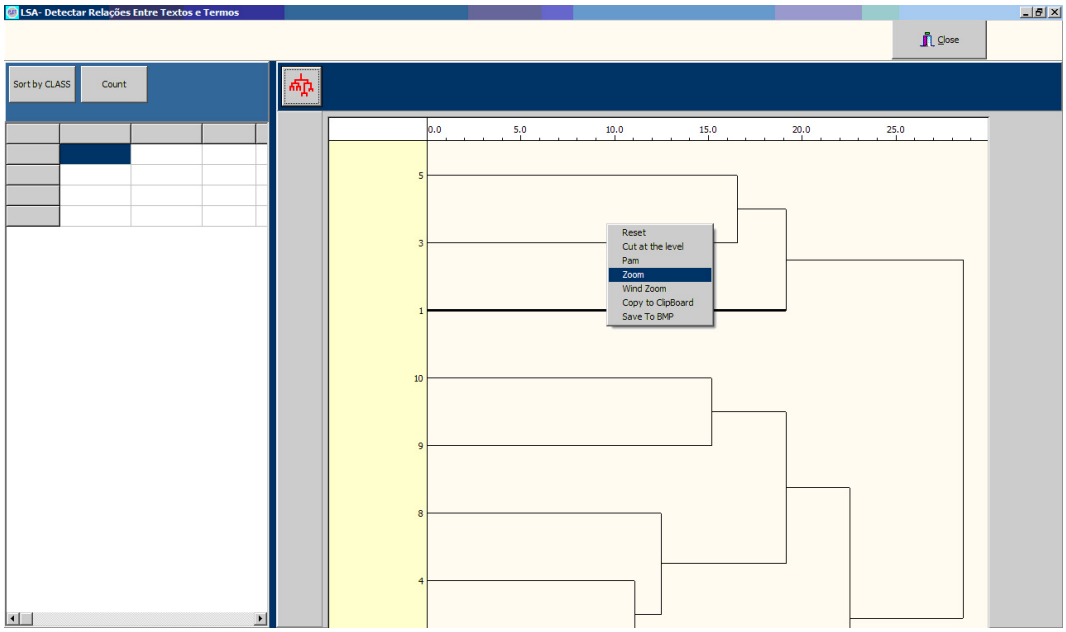


Figura 5.2. Com o botão direito do rato tem acesso a um menu flutuante que permite inspecionar a árvore e cortá-la a um nível arbitrário.

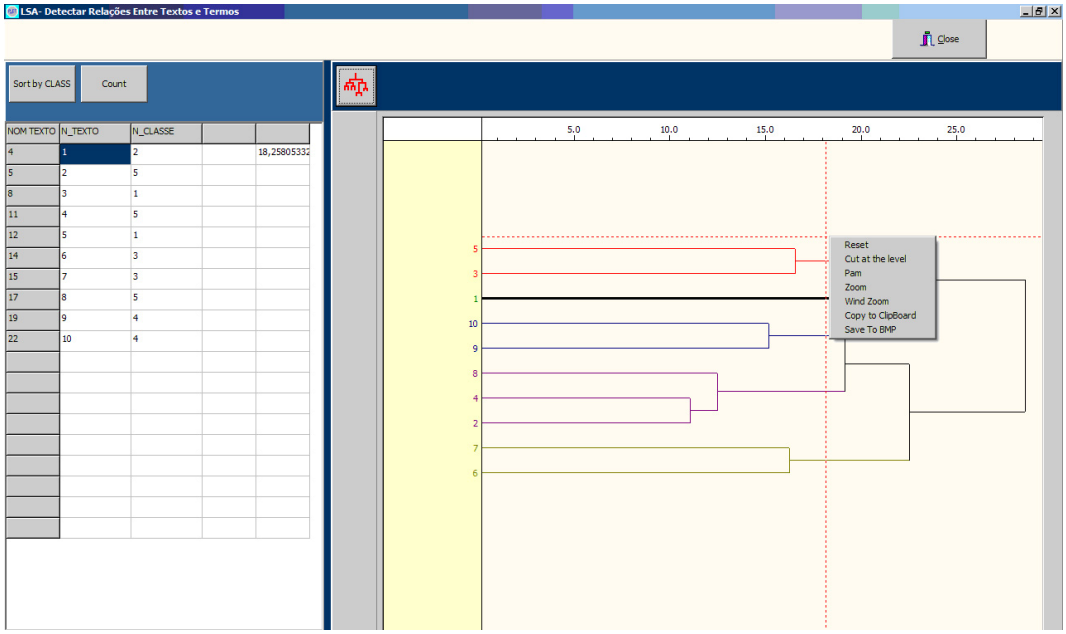


Figura 5.3. Corte da árvore ao nível 18 de dissemelhança. A árvore fica cortada em cinco ramos cujo conteúdo aparece na grade do lado esquerdo.

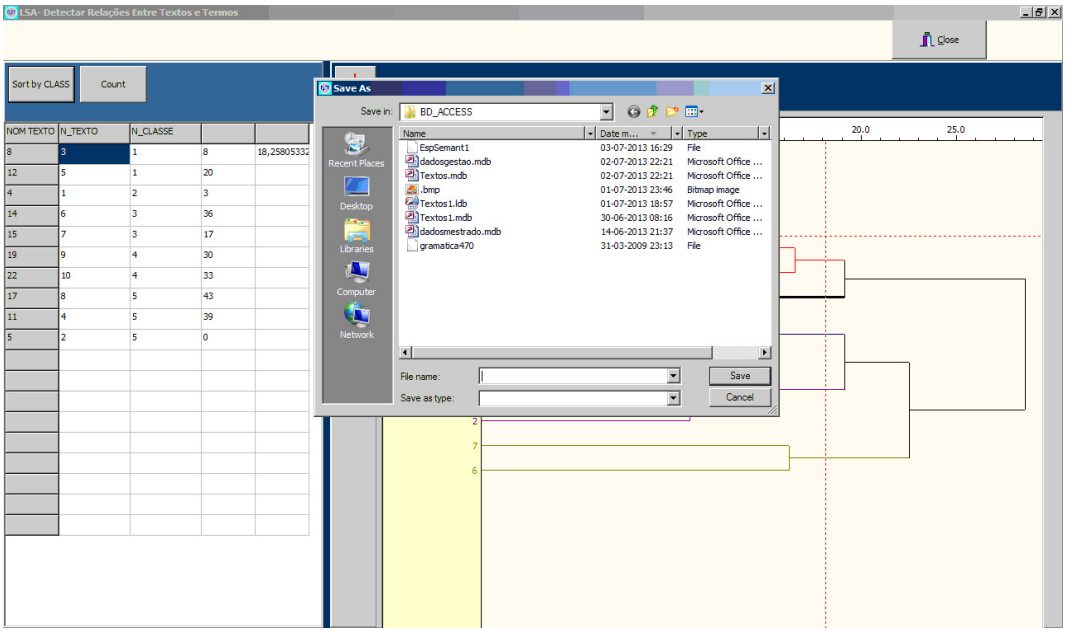


Figura 5.4. Guardar num ficheiro de texto do resultado da análise de clusters.

6. Janela de Classificação Automática de Textos de Resposta a Questões Abertas.

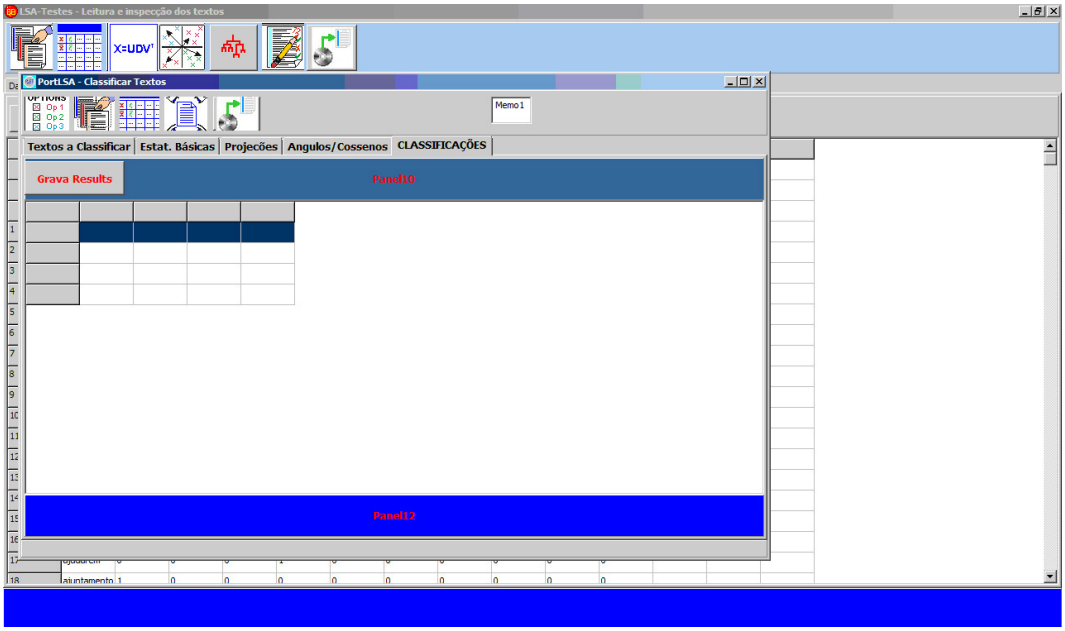


Figura 6.1. Janela para classificação de textos de resposta a questões abertas.

Esta janela é invocada pelo botão  na **janela 2**.

FUNÇÕES – Este botão invoca o sistema automático de avaliação dos textos (SAAT) produzidos por estudantes em resposta a questões abertas, posicionando os textos escolhidos no espaço semântico corrente.

As funções dos botões são as seguintes:

Permite que o utilizador fixe os valores dos parâmetros seguintes:

- **Método de Classificação** (na versão atual apenas está disponível o Método 1).
- **Percentagem de Variância**.

  Permite escolher a percentagem de variância captada pelo espaço semântico a partir da tabela de frequências.

  Um valor muito elevado é contraproducente (introduz ruído).

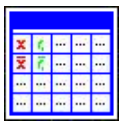
  Um valor demasiado baixo não tem dimensionalidade suficiente.

  O valor por omissão é 85% que corresponde a uma dimensionalidade *d* próxima de (1/2)/*p* em que *p* é o número de textos.
- **Percentagem da Amostra de Treino**.

  Da amostra de textos selecionados para classificação, uma certa parte é reservada para o treino (AT) do classificador e o resto para testar o classificador (T).
- **Número de vizinhos mis próximos**.

  É o número (*k*) de vizinhos no ES de um certo teste que vão ser usados para definir o resultado a atribuir a um certo texto a classificar (*k*= 1, 2, …, 8).

Escolha da base de dados e dentro desta a tabela onde se encontram os textos classificados por um professor a usar para amostra de treino (AT) e amostra de teste (T). Uma vez apresentados os textos, aqueles – distintos dos usados para definir o Espaço Semântico (ES) – são marcados com auxilio do rato (apontando os que interessam) e as teclas Shift ↑ (série contínua) ou Ctrl (marcar um a um). Veja **figura 6.2.**



Esta tecla lê os textos marcados e procede à contagem de palavras e suas frequências segundo um processo idêntico à função equivalente quando da seleção de textos para o ES. Ver **figura 6.3.**

No caso de não terem sido marcados quaisquer textos, o programa pede ao utilizador que diga qual o ficheiro de texto onde estão identificados os textos a usar – e que foram guardados no final de uma sessão anterior de classificação.

No final são apresentadas as frequências das palavras nestes textos para inspeção do utilizador. Veja **figura 6.4.**

Treino do classificador e sua utilização para classificar a amostra de teste (T). Tendo em conta os parâmetros escolhidos (percentagem de variância, percentagem da AT, número de vizinhos mais próximos) o SAAT vai determinar os cossenos dos ângulos entre cada texto a classificar na amostra T e cada um dos textos na AT. Em seguida, considerando os *k* mais próximos (maiores cossenos) obtém-se uma classificação Clsys. O resultado aparece na **figura 6.5.** em que as colunas Clsys e ClProf têm os resultados atribuídos pelo sistema e as classificações atribuídas pelo professor à amostra de treino (AT) correspondentes à escolha de parâmetros correntes.

A coluna "correlação" contém a correlação entre as notas atribuídas pela máquina (Clsys) e as notas atribuídas pelo professor à amostra de teste (T).

Este botão grava os resultados num ficheiro de texto que pode ser lido por EXCEL, SPSS para estudo estatístico aprofundado.

Permite gravar num ficheiro de texto os identificadores dos textos usados num certo estudo, para uso futuro.

Sugere-se que o nome destes ficheiros comece por "T_" afim de facilitar futuras consultas.

Exemplo: T_yyyy.txt

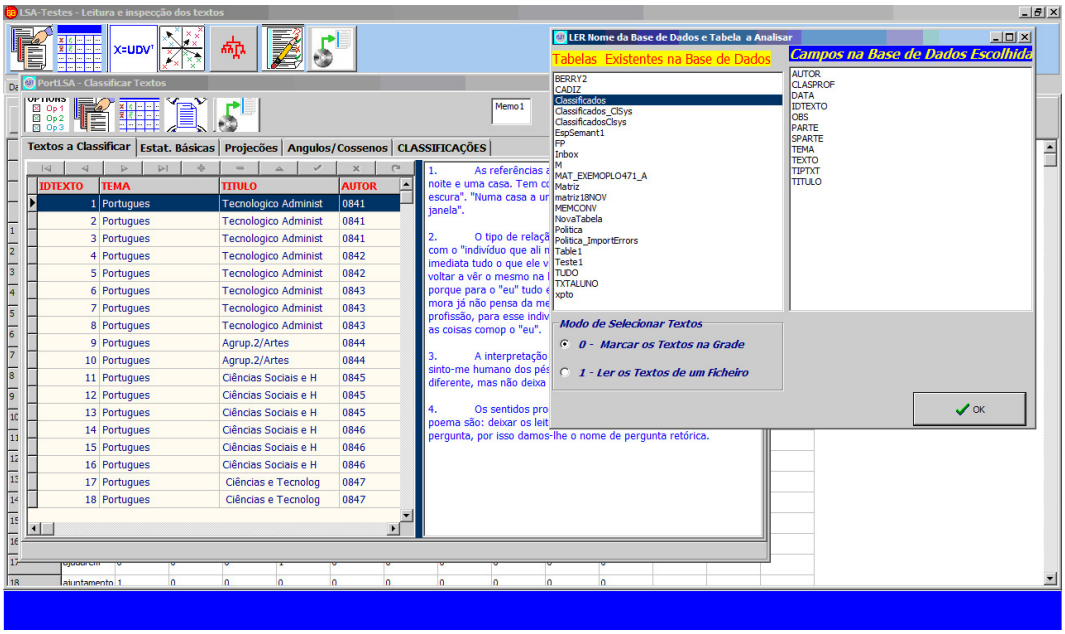


Figura 6.2. Apresentação das frequências dos textos escolhidos para a constituição da amostra de treino e de teste.

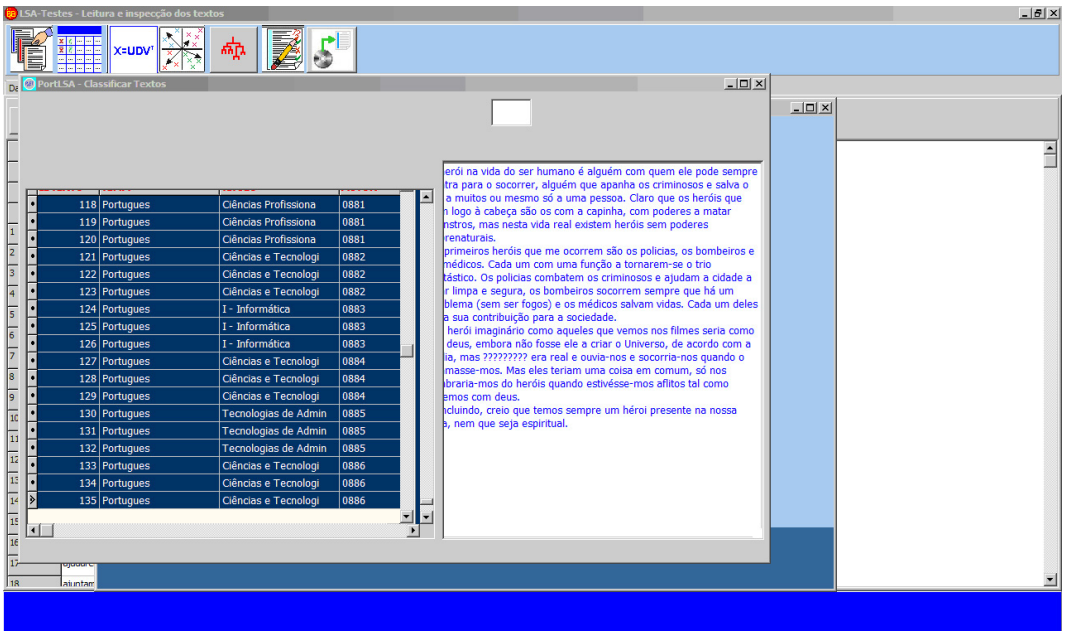


Figura 6.3. Seleção manual dos textos a usar na AT e na amostra T.

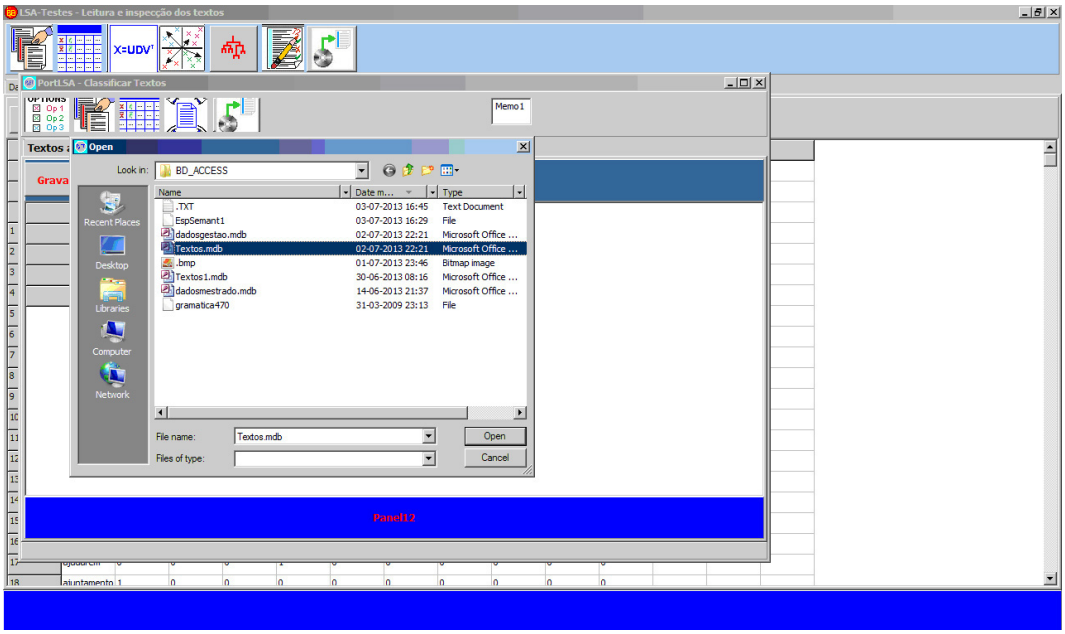
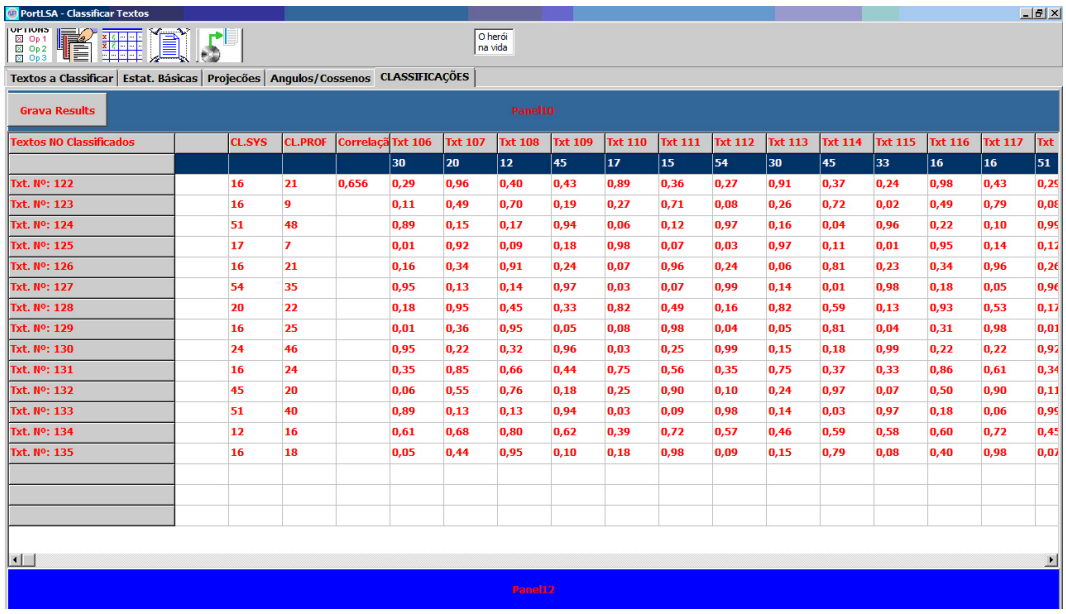


Figura 6.4. Indicação do ficheiro de texto onde estão os identificadores dos textos a usar no treino.



Textos NO Classificados		CL.SYS	CL.PROF	Correlaçã	Txt 106	Txt 107	Txt 108	Txt 109	Txt 110	Txt 111	Txt 112	Txt 113	Txt 114	Txt 115	Txt 116	Txt 117	Txt
					30	20	12	45	17	15	54	30	45	33	16	16	51
Txt. Nº: 122		16	21	0,656	0,29	0,96	0,40	0,43	0,89	0,36	0,27	0,91	0,37	0,24	0,98	0,43	0,25
Txt. Nº: 123		16	9		0,11	0,49	0,70	0,19	0,27	0,71	0,08	0,26	0,72	0,02	0,49	0,79	0,08
Txt. Nº: 124		51	48		0,89	0,15	0,17	0,94	0,06	0,12	0,97	0,16	0,04	0,96	0,22	0,10	0,95
Txt. Nº: 125		17	7		0,01	0,92	0,09	0,18	0,98	0,07	0,03	0,97	0,11	0,01	0,95	0,14	0,12
Txt. Nº: 126		16	21		0,16	0,34	0,91	0,24	0,07	0,96	0,24	0,06	0,81	0,23	0,34	0,96	0,26
Txt. Nº: 127		54	35		0,95	0,13	0,14	0,97	0,03	0,07	0,99	0,14	0,01	0,98	0,18	0,05	0,96
Txt. Nº: 128		20	22		0,18	0,95	0,45	0,33	0,82	0,49	0,16	0,82	0,59	0,13	0,93	0,53	0,17
Txt. Nº: 129		16	25		0,01	0,36	0,95	0,05	0,08	0,98	0,04	0,05	0,81	0,04	0,31	0,98	0,01
Txt. Nº: 130		24	46		0,95	0,22	0,32	0,96	0,03	0,25	0,99	0,15	0,18	0,99	0,22	0,22	0,92
Txt. Nº: 131		16	24		0,35	0,85	0,66	0,44	0,75	0,56	0,35	0,75	0,37	0,33	0,86	0,61	0,34
Txt. Nº: 132		45	20		0,06	0,55	0,76	0,18	0,25	0,90	0,10	0,24	0,97	0,07	0,50	0,90	0,11
Txt. Nº: 133		51	40		0,89	0,13	0,13	0,94	0,03	0,09	0,98	0,14	0,03	0,97	0,18	0,06	0,95
Txt. Nº: 134		12	16		0,61	0,68	0,80	0,62	0,39	0,72	0,57	0,46	0,59	0,58	0,60	0,72	0,45
Txt. Nº: 135		16	18		0,05	0,44	0,95	0,10	0,18	0,98	0,09	0,15	0,79	0,08	0,40	0,98	0,02

Figura 6.5. Resultados do classificador.

7. Estrutura da Tabela para Guardar Textos em Tabelas ACCESS.

A ordem dos campos é importante. Outros campos podem ser adicionados depois de ClasProf.

IDTEXTO – Number

TEMA – Text (20)

TITULO – Text (20)

AUTOR – Text (20)

DATA – Date/Time

TIPTEXT – Text (10)

PARTE – Text (10)

SPARTE – Text (10)

TEXTO – MEMO

CLASPROF – Number