

# **TIPOLOGIA TEXTUAL EAGLES\***

**John Sinclair**  
**Universidade de Birmingham**

## **Introdução**

### **Preâmbulo**

Este artigo reúne as recomendações para uma tipologia textual apresentadas ao Grupo de Trabalho da Comissão Europeia EAGLES (Expert Advisory Group for Language Standards), em Janeiro de 1996. Trata-se de um texto baseado no trabalho pioneiro do NERC (o projecto da UE Network of European Reference Corpora – cf. Calzolari *et al.* 1994), e foi revisto à luz de comentários solicitados a colegas do EAGLES. Oferece-se, agora, a um grupo mais lato de académicos, como um passo na direcção de uma maior harmonia dos esforços colectivos tendo em vista a constituição de corpora. Numa altura em que o estabelecimento de corpora conhece um incremento cada vez maior, e em que a investigação linguística se apoia cada vez mais em dados provenientes de corpora, torna-se necessário adoptar a maior clareza possível na sua descrição.

Este artigo, bem como as recomendações nele contidas, devem ser lidos em conjunto com as recomendações do EAGLES sobre Tipologia do Corpus (Sinclair, 1994).

Em discussões sobre o formato das Recomendações, foi explicado aos fornecedores de dados que o formato desejado consiste numa hierarquia de atributos, cada um com um conjunto de valores, visto este

formato convir à organização da base de dados, bem como ao processamento computacional. Por essa razão, tal formato foi adoptado sempre que possível. No entanto, em algumas circunstâncias, não constitui um modelo apropriado dos padrões de linguagem em comunicação; em tais casos, recorreu-se a outros formatos. Em outros casos ainda, a ausência de um conjunto de valores completo revela apenas falta de experiência na aplicação da classificação tipológica; noutros, o conjunto de valores depende daquilo que os textos dizem de si próprios (cf. categoria 'reflexa', mais abaixo), não estando, portanto, sob o controlo de classificadores; noutros casos ainda, os critérios determinantes são internos, e seria um grave erro de julgamento adoptar um conjunto de valores externos, que obscureceriam os internos.

Por vezes, o modelo valor/atributo é simplesmente irrelevante – seria extremamente improvável que um modelo tão simplista como este se mostrasse sempre adequado à descrição da linguagem humana. A atenção desta tipologia vai para a língua escrita. No entanto, é desejável que uma certa quantidade de material transcrito proveniente de língua falada faça parte, normalmente, de corpora de referência. Como tal, tomaram-se medidas básicas para que da Classificação constasse, também, a classificação de dados de língua falada. Na parte final do Relatório há uma curta Nota sobre Textos de Língua Falada.

### **Text Encoding Initiative (TEI) – Iniciativa de Codificação de Textos**

As linhas de orientação da TEI sugerem uma classificação textual baseada unicamente em critérios externos (i.e., critérios não-linguísticos). As categorias e subcategorias especificadas são simultaneamente exaustivas e de aplicação muito pesada. Na prática, isto é, em termos do número de horas necessárias para codificar cada texto, assim como dos recursos informáticos necessários para armazenar os cabeçalhos, o processo não parece ser viável para corpora do tamanho preconizado pelo EAGLES (cf. artigo sobre Tipologia do Corpus). No nível mínimo de codificação do TEI, anotam-se os pormenores relativos à forma como o texto foi informatizado (método utilizado, pessoas responsáveis, data, etc.), pormenores relativos à publicação (data, endereço da editora, situação relativa ao copyright, etc.) e descrição bibliográfica (informação sobre o autor, por exemplo, idade, sexo, lín-

gua materna, nacionalidade, etc., e pormenores sobre a forma de publicação do texto – série, revista, etc.). Recomenda-se, ainda, que o tamanho do ficheiro, bem como afirmações relacionadas com a edição e série sejam também armazenados. No âmbito destas categorias pode ser armazenada toda a informação de natureza não-linguística acerca do texto.

A lista é interminável, não existindo nela nenhuma alínea de relevância linguística, que nos ajude a distinguir entre as componentes externas que possam vir a tornar-se úteis ao utilizador, e as que apenas têm interesse do ponto de vista do arquivo. Quase todas as componentes são susceptíveis de se tornarem relevantes no futuro, o que não constitui, contudo, razão suficiente para incorrer no custo que representa a identificação e posterior armazenagem de cada componente, para cada texto do corpus. De qualquer forma, a prática já demonstrou que uma tal pre-categorização será certamente inadequada, basicamente devido a duas razões: a primeira deriva da nossa experiência nestes casos – os utilizadores pretendem sempre componentes que não se encontram rapidamente disponíveis; a segunda diz-nos que as razões que salientam uma determinada componente do contexto ou situação subjacentes a um determinado texto, envolverão uma reconceptualização da relação entre língua e situação, o que irá marginalizar qualquer classificação prévia nesse sentido.

Considere-se o caso da «correção política», típico de um parâmetro de classificação da língua que não existia há alguns anos, mas que poderia hoje ser objecto de sérios estudos. Vários tipos de linguagem ofensiva têm sido tradicionalmente recenseados; a blasfémia, por exemplo, tem uma longa história, assim como a pornografia (enquanto conceito oposto ao erotismo); a linguagem racista tem sido, mais recentemente, identificada e classificada, assim como a linguagem sexista. Mas a correção política requer uma re-conceptualização da relação entre o escritor, o texto, o leitor, e uma terceira parte, relação essa que não aparece nas classificações anteriores; daí o seu interesse. Até dispormos dos conceitos que fornecem uma razão adequada para uma determinada classificação, esta não é necessária.

Logo, não é boa política tentar adivinhar o futuro. Em vez disso, deveríamos deslocar a nossa atenção para as componentes externas, que se sabe, ou suspeita, influenciarem, sistematicamente, determinadas escolhas linguísticas.

### **Critérios Internos e Critérios Externos**

Os primeiros corpora electrónicos foram desenhados com recurso a critérios externos – fazendo-se referência a tipos de texto institucionalizados, ou a características do meio não-linguístico ou social em que os textos ocorriam. Mais recentemente, alguns critérios internos – características diferenciadoras do tipo de língua de cada texto – têm sido avançados por alguns investigadores. Este trabalho sugere que uma classificação rigorosa dos textos, uma tipologia adequada, consistiria, eventualmente, numa combinação equilibrada dos dois tipos de critérios.

Muitos critérios internos e externos reflectem-se mutuamente. Um texto que exiba frases compridas provém, provavelmente, de um determinado tipo de livro, ou de revista. Para alguns parâmetros de classificação, a informação externa é fundamental – determinado texto aparece num livro, ou numa revista? É um texto escrito, ou falado?... Independentemente do tipo de língua que se venha a encontrar no documento, ou na transcrição, tais distinções – relativas à realidade – são claras. Grande parte da investigação efectuada sobre variedade e géneros linguísticos, entre outros, assume que as distinções externas se reflectem internamente; esta suposição é, de facto, o único factor que confere relevância aos critérios externos – pois, se estes não tivessem influência sobre a linguagem usada, não teriam qualquer interesse para os linguistas e outros protagonistas das indústrias da língua.

Para se decidir se determinado critério de desenho de um corpus deve ser expresso através de critérios internos ou externos, talvez o melhor indicador consista em saber-se se esse mesmo critério é expresso intensiva ou extensivamente. Um critério intensivo exprime-se através de uma espécie de generalização, que é partilhada por todos os membros do grupo, enquanto que um critério extensivo não passa de uma lista de todos os membros. Quando o grupo em questão é de dimensões modestas, e claramente delimitado (como, por exemplo, o dos «jornais de qualidade», no Reino Unido), não é preciso escolher entre listá-los, ou produzir uma generalização que os identifique. No entanto, quando o número de membros do grupo é grande, as técnicas de definição extensiva não funcionam.

Assim, quando nos deparamos – como numa classificação do tema através de critérios externos, listando-se os possíveis temas – com um exemplo de listagem extensiva, sem princípio – ou dimensão – pré-

-estabelecidos, estamos perante um caso em que seria desejável abandonar esta abordagem, e procurar, antes, um meio generalizado de classificar os textos, utilizando critérios internos e definições intensivas.

### **Tema**

O tema é uma das áreas da tipologia textual em que encontramos mais controvérsia. Nenhuma classificação externa parece ser satisfatória. Recomendamos que o tema seja sobretudo classificado como um assunto interno, que tenha a ver com coisas como as escolhas vocabulares de um texto, e não como um assunto externo, em que todo o Universo é incansavelmente dividido em sub-categorias. Devido à importância deste ponto, vamos antecipar a nossa própria posição sobre o assunto, e ilustrar a linha de argumentação que estamos a desenvolver na consideração desta área. Afirmamos o seguinte:

Na classificação do tema, a evidência interna é de importância primária.

Queremos com isto dizer que, para chegar a uma melhor classificação do tema, a evidência interna (como, por exemplo, a formação de grupos vocabulares) deve ser analisada em primeiro lugar, sendo a evidência externa abordada numa fase de maior minúcia.

Para consubstanciar esta afirmação, considere-se a variação temática em documentos e conversas. Um jornal cobre um vasto âmbito de temas, com alguma estabilidade de número para número, mas também com grande variação. Se não for anotada manualmente (e este Relatório prevê corpora grandes demais para serem anotados desse modo), a versão informatizada do texto de um jornal não é explícita no que respeita aos temas.

Esta posição não se aplica aos casos em que os jornais incluídos no corpus são analisados por peritos, e divididos pelos seus artigos de fundo, reportagens, notícias, etc., sendo a cada grupo atribuído um tema.

Os jornais dividem-se em Secções: por exemplo, a Secção Desportiva. Trata-se aqui do tipo de auto-classificação a que aludiremos mais tarde como 'reflexiva'. Se utilizarmos a auto-classificação, para que haja uma consistência temática, pode ser útil ver a estrutura de um jornal a duas dimensões – o Número e o Conteúdo.

	notícias	desporto	mulheres	ciência
Número 1	⌈	⌈	⌈	⌈
Número 2	⌈	⌈	⌈	⌈
Número 3	⌈	⌈	⌈	⌈

Quadro 1

A dimensão vertical do Quadro 1 exhibirá, provavelmente, maior consistência a nível de temas do que a sua dimensão horizontal. Do mesmo modo, um romance cobre toda uma variedade de temas, mesmo sem a segmentação do jornal; divide-se em capítulos, e não há qualquer requisito que obrigue o autor a mostrar consistência temática dentro de cada capítulo. Os romances populares (do tipo 'histórias de espionagem') dividem frequentemente cada capítulo num determinado número de curtos episódios, cada um subordinado a um tema diferente.

Uma conversa espontânea não tem compartimentos estanques, movendo-se de tema para tema, normalmente numa série de etapas que acabam por obscurecer a segmentação temática. Todo este processo é muito subtil (Hazadiah, 1993), e não se prevê a existência de nenhum método capaz de automatizar a análise do discurso de modo a torná-la eficiente neste domínio.

Deparamo-nos, assim, com três importantes fontes de dados para corpora linguísticos, nas quais a aplicação de uma análise temática externa não parece ser útil. Em alguns tipos de documento, a relação com o tema pode ser mais simples (revistas especializadas, relatórios formais, etc), mas, mesmo quando o material parece confinado a uma área temática restrita, seria imprudente manifestar demasiado optimismo.

Assim, embora certos sistemas de classificação geral tenham sido criados (Dewey, etc.), estes pura e simplesmente não reflectem a estrutura textual. Sem dúvida que textos muito curtos, assim como excertos muito curtos de textos – fragmentos – poderão ser classificados desta maneira, mas não a maioria dos documentos e conversas. As razões são óbvias, e serão apenas muito brevemente mencionadas. Uma delas diz respeito à necessidade de misturar temas, necessidade esta que está na origem de muitas situações de comunicação – por exemplo, quando é necessário referir a influência de X sobre Y, etc. A resposta a este problema poderia residir numa classificação múltipla,

mas isso levar-nos-ia a um entrecruzar de temas, inútil quando se pretende ser prático.

Outra razão para a natureza imprevisível dos temas é a necessidade social de manter o interesse e a atenção do receptor, através de uma frequente re-centralização de temas (em conversa). Esta característica, combinada com o requisito social de obscurecer o momento em que há mudança de tema, torna este último um critério de utilidade duvidosa para a classificação, seja ela externa ou interna.

A Tipologia de Temas constante do Anexo 2 foi criada a partir do maior número de fontes publicadas que conseguimos encontrar, na sequência do estudo NERC, e demonstra como é inadequado tentarmos basear-nos em práticas anteriores dadas como correctas (observe-se a sua inconsistência), ou tentar organizar uma hierarquia de etiquetas temáticas simples. As memórias de um missionário médico reformado, possuidor de uma importante colecção de quadros militares, nomeadamente telas que ilustravam pormenores de material bélico e uniformes militares; uma pessoa que se deleitava com as línguas e monumentos que tinha encontrado nas suas viagens, e que prestava grande atenção ao nível de sofisticação científica da agricultura dessas regiões, assim como aos problemas causados pela distância dos grandes centros... – um documento como este, que não é de modo algum fantástico, seria indexado sob, pelo menos, uma dúzia dos temas constantes do Anexo 2.

## **Género**

A classificação de textos em diferentes géneros parece ser mais conseguida através de critérios externos do que de critérios linguísticos. Bhatia (1993) afirma que a natureza do género pode ser especificada através de critérios externos, sendo esta a forma segundo a qual os sistemas de classificação parecem ser identificados. Os critérios externos normalmente especificados compõem-se de informações referentes ao emissor/escritor e ao seu público, assim como das que dizem respeito às relações entre estes dois grupos, aos objectivos do autor, aos pormenores históricos, socio-culturais, filosóficos e ocupacionais. Para determinar o género de um texto, Bhatia relacionaria ainda determinado texto com outros do mesmo género, identificando o seu tema, mais uma vez, através do recurso a critérios externos.

Uma distinção básica consiste em saber se o texto é escrito ou oral, embora Sager (1990) identifique tipos de texto dentro da 'linguagem da ciência e da tecnologia' que podem pertencer a ambas as categorias, tais como o relatório, o memorando e o inventário, em que o meio de difusão não determinaria necessariamente o tipo de texto. Este é, antes, determinado através do estatuto dos participantes, e do seu nível de conhecimento ou autoridade em relação ao tema do próprio texto. O tipo de texto, ou modo de transmissão, baseia-se, também, no nível de preparação prévia, ou seja, há que ver se o texto é espontâneo ou preparado. Sager afirma, ainda, que as intenções informativa e avaliadora também constituem critérios para a subdivisão textual. (Esta última é a classificação básica da parte escrita do British National Corpus – que se divide, em primeiro lugar, entre escrita informativa e escrita imaginativa – sendo a intenção aquilo que, fundamentalmente, divide os diferentes tipos de texto. A escrita informativa inclui os seguintes campos: ciências, artes, comércio e finança, crença e pensamento, tempos livres, ciências naturais e ciências puras, ciências sociais e temas mundiais. Quanto à escrita imaginativa, esta é composta por obras criativas e obras literárias. Estes géneros são, então, categorizados de acordo com outros critérios, tais como o meio de difusão, a data de publicação e o tema).

É importante rejeitar uma distinção antiga entre texto imaginativo e texto informativo. Até a uma fase relativamente tardia de classificação, a distinção entre facto e ficção é relevante, mas, em todo o caso, trata-se de uma distinção de tipo entrecruzado. Exageros de simplificação, respeitantes ao conceito de informação, são abundantes na própria ciência da informação – de facto, toda ela se baseia numa simplificação exagerada, que preconiza que a informação pode, de algum modo, ser separada do resto da comunicação. Hunston (1989) demonstrou que o mais neutro e objectivo dos relatórios científicos esconde uma abundância de avaliação.

Em face de tudo isto, na tipologia proposta, a categoria 'informação' é uma das categorias que indica 'resultado', sendo restringida a obras de referência sem relevância prática.

Halliday e Martin (1993) não fornecem um sistema de classificação textual formal: identificam, através de critérios internos, áreas como campo, modo e conteúdo. Dado que a linguagem varia conforme a situação (o que é, presumivelmente, a razão fundamental para a exis-



tência de uma classificação textual), podemos identificar certo tipo de linguagens através de uma especificação de determinados critérios linguísticos internos – isto é, a frequência de características léxico-gramaticais. Halliday e Martin associam o texto científico a passivas, nominalizações, etc. No entanto, parece-nos que, aqui, seria desejável a intervenção de um tipo de classificação externa inicial, pelo menos à partida, de modo a especificar as características de um género em particular – neste caso, o ‘texto científico’. Os critérios linguísticos internos do texto seriam analisados após a selecção inicial, baseada em critérios externos. Os critérios linguísticos seriam subsequentemente confirmados como sendo específicos do género ‘texto científico’. Este tipo de classificação começa com a classificação externa, evoluindo depois para os critérios linguísticos; se estes se voltam a ligar a uma classificação externa, ajustando-se as categorias de acordo com esta mudança, seguir-se-á um processo cíclico, que continuará até se atingir um nível de estabilidade.

Biber (1989), por outro lado, rejeita tipologias determinadas pelo isolamento inicial de importantes diferenças externas entre textos, tentando-se subsequentemente identificar as características linguísticas associadas a essa distinção. Deste modo, Biber estabelece uma tipologia textual baseada apenas em critérios linguísticos internos, interpretando, em seguida, os resultados em relação a funções. Os critérios internos de Biber são retirados de estudos sobre variação linguística já publicados; tais critérios foram, portanto, aprovados no nosso teste de relevância. A tipologia daí resultante distingue, de forma definitiva, oito tipos de texto, baseados em cinco «dimensões», sendo estas últimas identificadas automaticamente através de uma análise de associações de palavras, como iremos ver em seguida.

Esta classificação de textos, baseada tão-só em critérios internos, também parece insatisfatória, já que iria alienar o texto do seu ambiente sociológico, destruindo, assim, a relação entre os critérios linguísticos e os critérios não-linguísticos. Atkins *et al* (1992) acreditam que: «é impossível equilibrar um corpus baseando-nos, apenas, nas suas características extra-linguísticas», mas referem, também, que: «um corpus inteiramente seleccionado com base em critérios internos não ofereceria informações sobre a relação da linguagem com o respectivo contexto situacional». Parece extremamente viável começar por seleccionar textos baseando-nos em critérios externos; depois,

quando as características linguísticas particulares de um 'tipo de texto' tiverem sido estabelecidas através da análise de critérios internos, estes últimos poderão, então, ser usados na selecção e classificação de textos. Biber sugere, mais tarde (1993), que é necessário promover um processo cíclico de redefinição entre critérios externos e internos.

### **CrITÉRIOS EXTERNOS**

Tanto os critérios externos, como os critérios internos, devem, portanto, ser considerados, aquando da classificação textual. Uma tipologia de textos a serem incluídos num corpus não pode ser inteiramente baseada em apenas um tipo de critérios, sejam eles externos ou internos. Não é produtivo, nem desejável, classificar textos apenas através de evidência interna, se se pretender conhecer as relações entre características linguísticas e características não-linguísticas. Do mesmo modo, num sistema de classificação baseado exclusivamente na evidência externa, não se atingiria, necessariamente, o agrupamento de textos linguisticamente próximos, impondo-se, antes, divisões textuais que poderiam não reflectir importantes diferenças linguísticas.

Muitos critérios internos e externos podem estar, de algum modo, relacionados, e nem sempre é fácil decidir quais os fenómenos que serão tratados mais adequadamente por um ou outro tipo. Para complicar ainda mais a situação, existe uma categoria intermédia, que poderíamos designar por 'reflexiva' (Lyons, 1977). Neste último caso, o texto 'fala' de si próprio, propondo a sua própria classificação. A reflexividade é uma propriedade de todas as línguas, sendo a base de muito daquilo que é convencionalmente encarado como 'critérios externos'. O frontispício de um romance, por exemplo, apresentará o nome do autor, podendo a data aparecer também. Tais elementos serão tidos como 'factos' externos, a não ser que isso venha a ser, de algum modo, contestado. O mesmo frontispício pode também trazer a menção «Romance» que, sendo mais polémica, pode provavelmente ser aceite enquanto etiqueta de género. No entanto, a existência de tais propostas dentro do próprio texto não é prova suficiente da exactidão da classificação.

Nesta situação, deveríamos dividir os critérios externos em duas variedades:

*circunstanciais*, em que a evidência é exterior ao texto, e

*reflexivos*, em que a evidência provém de afirmações contidas no texto.

A evidência reflexiva deve, sempre que possível, ser confirmada pela evidência circunstancial, a não ser que tenha sido apontada como a única fonte de termos numa determinada categoria.

É conveniente reutilizar termos já comumente aceites como categorizadores pela crítica literária, pela retórica, etc, ainda que tais termos possam entrecruzar-se na organização hierárquica desta tipologia. Por exemplo, um romance é um livro (E.2...), sendo o seu receptor o público em geral (E.3.1...); trata-se de um texto do tipo literário-recreativo (E.3.2..). A termos como romance é aposto um asterisco, de forma a indicar que se trata de Termos Definidos Externamente (TDEs).

#### **Categorias Aceites** (normalmente retiradas do NERC)

**Género Literário** Dentro da categoria 'recreação', propõe-se uma distinção entre 'facto' e 'ficção', e não entre 'literário' e 'não-literário'. A «factualidade» é encarada como uma variedade da ficção, e, para evitar controvérsias, às obras religiosas é dada uma categoria separada. Dentro da ficção, mantém-se a distinção entre poesia e prosa, assim como entre texto dramático e texto não-dramático. Na prosa, encontram-se sagas, ciclos, trilogias e vários outros tipos de relações em série, que propomos sejam ignoradas num corpus com as actuais dimensões preconizadas. O romance, a novela e o conto constituem três bem conhecidas dimensões de prosa ficcional. Os romances podem ainda assumir outras classificações, como as de romance histórico, psicológico, de humor, de ficção científica, de amor, de suspense, policial, de espionagem, etc, havendo possibilidades alarmantes da existência de classificações duplas – romance de suspense e de espionagem, romance de amor histórico, etc. Aqui é conveniente utilizar a já referida categoria reflexiva – se um romance declara ser do tipo X, seria, então, razoável, assim como económico, classificá-lo de acordo com isso; no entanto, se essa auto-classificação não existe, dispender-se-iam esforços inúteis na tomada de uma decisão, e surgiriam controvérsias sem fim.

Biografia e auto-biografia classificam-se como factuais, juntamente com outras categorias, tais como obras académicas, livros de leitura e guias práticos, mas o seu fim é recreativo, e não institucional.

**Meio de difusão** A distinção entre livros, revistas, etc, encontra-se definida de forma praticamente universal, e é aqui seguida. Trata-se de uma distinção que se entrecruza, de algum modo, com os géneros literários, visto que, enquanto que um romance é quase de certeza um livro, um poema, normalmente, não o é – no entanto, um livro de poesia é um objecto familiar. Os romances de Dickens foram publicados por partes, o mesmo acontecendo com algumas obras modernas. Um obituário, no jornal, é um tipo de biografia, ao nível de uma curta peça jornalística; etc. No entanto, a percepção normal da classificação preconiza que se mantenha unido o grupo literário, apesar de isso acarretar o risco de duplicações, em detrimento de uma mistura com material não-literário, preconizada por outros critérios.

**Estilo** Esta é uma área sempre propícia a acesas disputas, e em que não se dispõe de um conjunto de valores consensual. Aspectos que têm a ver com o receptor de uma obra – o público leitor – podem afectar a linguagem, assim como a relação entre o escritor e o leitor. O estilo é uma categoria característica de um tipo de classificação, que tem na base critérios internos, em detrimento dos externos, e o estado insatisfatório das classificações feitas até aqui (por ex., Joos, 1961) sublinha esta situação. Mais tarde, no âmbito dos critérios internos, tratar-se-á com mais pormenor esta categoria.

**Modo de transmissão** Esta categoria divide-se, normalmente, entre 'escrito' e 'falado'. Adicionamos um terceiro elemento, 'electrónico', para sublinhar a nossa posição de que a linguagem transmitida através de meios electrónicos não coincide exactamente com a que se enquadra nos modos tradicionais.

O oral inclui-se na tipologia constante deste documento, e será tratado com mais pormenor mais abaixo; (cf. também Leech *et al.*, eds, 1995, para um comentário actual da problemática dos corpora de língua falada).

Aconselha-se que esta tipologia siga as recomendações do Relatório de Tipologia do Corpus, evitando confundir modos de transmissão

diferentes. Tal como referiu Firth (1957), a língua escrita comporta «a implicação da fala», e, do mesmo modo, qualquer polícia sabe que tudo o que se disser pode vir a ser escrito. Como tal, a única forma de manter os dois modos separados é classificar de acordo com o modo exibido pelo material aquando da sua inclusão no corpus. Assim, descrições do tipo 'escrito para ser falado' não constam da tipologia; tais descrições encontram-se demasiado próximas das operações da Falácia Intencional.

Pensa-se frequentemente que a língua falada é menos formal, mais espontânea, do que a língua escrita. Esta suposição pode ser estatisticamente verdadeira, numa perspectiva lata e geral, mas há grande abundância de língua falada formal, na rádio ou na televisão, em círculos oficiais e legais, bem como na área da diplomacia. Do mesmo modo, existe ainda uma abundância de correspondência informal, tendo, nesta área, a invenção do e-mail sido responsável, possivelmente, pelo aparecimento de uma grande quantidade de língua escrita informal.

Este ponto é aqui ilustrado pela fig.1 (*apud* Sinclair, *in* Leech *et al*, 1995), em que se dá conta da relação entre géneros de material escrito e falado e dimensões do público.

oral	GRANDE	escrita
rádio/televisão	1 000 000s	jornais
rádios locais	100 000s	revistas/livros
comícios	1 000s	avisos
conferências	100s	publicações locais
aulas	10s	actas de reuniões
debates	5s	circulares
entrevistas	3s	grupos de trabalho
conversas	2s	cartas pessoais
PEQUENO		

Fig. 1 – Dimensões de público.

Na sequência das recomendações constantes da Tipologia de Corpus, esforçámo-nos, aqui, por evitar categorias mistas, já que a classificação daí resultante se torna arbitrária e perde o seu significado. O que é dito pode ser passado à forma escrita, e o que é escrito pode vir a ser lido; quer a fala, quer a escrita, podem ser convertidas para

um formato electrónico com facilidade. Actualmente, cada vez mais textos se encontram disponíveis em mais do que um formato; por exemplo, a correspondência composta num processador de texto pode ser posteriormente impressa, os arquivos de um jornal podem ser inseridos num CD-ROM.

No entanto, é raro os textos serem compostos tendo em vista a sua transmissão através de mais do que um modo, sendo normalmente fácil identificar o modo original, quando existe mais do que um. A conversa, gravada, transcrita e digitada num computador, existe nos três modos, mas se foi seleccionada por se tratar de uma conversa, então irá preencher um lugar na dimensão falada do corpus. A transcrição escrita pode normalmente ser eliminada, se não contém mais informação do que a versão electrónica, mas a gravação sonora original irá normalmente ser preservada, dado que nenhuma transcrição a substitui adequadamente (Leech *et al*, 1995).

Para obter, então, uma classificação significativa, é necessário voltar ao desenho do corpus, e ao ponto preciso, nesse desenho, que determinado texto irá suprir. A pessoa que constitui o corpus deve ver claramente em que modo de transmissão se enquadra o texto, etiquetando-o de acordo com isso.

Todos os textos que compõem um corpus acabam, é claro, por assumir um formato electrónico, embora não seja essa a razão do estabelecimento da categoria 'modo electrónico' nesta tipologia. A razão prende-se, antes, com o facto de actualmente se verificar a existência de uma grande quantidade de material textual que aparece apenas sob formato electrónico, ou que aparece primariamente nesse formato. As características comunicativas deste novo meio de expressão conduzem a escolhas singulares, a nível da forma e do estilo, escolhas essas que não se encontram nos textos escritos ou falados.

Existem, e sempre existiram, categorias controversas a meio caminho entre a escrita e a fala – por exemplo, um argumento teatral. Embora escrito, e muitas vezes impresso, ninguém discute que o objectivo do argumento teatral é, eventualmente, o de vir a ser dito em representação. Contudo, uma peça bem sucedida será produzida muitas vezes, com muitas alterações ao texto; cada representação, em cada produção diferente, é uma experiência artística distinta, havendo boas razões para a sua gravação e arquivação. O texto original da peça não é, no entanto, afectado por tais representações faladas (a não ser que o autor venha a alterar o texto, criando, assim, novas edições).

Intimamente aparentado com o argumento teatral, encontramos o diálogo, nos romances – um tipo de representação da língua falada, mas cujo primeiro objectivo não é o de surgir em modo falado; se, como acontece com frequência actualmente, o autor pensar numa adaptação ao cinema, ou à televisão, um novo texto será preparado, provavelmente por outra pessoa, que não o autor, podendo o diálogo mudar substancialmente. Alguns autores, como, por exemplo, John Steinbeck, escrevem romances em muito semelhantes a argumentos teatrais, outros, tais como Faulkner e Runyon, produzem obras cuja linguagem é escrita para representar o discurso do narrador.

Tais categorias são complexas, e passíveis de grande variação; muitos autores experimentam novos formatos, e já existem romances electrónicos, em que o leitor participa na escolha do texto que lê (também há notícia de pelo menos um romance em suporte de papel que segue o modelo de «descoberta programada», em que o leitor deve optar por vias diferentes em dados pontos). Na presente tipologia prevêem-se categorias 'especiais', já institucionalizadas e separadamente definidas, podendo ser utilizadas com vista à manutenção da clareza da classificação, sem que seja necessário optar por categorias mistas gerais.

## **Tipologia**

**E.1. origem** – assuntos relacionados com a origem do texto, que se pensa afectarem a sua estrutura ou conteúdo.

**E.2. estado** – assuntos relacionados com a aparência do texto, com a sua forma de edição, e com a sua relação com matéria não-textual, na altura da sua selecção para o corpus.

**E.3. objectivos** – assuntos relacionados com as razões para a produção do texto, e com o efeito que este é suposto provocar.

O estudo dos parâmetros internos encontra-se numa fase muito menos avançada do que aquele que se reporta aos parâmetros externos. Consideram-se parâmetros internos principais:

I.1. **tema** – o assunto e o(s) domínio(s) de conhecimento do texto.

I.2. **estilo** – os padrões de linguagem que se considera estarem relacionados com os parâmetros externos.

Por vezes, determinado aspecto da classificação é reflexo, já que contém uma afirmação quanto à sua origem, público, etc. Se isto for utilizado como base para a classificação externa, recomenda-se a aposição de «R» após a classificação. Uma declaração no texto de um documento ou transcrição não é necessariamente aceite como correcta, mas deve ser sempre notada; qualquer discrepância entre a classificação reflexa e a que foi seleccionada para a tipologia deve ser justificada.

#### E.1. **origem**

Os principais parâmetros que exprimem a origem de um texto são:

E.1.1 **intervenientes** – os vários intervenientes, cujo trabalho ajudou a dar ao texto a sua forma definitiva.

Os papéis relevantes incluem os de:

E.1.1.1 **autor** – a pessoa que escreveu o texto, que produziu o trabalho original.

E.1.1.2 **editor de provas** – qualquer pessoa que tenha alterado ou aconselhado alterações ao texto, uma vez este último produzido, ou quem o prepara para uma mudança de formato.

E.1.1.3 **editor** – a pessoa identificada como responsável legal pelo acto de publicação.

E.1.1.4 **detentor do copyright** – a pessoa com direitos legais sobre o texto publicado.

E.1.1.5 **tradutor** – a pessoa que traduz determinado texto para outra língua.



- E.1.1. **adaptador** – a pessoa que altera o texto, de modo a torná-lo adequado a outro género artístico (p.ex., guião cinematográfico, série televisiva, livro de banda desenhada).

Qualquer um destes papéis pode ser atribuído a mais do que um interveniente, assim como vários papéis podem pertencer a um só interveniente. Para cada interveniente pode ser relevante conhecer os seguintes pontos:

- E.1.1.~1 **idade** na altura da composição; entre os 16 e os 60 anos uma indicação geral da década é suficiente; o trabalho efectuado por crianças ou pessoas da terceira idade deve ser identificado como tal, dado que esta característica parece comportar diferenças, a nível linguístico, do texto produzido por adultos em geral.

- E.1.1.~2 **sexo** na altura da composição

- E.1.1.~3 **influência linguística anterior**. Língua-mãe, seguida de indicação de qualquer outra língua aprendida na infância. Outras línguas relevantes para o texto devem também ser indicadas.

- E.1.1.~4 **domicílio** na altura da composição, se relevante e se se tratar de uma informação possível de obter.

- E.1.2 **processos** – ou seja, os processos de produção que se julgue terem tido influência no texto. Trata-se de uma categoria geral e vaga, que provavelmente não será muito utilizada, já que só raramente os processos de produção de um texto o transformam.

- E.1.3 **circunstâncias** – ou seja, qualquer material ou circunstâncias que se julgue serem relevantes para a estrutura ou conteúdo do texto. Trata-se de outra categoria de âmbito geral, e sem valores fixos.

- E.1.4 **planificação** – ou seja, questões como datação e «timing» que sejam relevantes para a construção do texto, normalmente de preenchimento obrigatório.

**E.2 estado**

Os textos são seleccionados a partir do fluxo natural da fala e da escrita, havendo, no momento da selecção, certos factores externos que podem ser relevantes.

**E.2.1 modo** – trata-se do modo de transmissão, ou seja, indicação se se trata de um texto falado, escrito, ou electrónico, essencialmente.

**E.2.1.1 falado** – a linguagem, quando seleccionada, encontra-se em formato falado, ou seja, originalmente, como uma onda sonora. O material incluído nesta categoria necessita de ser transcrito a partir da onda sonora, ainda antes de ser classificado.

**E.2.1.1.1 nível de percepção do informante** – indicação do grau de percepção do informante em relação à gravação.

**E.2.1.1.1.1 sub-reptício** – os informantes desconhecem por completo que estão a ser gravados.

**E.2.1.1.1.2 avisado** – os informantes sabem que há uma gravação a ser feita, mas não se encontram completamente na posse de todos os pormenores.

**E.2.1.1.1.3 consciente** – os informantes estão completamente conscientes do facto de o que dizem estar a ser gravado, e de a gravação vir a ser utilizada em análises linguísticas.

**E.2.1.1.2 localização** – ou seja, a localização dos informantes no momento da gravação

**E.2.1.1.2.1 estúdio** – a gravação faz-se no ambiente controlado de um estúdio de gravações.

**E.2.1.1.2.2 no exterior** – por exemplo, em casa, no trabalho, em viagem, num centro de ocupação dos tempos livres, etc.

- E.2.1.1.2.3 **telefone**
- E.2.1.2 **escrito** – a linguagem, quando seleccionada, encontra-se num formato escrito. Este material requer leitura óptica ou digitação, a partir de uma imagem, antes de poder ser classificado.
- E.2.1.2.1 **material impresso** – a linguagem, quando seleccionada, encontra-se em formato impresso, e deve ser digitada ou submetida a leitura óptica, com vista à obtenção de uma versão electrónica.
- E.2.1.2.1.1 **livros** – objectos com um número de depósito geral (ISBN).
  - E.2.1.2.1.1.1 **não-ficção**
  - E.2.1.2.1.1.2 **ficção**
  - E.2.1.2.1.1.1.~.1 **artes**
  - E.2.1.2.1.1.1.~.2 **economia**
  - E.2.1.2.1.1.1.~.~ **etc.**
- E.2.1.2.1.2 **jornais** – normalmente diários; habitualmente fáceis de distinguir da próxima categoria; com uma grande audiência, contendo notícias de actualidade.
- E.2.1.2.1.2.1 **notícias em geral** – trata-se da secção principal do jornal, onde os cabeçalhos são desenvolvidos, e os assuntos correntes relatados.
- E.2.1.2.1.2.2 **negócios/economia** – secções do jornal marcadas especificamente com a menção negócios/economia
- E.2.1.2.1.2.3 **desporto** – secções do jornal marcadas especificamente com a menção desporto
- E.2.1.2.1.2.4 **arte** – secções do jornal marcadas especificamente com a menção arte
- E.2.1.2.1.2.5 **especialidades** – secções do jornal marcadas de acordo com as especialidades a que se destinam.
- E.2.1.2.1.3 **revistas** – normalmente semanais ou mensais; a impressão é de melhor qualidade do que a dos jornais, sendo as revistas normalmente mais caras.

- E.2.1.2.1.4 **publicações de carácter efémero** – folhetos, panfletos, brochuras, revistas da paróquia, folhetos colocados na caixa do correio.
- E.2.1.2.1.5 **correspondência** – está a considerar-se uma sub-classificação para esta categoria, que envolva as noções de automático e humano, oficial e pessoal.
- E.2.1.2.2 **material digitado** (incluindo aquele que resulta de processamento de texto) – todo o tipo de relatórios e documentação.
- E.2.1.2.3 **manuscrito** – textos escritos manualmente.
- E.2.1.3 **electrónico** – a linguagem, quando seleccionada, encontra-se num formato electrónico (e-mail, boletins electrónicos, páginas da Internet, etc.), ou seja, não requer conversão, nem a partir de uma versão oral, nem a partir de uma imagem visual. Pode encontrar-se, eventualmente, já classificada, caso em que a classificação necessitará de ser adaptada às convenções utilizadas no corpus.
- E.2.2 **relação com o modo** -se escrito, de que tipo é o papel, a impressão, etc.; se falado, informação sobre as condições acústicas, etc.
- E.2.3 **relação com matéria comunicativa extra-linguística** – diagramas, ilustrações, outros meios associados à linguagem numa situação de comunicação.
- E.2.4 **aparência** – pode haver, por exemplo em folhetos publicitários, aspectos na apresentação do documento de características excepcionais ao nível do design, que possam exercer um efeito importante, no que diz respeito à linguagem.

**E.3 objectivos**

**E.3.1 público** – a pessoa ou grupo de pessoas para quem o texto é criado.

**E.3.1.1 público imediato** -pessoas que formam o evento comunicativo, com oportunidade, ainda que teórica, de participar.

**E.3.1.2 público mais lato** – por exemplo, os leitores, os espectadores.

Entrecruzando-se com estas categorias, encontramos:

**E.3.1.~.1 – dimensões do público** – Por exemplo, menos de 5 / 5-20 / 21-50 centenas / milhares.

**E.3.1.~.2 – constituição do público**

**E.3.1.~.2.1 membros do público em geral**

**E.3.1.~.2.2 leigos bem-informados**

**E.3.1.~.2.3 profissionais**

**E.3.1.~.2.4 especialistas**

**E.3.1.~.2.5 estudantes, estagiários**

**E.3.1.~.2.3.1- n lista das profissões**

**E.3.1.~.2.4.1- n lista das especialidades**

**E.3.1.~.2.5.1- n lista dos cursos e estágios**

Nota: Se assim convier, estas listas podem ser relacionadas com listas internas de temas.

**E.3.1.~.3 – relação autor – público**

**E.3.1.~3.1 distante** – autor e público não se conhecem pessoalmente, encontrando-se ainda mais separados pelos papéis institucionais que representam, e que os despersonalizam.

**E.3.1.~3.2 neutra** – autor e público não se conhecem pessoalmente, mas, tanto o autor, como o membro do público, são vistos enquanto indivíduos.

- E.3.1.~3.3 **próxima** – autor e público conhecem-se pessoalmente, ou assumem essa atitude. (Note-se que esta é uma categoria em que pode haver discordância entre uma possível classificação reflexa pré-existente no texto, e o julgamento do investigador, por ex., no caso das circulares comerciais, que, embora pareçam personalizadas, devido a automatizações em lista efectuadas por computador, não o são na realidade).

### E.3.2 **resultado pretendido**

- E.3.2.1 **informação** -é um resultado pouco provável, já que os textos são raramente criados tendo apenas esse fim em vista. Aparece principalmente em compêndios de referência.

- E.3.2.2 **discussão** -afirmações formais que envolvam polémicas ou tomadas de posição; discussões.

- E.3.2.3 **recomendação** – relatórios, documentos legais, regulamentares, ou de consultoria.

### E.3.2.4 **recreação**

- E.3.2.4.1 **ficção** – incluindo factuality

- E.3.2.4.1.1 **romance/novela/conto**

- E.3.2.4.1.2 **histórica (reflexiva), romance de suspense, etc.**

### E.3.2.4.2 **não-ficção**

- E.3.2.4.2.1 **biografia**

- E.3.2.4.2.2 **autobiografia**

- E.3.2.4.2.3 **cartas** – da variedade publicada, e não correspondência.

- E.3.2.5 **religião** – livros sagrados, livros de oração, missais. Além de uma categoria textual pertencente aos critérios externos, a religião é também um Tema importante – integrado nos critérios internos. Assim, se não houver cuidado, esta categoria poderá sobrepor-se a outras nesta hierarquia, pelo que se recomenda o seu uso apenas para textos que não possam ser classificados sob outras etiquetas. Não basta que o tema seja religioso (por ex., a biografia de uma personalidade religiosa) para poder ser classificado pela categoria em epígrafe.

#### E.3.2.6 **instrução**

- E.3.2.6.1 **obras académicas** – escritas para um público especialista (cf. E.3.1~2.4), em que o autor é também um especialista.
- E.3.2.6.2 **livros de estudo** – escritos para um público estudantil, tendo, portanto, o autor, uma reputação de experiência apropriada, quer académica, quer profissional.
- E.3.2.6.3 **livros práticos** – estes são escritos para ensinar e guiar no exercício de trabalhos práticos.

### **CrITÉRIOS INTERNOS**

Desenvolver-se-á aqui a posição segundo a qual dois parâmetros centrais na classificação de textos serão descritos mais adequadamente utilizando-se critérios internos, ou linguísticos, por oposição a externos, ou socio-culturais.

Esses parâmetros são o tema e o estilo. As tentativas de os analisar, de forma a criar uma base de classificação, são normalmente expressas em termos de critérios externos, resultando insatisfatórias a vários níveis. Não sendo normalmente apoiadas por evidência científica, são impossíveis de repetir e, não beneficiando de consenso geral, redundam numa co-existência de diferentes versões, sem que haja critérios sólidos para poder escolher uma em detrimento das restantes. Não há qualquer controlo sobre a granularidade da classificação, nem dimensões suficientes para a grande quantidade possível de classificações entrecruzadas. A possibilidade de tema e estilo serem adequados à

análise atributo/valor é extremamente remota, e inatingível no presente estado do nosso conhecimento.

Os critérios externos seguem distinções e classificações já existentes na nossa cultura, quer de documentos, quer de eventos falados. Podem não utilizar toda uma classificação pré-existente, assim como podem utilizar factos culturais que não são normalmente aplicados na classificação textual, mas a sua característica definidora revela que os critérios que compõem uma análise criteriosa externa se encontram na nossa cultura, e não na nossa língua.

Considere-se, por exemplo, a categoria «jornal de qualidade». Trata-se de um termo informal, em distribuição complementar com o termo «tablóide»<sup>1</sup>. Outras categorias poderão, no entanto, surgir. No passado, no Reino Unido, o preço funcionava como uma importante distinção, mas actualmente trava-se uma guerra de preços que invalida esse critério – há alguns anos os jornais de qualidade não traziam notícias na primeira página, apenas anúncios e avisos. A folha de um tablóide é aproximadamente metade da de um jornal de qualidade; o editorial de um tablóide consiste em meia-dúzia de frases respeitantes a um determinado assunto, enquanto que nos jornais de qualidade o editorial pode ser um ensaio erudito de duas ou três mil palavras, exprimindo o ponto de vista do director.

Todos estes pontos constituem critérios externos. O que se passa, é que não se olha para o interior dos jornais, para o tipo de língua que usam, de forma a classificá-los. Encontrar-se-ia, por exemplo, um estilo bem conhecido de escrever cabeçalhos, em alguns tablóides – utilizando palavras curtas e muito poucas palavras gramaticais (Lorde Nu em Tempestade Sexual), o que poderia constituir um critério interno; este procedimento, no entanto, não tem sido considerado.

## **Tema**

O tema é o aspecto lexical da análise interna de um texto. Em termos externos, o problema da classificação prende-se com o facto de haver demasiados métodos de utilização possível, e nenhum consenso ou estabilidade nas várias sociedades, ou a um nível superior a estas últimas, sobre o qual se possa construir uma classificação. Existem classificações semânticas (tal como o célebre Thesaurus de Roget



(Roget, 1962)), ou classificações bibliográficas (tal como a de Dewey). Os sistemas educativos dividem o conhecimento numa miríade de hierarquias confusas, mudando de ideia a cada passo. Hierarquias e outras organizações terminológicas podem ser utilizadas como identificadores temáticos; os terminólogos organizam bancos de termos baseados em princípios de entrecruzamento de áreas, fornecendo assim uma forma de organização do conhecimento conceptual.

Tentar produzir um modelo temático que englobe tudo aquilo sobre o que se pode escrever ou de que se pode falar, convenientemente arrumado em caixas distintas, de tal forma que cada texto possa ser colocado numa só caixa, com apenas uma pequena percentagem de sobreposição ou dúvida, é, de facto, um acto de excessiva simplificação.

No thesaurus de Roget (1962), a língua inglesa está dividida essencialmente em seis categorias globais: Relações Abstractas, Espaço, Mundo Material, Intelecto, Volição e Afectos. A partir daqui surgem subcategorias que dividem o mundo em áreas subsequentes, de mais fácil compreensão; por exemplo, a categoria «Afectos» divide-se em Afectos Gerais, Pessoais, por Simpatia, Morais e Religiosos. A este nível pode já ser-nos difícil aceitar diferenças fundamentais entre, por exemplo, afectos morais e religiosos, especialmente se os nossos valores morais estão profundamente ancorados na religião. As fronteiras tornam-se ainda mais vagas se dividirmos a linguagem em ramos menos abrangentes: sob as subcategorias surgem novos agrupamentos temáticos. Na sequência da categoria «Afectos», e, por exemplo, da subcategoria «Pessoais», encontramos os agrupamentos «Passivos», «Discriminadores», «Em Perspectiva», «Contemplativos» e «Extrínsecos». Estes dividir-se-ão, por seu turno, até totalizar os 1000 temas por meio dos quais podemos convenientemente categorizar a linguagem.

A este nível de categorização, as fronteiras entre os temas propostos tornam-se muito difíceis de distinguir, e as referências cruzadas também aumentam grandemente. Ao avançarmos na hierarquia encabeçada por «Afectos – Pessoais» encontramos, sob o nível «Passivos», temas como Alegria, Sofrimento, Agradabilidade, Dor, Contentamento, Descontentamento, etc. A este nível, contudo, como se pode diferenciar entre o afecto classificado como «Afectos, Pessoais, Em Perspectiva, Antipatia» e aquele indexado com os rótulos «Afectos, por Simpatia, Sociais, Ódio»? Ou ainda, como distinguir entre linguagem classificada

como «Comunicação de Ideias, Meios de Comunicar Ideias, natural, convencional, carta», e aquela rotulada «Comunicação de Ideias, Meios de Comunicar Ideias, língua escrita, correspondência»? Será desejável, na classificação temática da língua, fazer tais distinções?

As fronteiras entre temas são, de facto, indistintas, e defenderíamos que a classificação temática de corpora se fizesse a um nível superior, com poucas categorias temáticas que se alterariam com a inclusão de novos dados linguísticos.

Há numerosas formas de classificar textos de acordo com o tema. Cada projecto de corpus possui as suas políticas e os seus próprios critérios de classificação, situação que funcionou, de facto, como mola propulsora subjacente a este artigo, em que se pretendem oferecer hipóteses de linhas de força para a classificação textual. O facto de existirem tantas abordagens diferentes para esta questão, aliado ao facto de diferentes temas classificativos serem identificados por diferentes grupos, indica que as classificações existentes provêm de fora, do linguista, por oposição à linguagem em si. A categorização subjectiva da linguagem conduzirá inevitavelmente a diversidades nas categorias estabelecidas, já que qualquer tipologia daí resultante será apenas uma visão da linguagem, entre as muitas possíveis.

O relatório NERC inclui um resumo dos sistemas de classificação utilizados pelos principais projectos de corpora da Europa. Embora tal resumo possa parecer demasiado extenso, ele limita-se, de facto, a oferecer uma visão consolidada das categorias comuns a todos os projectos. Alguns sistemas classificativos são mais pormenorizados do que outros. No Corpus Dinamarquês encontramos uma extensa lista de temas, desde «transporte» até «música», desde «negócios» a «ambiente» (cf. Norling-Christensen, 1995, Anexo 1); há até uma área temática separada para a «UE». Neste tipo de sistema classificativo, como se pode justificar a classificação de um texto segundo uma das 66 áreas temáticas propostas? A maior parte dos textos pertencerá a mais do que um tema; por exemplo, «lei», «crime» e «sociedade» são considerados temas separados, quando, de facto, estão inextricavelmente ligados.

Tal como nos é explicado, «Foram distinguidos textos específicos de textos de aplicação geral» (Calzolari *et al.*, 1995). Embora esta tabela ilustre apenas as categorias temáticas de «aplicação geral», há grande variação entre as classificações técnicas. Observando a lista de

categorias disciplinares propostas, encontramos «Lazer» enquanto categoria separada de, por exemplo, «Desporto»; a categoria «Ciência» e depois, separadamente, «Física», «Biologia», «Química»; «Finanças» e «Economia». Certamente que surgiriam problemas na classificação de textos a partir de tais etiquetas temáticas, já que as fronteiras entre elas não são, de modo algum, claras. A decisão acerca da etiqueta a aplicar a cada texto acabaria por ficar a cargo de uma pessoa, ou de um grupo de pessoas, o que não é um procedimento satisfatório.

Recomenda-se aqui que o tema passe a ser determinado através do uso de critérios internos, de modo a fornecer uma justificação linguística para qualquer categorização textual que possa vir a ser efectuada com base no tema. As linhas de orientação para esta tarefa serão baseadas em resultados provindos de uma análise objectiva, assistida por computador, dos textos a serem incluídos no corpus. Proceder-se-á a uma recensão do trabalho que tem vindo a ser executado, até à data, sobre a análise informática de características linguísticas de textos. Este procedimento ajudará a estabelecer um conjunto adequado de temas, segundo os quais se poderão classificar os textos dos vários corpora, bem como seleccionar novos textos de forma coerente, de modo a manter o equilíbrio desejado. Uma importante parte da investigação nesta área tem sido efectuada por Martin Phillips, sendo o seu trabalho aqui recenseado, já que se trata do tipo de metodologia de determinação de temas que recomendamos.

Martin Phillips (1983) propõe uma linha de conduta para a determinação do tema num texto que passa por uma metodologia distribucional objectiva e quantitativa. Aquilo que na presente tipologia temos vindo a denominar «tema», na classificação textual, Phillips denomina «acerca de», isto é, «a percepção psicológica da disciplina-assunto».

Na sua óptica, a qualidade «acerca de» de determinado texto deve-se aos padrões globais que o constroem, ou seja, a sua «macroestrutura». Na sua tese, analisa a macroestrutura de um dado número de textos recorrendo a meios informáticos, para que os resultados derivem do próprio texto, e não de estruturas externas. Sublinha, ainda, aquilo que temos vindo a defender no presente artigo, ou seja, que qualquer afirmação sobre o assunto, ou sobre a qualidade de «acerca de» de um texto, deve basear-se numa análise objectiva das suas características linguísticas. Seria um erro tentar esquematizar a

língua com base em estruturas externas – trata-se de um método reconhecidamente impossível, como vimos no relatório NERC.

Para poder emitir uma opinião sobre o tema de determinado texto, o leitor tem primeiro de compreender esse texto, o que abre as portas a interpretações subjectivas, do tipo que queremos afastar da classificação de textos a serem incluídos em corpora. Uma investigação objectiva das características linguísticas que compõem o tema de um texto é uma opção realista, na actualidade, dado o advento dos corpora informatizados, e do sofisticado software de análise. Com este tipo de análise objectiva, Phillips salienta que, não só todos os resultados provêm do próprio texto, sem interferências do exterior, como se trata de um processo que pode repetir-se para qualquer número de unidades linguísticas (para Phillips, equivalentes a capítulos, mas que poderiam também representar qualquer dimensão dentro de um texto).

Phillips explica ainda que a qualidade «acerca de» de um texto não dependerá directamente de questões linguísticas formais, já que qualquer leitor pode sempre resumir um texto, e explicar aquilo de que o texto trata, sem fazer referência directa a itens lexicais específicos nele utilizados. Todavia, em última análise, a compreensão de um texto e a evocação do seu tema podem ser retirados apenas do próprio texto. O tema de um texto parece ser independente da sua representação linguística, mas, ao mesmo tempo, é evocado pela nossa consciência apenas através do simbolismo da linguagem. Phillips ilustra, portanto, a nossa necessidade de olhar aos padrões globais num texto, sendo através da análise de tais padrões globais que teremos acesso ao seu tema. Tal standardização global constitui a macroestrutura do texto, e é a este nível superior, ao nível da macroestrutura, que o texto é analisado através de software sofisticado, para dele se extrair informação sobre o seu tema.

No âmbito deste relatório é, pois, importante, determinar, em primeiro lugar, as características relevantes para o conceito de tema, ou assunto e, em segundo lugar, discutir as possibilidades de análise de tais características. Análises a nível da macroestrutura têm sido executadas, com sucesso, no âmbito do projecto Aviator (Universidade de Birmingham), projecto esse que funcionou como uma continuação informática dos resultados do trabalho desenvolvido por Phillips.

Phillips afirma que «em qualquer discussão relacionada com o 'tema', a noção de 'situação' é importante, e é provável que relações

distintas entre os textos e os seus contextos possam constituir a base de uma tipologia textual». Este objectivo «Firthiano» situa solidamente o texto em relação a um nível de restrição mais elevado, isto é, o contexto situacional. Quer os textos, quer os contextos situacionais são demasiadas vezes classificados através de métodos externos, e associados a categorizações que reflectem, principalmente, uma interpretação e uma classificação subjectivas do nosso meio. Se o tema de um texto depende da sua relação com o contexto, devemos, então, determinar quais são as características relevantes, analisando-as objectivamente, através de uma análise informática, de modo a evitar distorções e imposições desnecessárias.

Phillips defende que, embora o tema de um texto não dependa directamente da representação linguística, os conceitos da linguagem são, em última análise, construídos por itens lexicais, ou associações de itens lexicais. Para este autor, o tema de um texto pode ser determinado por padrões a um nível superior da análise. Os padrões de um texto constituem aquilo a que Phillips se refere como sendo a sua macroestrutura, a partir da qual se pode chegar às características que compõem um tema:

I contend that with all semantic, syntactic and lexical markers of a science text neutralised, the meaning of the text is not exhausted but that a distributional analysis of what remains will reveal the presence of global patternings, which I call macrostructure.

Analisa-se a macroestrutura de um texto através da observação das associações de tipo combinatório presentes nesse texto, e que revelem ser características de um certo estilo, tema, ou género. A metodologia de Phillips é uma análise distribucional de textos (tratando-se de uma análise de tipo distribucional, recomenda-se a inclusão de textos completos no corpus, para que todas as características linguísticas do texto estejam representadas). O primeiro passo na metodologia de Phillips consiste na identificação da frequência de ocorrência de cada par nó/co-ocorrente numa amplitude de quatro palavras de cada lado do nó. O resultado é achado ao tomar-se cada palavra no corpus como nó, e as palavras na amplitude determinada como co-ocorrentes. As frequências dos co-ocorrentes foram, então, comparadas com o nível do lema (embora a justificação para esta decisão seja agora discutível). Associações particulares entre lemas foram identificadas ao longo dos capítulos (pertencentes, aqui, a livros científicos), e a diferença na

macroestrutura de cada um foi avaliada. Desta forma, a macroestrutura foi definida através da análise dos padrões intercombinatórios.

Phillips emprega o método de Ward de análise de grupos para observar as atracções entre lemas, ampliando a noção de associação para que esta possa incluir associações que, normalmente, têm, pelo menos, dois lemas em comum. Depois de estabelecido um limiar de separação (de modo a evitar que os dados não se reúnam, na sua totalidade, num só grupo), a cada nó é alocado um 'coeficiente de similaridade', baseado no seu ambiente combinatório. Os nós com 'coeficientes de similaridade' aproximados são reagrupados, e o processo repetido com o novo grupo. Assim se analisa, portanto, a similaridade de redes de lemas, que formam a macroestrutura do texto. Para duas redes serem similares, devem conter pelo menos dois lemas iguais, e pelo menos um lema deve ser membro do conjunto nuclear de nós que compõe cada rede.

A análise demonstrou a existência de padrões de associação entre os capítulos. Phillips indica como as duas maiores descobertas do seu estudo, primeiramente, a descoberta da existência de «conjuntos lexicais sintagmáticos» e, em segundo lugar, o facto de «os conjuntos identificados na análise serem significativos» se interpretada a organização sintagmática das palavras em conjuntos lexicais estabelecidos pela análise como sendo as linhas de orientação conceptuais do texto.

Esta análise está na base da noção de macroestrutura lexical e da descoberta de restrições combinatórias; a interpretação semântica de tais factores levar-nos-ia à compreensão daquilo de que trata o texto.

O trabalho de Phillips constituiu a base para o sector do projecto AVIATOR, coordenado por Antoinette Renouf na Universidade de Birmingham, que pretendia desenvolver software para monitorizar a existência de padrões na língua. O projecto tinha dois objectivos principais: um, o de monitorizar o estado de permanente evolução da língua, isto é, identificar palavras e combinatórias novas; o outro, extremamente relevante para o tema deste artigo, o de identificar grupos de palavras num texto que reflectam, até certo ponto, o seu conteúdo conceptual. Dado que o objectivo inicial era o de monitorizar principalmente uma evolução, dispos-se de um fluir constante de dados linguísticos (provenientes do jornal nacional *The Times*) ao longo dos três anos de duração do projecto.

O software desenvolvido no âmbito do projecto Aviator é útil para a identificação do tema, ou qualidade «acerca de» de um texto. Foi desenhado para identificar padrões no texto e similaridades de padrões intra-textuais o mais automaticamente possível. A metodologia seguida foi praticamente a mesma de Phillips. Uma das primeiras fases foi o estabelecimento de uma lista de palavras a excluir, constituída pelas palavras mais frequentes no corpus, e que incluía as palavras consideradas como «não-lexicais». Esta lista foi gerada automaticamente. Tais palavras foram, portanto, abandonadas, e consideradas não-significativas ou indicativas da qualidade «acerca de» de um texto. Criou-se, em seguida, um banco de combinações formado pelas palavras restantes.

O programa de agrupamento de palavras processou, então, os nós mais frequentes, de modo a criar grupos, a partir dos quais se pode identificar a qualidade «acerca de» de um texto. O resultado da aplicação deste programa era também comparado com os resultados atingidos por um grupo de pessoas a quem se tinha atribuído a tarefa de resumir um texto ou dizer de que tratava. O programa criou listas e grupos de palavras-chave indicativos do tema do texto. Com base nos resultados do projecto Aviator, parece possível determinar o tema de textos através da análise de critérios linguísticos com software como o utilizado por Phillips (1983) e Renouf et al (1993). Uma tipologia textual efectuada com base nos temas resultantes deste tipo de análises poderia, então, ser estabelecida.

Não se prevê, no entanto, que os tipos de tema identificados através desta metodologia constituam algum dia um número finito. Na sequência da investigação no âmbito do programa Aviator, seriam adicionados dados ao corpus sempre que possível. As diferentes áreas temáticas identificadas teriam, portanto, que acompanhar um modelo dinâmico de investigação linguística. Se, tal como aqui se recomenda, passarmos a dar maior relevância aos corpora monitores, é do nosso interesse utilizar software para classificar textos da maneira mais automática possível, bem como desenvolver software que possa ser utilizado em várias línguas.

O tema de um texto – aquilo de que o texto trata – é, provavelmente, exclusivo desse texto e, em última análise, trata-se do próprio texto. Todavia, por mais exacta que seja, esta constatação não é, na prática, muito útil, já que, por motivos vários, é necessário contar com um meio de associar, ou diferenciar, textos, de acordo com o tema.

Esta necessidade tem sido desde há muito implicitamente reconhecida, e é vulgar verem-se classificações textuais baseadas em grosseiros sistemas de identificação de palavras-chave. É comumente solicitado aos autores que forneçam, juntamente com os seus manuscritos, um pequeno número de palavras (incluindo curtos sintagmas), através das quais o texto possa ser indexado. É muito provável que tais palavras e sintagmas apareçam proeminentemente no próprio texto, o que faz deste método uma análise grosseira e imediata de critérios internos.

Em relação com o método das palavras-chave está um método de classificação muito mais sofisticada – o do resumo, também rotineiramente fornecido pelos autores. Um resumo é normalmente um curto texto que ilustra a mensagem principal contida no texto total; assim, é provável que contenha linguagem característica deste último.

Na época dos corpora de dimensões ilimitadas, que incluem textos completos de dimensões variadas, é necessário que os critérios internos sejam de natureza automática e formal. Resumir um texto é dispendioso e lento, quando comparado com a velocidade de processamento disponibilizada por uma máquina. Assim, recomendamos que operações como o resumo e o fornecimento de palavras-chave sejam desenvolvidas, para a classificação temática interna.

As palavras-chave fornecem um conjunto de etiquetas ou títulos para indicação do que se encontra no interior de um texto, e do modo como esse material se relaciona com outros textos. O resumo dá indicações quanto ao tipo de frases e argumentos que aí se podem encontrar. É isto que as pessoas pretendem e julgam de fácil entendimento, embora a relação com os textos originais não seja estabelecida com prontidão.

No entanto, o resumo automático é objecto de um certo número de projectos de investigação, dos quais destacamos um tipo em particular, em que a máquina selecciona e agrega frases do texto original para constituir o resumo (Hoey, 1991 pp. 113-4, 142, 160). Nestes sistemas, a relação entre texto e resumo é simples e explícita.

A selecção automática de palavras-chave é outro passo óbvio, e aqui a noção de Phillips de «acerca de» tem uma importância prática considerável. Em princípio, usar esta noção poderia levar a uma hierarquia de classificação textual que produziria, ao seu mais «alto» nível, palavras-chave aceitáveis, seguidas de uma pirâmide de análise



lexical cada vez mais detalhada. Os textos poderiam ser comparados automaticamente e a sua sobreposição poderia ser apresentada em termos de fácil entendimento para o homem.

O conjunto da Pirâmide Temática e de um resumo abstracto de uma dada granularidade fornecerá os instrumentos mais adequados à navegação de corpora e outros arquivos. Embora nenhum destes instrumentos esteja ainda completamente disponível, a investigação nesse sentido é vigorosa. Tratar o tema através do método «acerca de» tem a vantagem de não se ser dependente da linguagem; não se sabe ainda se um sistema de resumo independente da linguagem dará resultados suficientemente bons para permitir a sua utilização.

O trabalho de Phillips relaciona-se, em primeiro lugar, com a linguagem especializada da ciência e da tecnologia, onde se assume que há menos variedade vocabular do que no texto em geral. Paralelamente a este encontra-se o trabalho de Yang Hui-Zhong (1986), que construiu métodos para a detecção de termos técnicos em texto livre. Dado que a estratégia envolve a distribuição de termos através de textos, é de aplicação particularmente desejável em corpora, e embora o estudo publicado diga respeito ao termo individual pode, inversamente, ser utilizado para classificar textos de acordo com a distribuição de termos técnicos neles contidos.

### **Recomendação Provisória – Tema**

É provável que transcorram alguns anos até que sejam criados métodos automáticos de atribuição temática, testados em material suficiente e variado, proveniente de várias línguas, e acreditado por um organismo como o EAGLES. Até lá, a comunidade utilizadora de corpora deve concordar numa posição interna que permita aos investigadores a compreensão e utilização dos respectivos trabalhos, evitando esforços inúteis. Depois do estudo e da análise das práticas correntes, propõe-se a seguinte lista, que indica o nível apropriado de classificação e itens organizadores apropriados; quando aplicada, deve ser variada e aumentada de acordo com as prioridades dos investigadores:

- I.1. a vida mental
- I.2. cultura
- I.3. o mundo físico
- I.4. coisas vivas

- I.5. sociedade
- I.6. manufactura
- I.7. comunicações

## **Estilo**

«Estilo» é um termo conspícuo, dado que é usado de maneiras muito distintas por investigadores de disciplinas muito diversas, possuindo, além do mais, significados populares. É utilizado aqui como uma maneira de diferenciar textos sem recorrer ao tema, principalmente através da escolha da presença ou ausência de um vasto leque de características estruturais e lexicais. Algumas destas características estão em distribuição complementar (por exemplo, verbos na voz activa ou passiva), e algumas são preferenciais, por exemplo, formas de cortesia e de eufemismo.

Tal como no caso do tema, não existem esquematizações institucionalizadas que possamos considerar. Embora muito se diga sobre o estilo, e existam vários parâmetros organizacionais, nenhum destes últimos apresenta modelos acreditados. Assim, por exemplo, muitos estudiosos da língua pensam ser necessário existir um parâmetro de formalidade, e termos como «formal», «informal», «coloquial», etc. são comumente utilizados, apesar de nem sempre bem definidos. Numa das propostas mais influentes (da sua época), Martin Joos estabeleceu cinco níveis (1961) – gelado, formal, informal, coloquial, íntimo. Esta classificação provou ser muito útil, mas Joos podia ter igualmente avançado com quatro, seis, sete ou vinte níveis, dado que a motivação para serem cinco foi, principalmente, da esfera da conveniência.

Halliday (1964) distingue diferentes registos baseando-se em três dimensões: «área de discurso», «modo de discurso» e «estilo de discurso» (em trabalhos posteriores transformaria a etiqueta «estilo» em «teor»). O estilo é aqui utilizado com referência às relações entre os participantes. Halliday sugere uma primeira distinção entre coloquial e cortês, distinção essa que é adoptada por muitos dos dicionários actuais. Afirma também que os estilos de discurso devem ser tratados como um contínuo, com categorias tais como «informal», «íntimo» e «deferente».

Muitos dos dicionários existentes na actualidade classificam certas palavras de acordo com o estilo. A distinção mais comum é entre

linguagem formal e informal. Mesmo dentro de categorias tão abrangentes como estas, é possível que um dicionário classifique determinada palavra como formal, e que outro o não faça. Dado que tais decisões pertencem ao lexicógrafo, não é surpreendente que existam inconsistências entre as várias interpretações do estilo de uma palavra. «Formal» é, no Longmans (1993), uma etiqueta que «normalmente significa tratar-se de uma palavra provavelmente utilizada em escrita muito formal ou em discurso formal lido». De facto, estas duas categorias são normalmente definidas pelo tipo de contexto ou situação nos quais a linguagem se encontra, por exemplo, requerimentos, declarações e outros documentos oficiais, etc.

Alguns dicionários (por exemplo, o Collins Robert 1987) possuem graus de formalidade ou informalidade, dando origem a sub-grupos de estilo dentro das etiquetas formal e informal. Encontramos, por exemplo, uma classe para palavras que, «não fazendo parte da norma, são usadas por todos os falantes educados em situações de descontração, mas que não seriam usadas numa composição formal ou numa carta, nem numa ocasião em que o falante quisesse causar boa impressão». Num lugar inferior da escala, encontram-se palavras que «indicam que a expressão é usada por alguns, mas não por todos os falantes educados, numa situação de grande descontração». Em que consiste uma «situação de descontração», e será que significa a mesma coisa para todos estes «falantes educados»?

No Webster's Dictionary (1976) as etiquetas de estilo são «calão», «não-normativo» e «sub-normativo», cada uma definida pela sua relação com um padrão ou norma. A maioria dos dicionários coloca etiquetas de estilo nas palavras que não fazem parte da norma. Alguns chegam até a atribuir etiquetas para identificação da linguagem específica em que a palavra é predominantemente utilizada. Sansoni (1988) oferece um vasto leque de etiquetas de estilo, que situam as palavras em diferentes linguagens específicas, como por exemplo, *linguaggio burocratico, commerciale, cinematografico, giornalistico, infantile, marinaro, scolastico, universitario*, etc. – a lista é substancial.

O estilo «literário» é, também, uma classe frequentemente presente em dicionários. Paralelamente, encontramos também o «estilo «poético», mas o modo como ambas as etiquetas se relacionam não é claro. Será o estilo poético uma sub-classe do estilo literário? Podemos

assumir que ambos podem incluir-se na classe «não-normativa» do Webster's? Classificar-se-ia, então, a classe «calão» abaixo da etiqueta «sub-normativo», com «informal» a servir de ponte entre as duas? Se assim fosse, onde colocaríamos a categoria «ofensivo», ela própria uma etiqueta de estilo comum? Estaria no extremo oposto a «cortês»? No entanto, sabemos que, segundo Halliday, a linguagem se divide entre «coloquial» e «cortês».

Parecem, portanto, existir tantas etiquetas de estilo diferentes utilizadas em dicionários, que se torna difícil perceber a maneira como se relacionam entre si, no sugerido contínuo entre coloquial e cortês, informal e formal. Há todo um grupo de categorias que acompanham a distinção informal/formal, embora a estrutura do valor hierárquico de cada uma não seja clara. Outras categorias de estilo utilizadas, tais como calão, coloquial, vulgar, cortês, rara, popular, não parecem nunca ser definidas pelos dicionários que as empregam, como se as suas fronteiras fossem de compreensão automática.

Que a linguagem não se mantém suficientemente estática para atribuir etiquetas de estilo fixas a palavras individuais, é uma certeza.

Na lexicografia prática, em que os compiladores diariamente tomam decisões relacionadas com a classificação estilística, é claro que não existe nenhum consenso automático quanto à formalidade de determinada expressão. Sendo as pessoas diferentes umas das outras, é natural que tenham diferentes padrões de julgamento, e atitudes diferentes quanto ao que é ou não próprio no uso da linguagem, bem como diferentes experiências linguísticas, às quais idade e região em que foram educadas não são alheias; não têm, além disso, a certeza quanto ao modo de aplicação da terminologia – é bom, mau, ou neutro ser coloquial? Quando se deve usar calão?

Há, também, uma outra questão latente, que nunca se resolve por si só: algumas escolhas feitas na escrita são consideradas informais, mas as mesmas escolhas, no oral, não passam de neutras. Será que há um deslocamento da língua falada em relação à língua escrita, que coloca a primeira sempre alguns pontos abaixo da segunda na escala da formalidade? Respostas definitivas a tais perguntas requerem um alinhamento de critérios externos e internos pelo qual devemos ainda esperar algum tempo, contribuindo isso, no entanto, para indicar a ausência de clareza que caracteriza a descrição do estilo.

O estilo é outro dos critérios que recomendamos sejam identificados através da análise de critérios internos. Mais uma vez afirmamos que tal se deveria alcançar recorrendo a meios automáticos, através da análise estatística de textos, do tipo da já utilizada pela linguística forense, nas questões de atribuição de autor. O que é importante para a tipologia textual é que os critérios utilizados na categorização estilística sejam estabelecidos objectivamente.

Biber (1988) propõe uma metodologia para o agrupamento objectivo de variações em textos ingleses através de análises estatísticas. A análise é baseada na identificação de diferentes grupos de características linguísticas num conjunto de textos de língua inglesa escrita e falada. Utilizando uma técnica designada por análise factorial, Biber identifica com sucesso as características linguísticas de textos através de uma análise objectiva de dados linguísticos. A sua afirmação fundamental preconiza que a co-ocorrência frequente de um determinado grupo de características linguísticas indica uma função subjacente partilhada por essas mesmas características. Assim, podemos identificar quais as características que co-ocorrem e quais as que estão em distribuição complementar. Podemos interpretar estes resultados de forma a estabelecer uma correlação entre variações em grupos linguísticos e função. Além disso, podemos utilizar este tipo de análise para definir objectivamente um conjunto de textos que pertençam a cada variante na língua inglesa, para, com esta classificação técnica, poder categorizar novos textos a incluir num corpus. Pensamos, no entanto, ser aconselhável sofisticar continuamente, e não fixar, as categorias, de modo a dar conta do fluir contínuo da língua para o interior e para o exterior do corpus.

Biber, que se propõe identificar variações fundamentais entre o oral e o escrito, trabalha com textos extraídos do corpus LOB (Lancaster-Oslo-Bergen), de inglês escrito, e com o corpus London-Lund, de inglês falado. Estes textos tinham já sido classificados em vários géneros aquando da sua inclusão nos corpora referidos, provavelmente através da utilização de critérios externos. As categorias exibidas pelo corpus LOB incluem géneros como reportagem jornalística, editoriais, recensões jornalísticas, religião, artesanato e hobbies, saber popular, biografias, etc.; no corpus London-Lund as categorias são comunicação directa, conversa telefónica, discursos planeados, transmissões rádiofónicas ou televisivas, etc.

Biber distingue «género», que utiliza para descrever a classificação obtida por critérios externos, de «tipo de texto», em que se refere ao agrupamento de textos de forma linguística similar, independentemente das categorias genéricas a que estes possam pertencer.

O primeiro passo, na metodologia de Biber, para a definição do tipo de texto através de critérios internos consiste em rever qualquer investigação de características linguísticas que, por sua vez, identifique características potenciais importantes. Aqui, o objectivo não é estabelecer quais são as características linguísticas mais importantes (já que isto deve ser feito objectivamente, através de uma análise estatística dos dados), mas, antes, apresentar um leque o mais lato possível de características linguísticas significativas. Biber identifica 67 características linguísticas de texto inglês que inclui na análise (cf. *ibid.* pp.86-87 a lista completa). Tendo seleccionado os textos com que se irá trabalhar, podemos, então, obter listas de frequência da totalidade das 67 características em cada um dos textos utilizados na análise. Estes números devem, depois, ser normalizados para um dado comprimento de texto (aqui, 1000 palavras), de modo a assegurar compatibilidade. As frequências média, mínima e máxima, a distância entre elas e o desvio-padrão são igualmente calculados antes de se dar início à análise factorial. Características com uma frequência de ocorrência muito baixa são abandonadas, por serem consideradas insignificantes.

O objectivo da análise factorial consiste em identificar grupos de características linguísticas que co-variam. Dizer que co-variam não é necessariamente dizer que co-ocorrem, e sim que há, de facto, uma correlação (ou correlação inversa) entre os seus valores de frequência nos textos. Isto significa que é tão importante saber que duas características co-ocorrem, como saber que duas outras características estão em distribuição complementar, ou que a presença de uma aponta para a ausência da outra.

Desta forma, Biber estabelece factores, constituídos por conjuntos de diferentes critérios linguísticos.

O factor incluirá quer «pesos» negativos (indicam características linguísticas cuja presença marca a ausência de outras), quer «pesos» positivos (indicam características linguísticas que co-ocorrem). As características mais significativas são as que exibem o maior «peso», indique ele atracção ou repulsão. Biber acaba por estabelecer 7 factores representativos dos textos em estudo, todos eles, é óbvio, em distribuição complementar.

Tratar-se-á agora da interpretação destes factores, identificados através da análise textual acima referida, e que serão, por seu turno, utilizados para identificar estilos de texto. Até agora, a identificação das características, o cálculo das suas frequências e a extensão da sua co-variação foram objectivos e automáticos. Os factores estabelecidos são, portanto, interpretados por Biber para ilustrar as funções comunicativas associadas a cada factor (cf. Biber, 1988, pp.104-114, para uma interpretação detalhada dos factores identificados). Por uma questão de conveniência podemos apresentar aqui alguns exemplos do tipo de resultados obtidos.

Biber descobre, por exemplo, que no primeiro factor há um alto peso de nomes, verbos pessoais, presentes, pronomes, interrogativas em QU, etc. Estas informações são-nos dadas pela estatística. Seguidamente, Biber interpreta estes padrões de características linguísticas em termos do estilo do texto. Uma elevada densidade de nomes (os principais portadores de significado referencial num texto), por exemplo, indica uma alta densidade de informação. Palavras mais compridas sugerem significados mais específicos e especializados, e a *ratio* palavra/ocorrência também aponta para uma alta densidade de informação, assim como para uma escolha lexical muito precisa, daí resultando uma apresentação exacta de conteúdo informacional. Encontramos também um estilo interactivo, visto existir um peso elevado de formas no Presente, indicando um estilo mais verbal do que nominal, acompanhado pela presença de pronomes e interrogativas em QU. A interpretação continua, mas este não é o local apropriado para entrar em tais pormenores. O que é importante é frisar a importância deste tipo de análise para a identificação do estilo de um texto.

Biber relaciona o factor com o texto, ao qual é, então, atribuída uma pontuação factorial. Isto permite relacionar os agrupamentos linguísticos do primeiro factor com, por exemplo, dois parâmetros comunicativos individuais: em primeiro lugar, aquele que pressupõe que o objectivo principal do escritor/falante é informacional, apesar de acompanhado de um estilo interactivo, afectivo e implicado e, em segundo lugar, que as circunstâncias de produção são de ordem a permitir cuidadosas possibilidades de edição, possibilitando escolhas lexicais precisas e uma estrutura textual integrada.

As dimensões identificadas por Biber são, então, relacionadas com os géneros já pré-estabelecidos para a classificação dos textos dos

corpora utilizados. É óbvio que os géneros não são coerentes em termos das suas características linguísticas. Algumas aplicações podem empregar-se simultaneamente com vários géneros, enquanto que dentro de um género determinado pode haver uma grande variação. Como exemplo, Biber refere que na «prosa académica» existe um vasto leque de variação, tal como no género «conversa», o que indicaria que não podemos encarar estes géneros como representativos respectivamente do inglês escrito e do inglês falado, assim como não podemos usar a análise de um só texto como representante de todo um género. O estudo demonstra também que não há qualquer diferença única e absoluta entre a escrita e o oral, existindo antes várias dimensões de variação manifestadas em ambos os tipos de discurso. As dimensões identificadas neste estudo conseguem definir um conjunto de relações entre textos que podem ser usadas numa tipologia textual global. Tal como afirma Biber, já que os textos em estudo cobrem um largo âmbito de tipos de discurso em inglês, assim como as características linguísticas de muitas funções comunicativas, as dimensões de variação por ele estabelecidas fornecem os parâmetros de variação linguística nos textos de língua inglesa como um todo.

As implicações de uma tal investigação são muitas. A metodologia aqui esboçada foi já aplicada a vários outros tipos de investigação, nomeadamente à comparação de dialectos do inglês, e à identificação da variação linguística entre o inglês britânico e o inglês americano, através da análise de características lexicais e sintácticas; fizeram-se também comparações estilísticas, não só entre autores dados, mas também no âmbito da evolução histórica do texto escrito na língua inglesa. Particularmente relevante na discussão do estilo é a investigação estilística efectuada por Biber e Finnegan (1988), que utilizam a técnica aqui delineada em conjunção com a análise de grupos vocabulares, de modo a identificar oito estilos de atitude em textos ingleses, por exemplo, «Cautelosa», «Afastada da Disputa», etc. No âmbito multilinguístico, as consequências deste tipo de análise são importantes, já que esta abordagem da análise textual pode ser utilizada noutras línguas, podendo-se assim contar com tipos de texto comparáveis de língua para língua.

Biber trabalha com um número considerável de características linguísticas; ao contrário, Nakamura desenvolveu técnicas para examinar o papel classificatório de características individuais, utilizando o



Método Quantificacional de Hyashi, Tipo Três. Numa série de artigos (1986, 1987, 1992, 1993) explicou como textos, gêneros e corpora podem ser diferenciados de acordo com a incidência de determinadas características.

A base do método de Nakamura está no modo através do qual um elevado número de observações individuais ou classificações da língua podem ser agrupadas de modo a revelar tendências gerais bastante latas. Embora o processamento estatístico possibilite a existência de 14 diferentes parâmetros de agrupamento (designados por «eixos»), na prática são os três primeiros que dão normalmente conta da maior parte dos dados; além disso, dado que os resultados se interpretam mais facilmente quando apresentados sob a forma de um diagrama, três dimensões constituem a mais complexa estrutura que se pode processar adequadamente. Os diagramas de Nakamura mostram graficamente as distâncias relativas dos itens linguísticos entre si, ou as distâncias relativas entre textos ou corpora que contenham esses itens.

As técnicas de Nakamura podem ser usadas para classificação num elevado número de circunstâncias; a distribuição dos pronomes nos textos, a distribuição de etiquetas gramaticais, ou a distribuição do vocabulário são possíveis; pode trabalhar-se com muito ou pouco material – Nakamura trabalha com corpora que vão de 1 a 200 milhões, aplicando quer um só critério, quer uma combinação complexa de critérios.

### **Recomendação provisória: estilo**

Tal como se referiu acima, um pequeno número de parâmetros estilísticos é frequentemente invocado, e servirá de classificação interna. Estes parâmetros são apresentados seguidamente, sem definição, e o utilizador deve adicionar os seus próprios usos dos termos empregues.

#### **I.2. estilo**

##### **I.2.1. Formalidade**

###### **I.2.1.1. informal**

###### **I.2.1.2. formal**

##### **I.2.2. Preparação**

###### **I.2.2.1. pensada**

###### **I.2.2.2. espontânea**

- I.2.3. Agrupamento comunicativo
  - I.2.3.1. grupo conversacional
  - I.2.3.2. falante/escritor e público
  - I.2.3.3. públicos remotos (p.ex. rádio, TV)
- I.2.4. Direcção
  - I.2.4.1. unilateral
  - I.2.4.2. interactiva

Estas categorias são, de forma geral, externas; tal como se refere acima, tais categorias genericamente aceites não são claramente definíveis, estando indubitavelmente relacionadas com pormenorizadas selecções de gramática e fraseologia ainda não descobertas. Um factor recém-chegado às categorias estilísticas é mencionado em seguida.

### **Combinatória**

Regista-se actualmente um grande interesse por esta característica linguística, constituída pela co-ocorrência de palavras num curto fragmento de texto. A combinatória é ao mesmo tempo linguisticamente poderosa e fácil de identificar através de processamento informático. Note-se que a combinatória é usada como critério tanto no caso do tema, como no do estilo, o que pode ajudar a explicar a sua popularidade.

No caso do tema, o agrupamento de co-ocorrentes ajuda a desambiguar as palavras individuais, e possibilita uma identificação mais exacta do tema do texto do que a que é disponibilizada por simples palavras-chave. Mais recentemente, a combinatória tem sido utilizada para classificar géneros, revelando que a mesma palavra está caracteristicamente associada a certos co-ocorrentes em determinados tipos de língua escrita e falada.

### **Língua falada**

(Uma grande parte desta secção foi retirada da primeira versão do Relatório do Sub-grupo de Corpora de Língua Falada do Corpus EAGLES, de Maio de 1995. O autor deste Relatório é o prof. Joaquim Llisterri da Universidade Autónoma de Barcelona, a quem os autores do presente relatório desejam agradecer.)

A maioria dos corpora gerais inclui um determinado número de transcrições de língua falada – a transcrição é, por vezes, ortográfica. Uma quantidade substancial de dados provenientes de língua falada, em especial gravações de conversas espontâneas entre pessoas, é encarada como uma das mais ricas fontes de conhecimento da língua.

Existem corpora constituídos apenas por material proveniente de língua falada, assim como também encontramos corpora restritos que apenas incluem língua escrita. Há também corpora bastante diferentes do habitual (por vezes designados por corpora de fala, para frisar a distinção), compilados por investigadores especializados em fonética.

A distinção crucial faz-se entre um corpus de língua falada que possa ser corrido em paralelo com um corpus de língua escrita, de modo a fornecer evidência geral sobre a gramática, o léxico, a fraseologia e o estilo de uma língua, e um corpus de língua falada, normalmente conhecido por Corpus de Fala, que é constituído para aprofundar a investigação da Comunidade Falante quanto à natureza da substância fonética.

Esta questão foi considerada em toda uma sequência de conferências e seminários (cf., por exemplo, Leech et al, 1995), que revelaram existir interesses comuns entre os dois grupos, interesses esses passíveis de aumentar ainda mais, dados a presente direcção da investigação e o desenvolvimento da tecnologia. No entanto, grande parte dos corpora de fala são especializados demais para poderem ser incluídos num corpus geral, e na tipologia de corpora EAGLES seriam considerados corpora especiais (1994 ¶ 2.4.2.2).

Em termos práticos, muitos investigadores constataam a necessidade de tentar incluir nos seus corpora alguma língua falada transcrita, ou de, pelo menos, tentar preparar a inclusão de algum material mais tardiamente. Assim, deste relatório constará uma breve apresentação das questões que devem ser colocadas – quais os aspectos, relativos aos textos falados, a considerar aquando da preparação de um corpus de material escrito.

## **Definições**

Um corpus é uma colecção de fragmentos de língua que são seleccionados e ordenados de acordo com critérios linguísticos específicos, de forma a serem utilizados como uma amostra da língua (ibid. 1994 ¶ 2.1).

Um corpus de língua falada é constituído por gravações de fala, acessíveis informaticamente, e que foram transcritas ortograficamente, ou recorrendo-se a uma notação fonémica ou fonética reconhecida.

### **Fases de Desenvolvimento**

Podem distinguir-se três fases no desenvolvimento de um corpus de língua falada:

**a fase prévia à gravação.** Nesta fase, os objectivos do corpus já se encontram articulados, e deles derivam as especificações dos tipos de gravação pretendidos, que incluem situação física, número e tipo de informantes, situação de comunicação e áreas temáticas a serem cobertas.

**a fase de gravação.** Nesta fase, planeia-se e controla-se a gestão da gravação – natureza de qualquer intervenção, bem como grau de percepção dos participantes de que estão a ser gravados. O equipamento de gravação é especificado, bem como outros pormenores técnicos, tais como a colocação de microfones.

**a fase posterior à gravação.** Nesta fase, a transcrição é especificada e efectuada, sendo em seguida processada para inclusão no corpus.

Nota: Na recolha de material destinado ao estudo directo do sinal de fala, os aspectos acima referidos podem ser substituídos por outros. A captação de situações de comunicação genuínas pode não ser importante, enquanto que ela é central para um corpus de língua falada em geral; as especificações técnicas podem ser muito mais elaboradas; intervenção directa e experimentação são comuns, estando excluídas do material a integrar num corpus geral. A transcrição e a análise da onda sonora podem ser extremamente detalhadas. Os tipos de dados de fala referidos em seguida são considerados corpora especiais na perspectiva de corpora de referência gerais; como tal, não voltarão a ser mencionados aqui.

fonemas isolados lidos em voz alta  
palavras isoladas lidas em voz alta  
frases isoladas lidas em voz alta

- fragmentos de texto lidos em voz alta
- discurso semi-espontâneo (por ex. números, expressões alfanuméricas)
- discurso espontâneo sobre um assunto pré-determinado (p.ex. voltar a contar uma história)
- experiências factoriais (os informantes são gravados em situações muito distantes do comportamento comunicativo normal, por ex., passando informações sobre mapas ou percursos através de uma barreira visual)

### **Informantes**

Idade e sexo são características já solicitadas para a língua escrita. Cf. acima (E.1.1.~1), «entre os 16 e os 60 anos uma indicação geral da década é suficiente; o trabalho efectuado por crianças ou pessoas da terceira idade deve ser identificado como tal, dado que esta característica parece comportar diferenças, a nível linguístico, do texto produzido por adultos em geral».

A abrangência demográfica é importante para estudos de pronúncia, mas marginal num corpus de referência normativo. Pode ser adjunta ao parágrafo E.1.1.~3, *Influência Linguística Anterior*.

### **Equipamento**

O padrão corrente indica o gravador DAT, capaz de uma gravação digital que pode ser processada directamente por computador. Devem ser usados microfones unidireccionais e omnidireccionais, colocados a diferentes distâncias, de modo a permitir a distinção entre enunciados sobrepostos. Espera-se a generalização do uso de aparatos microfónicos.

### **Convenções de transcrição**

O relatório NERC especifica um nível de transcrição adequado à entrada de dados em larga escala. Os informantes são minimamente identificados e utilizam-se convenções simples para segmentos de difícil compreensão, hesitações e pausas, sobreposição de enunciados,

risos, etc. Por motivos de conveniência, estas especificações encontram-se em Anexo.

### Armazenamento e Acesso

As transcrições devem normalmente encontrar-se sob a forma de simples ficheiros de texto, com anotações (cf. ponto acima) claramente identificadas num sistema compatível com o SGML. Existe já software para alinhamento do som digitalmente gravado com a respectiva transcrição ortográfica, sob a forma de protótipo, para várias línguas; espera-se que este tipo de software seja brevemente colocado no mercado, e alargado gradualmente a todas as línguas da Europa.

Não se recomenda o uso de bases de dados para armazenagem de dados provenientes de corpora de língua falada, em virtude das onerosas despesas na informatização do material, e do lento desempenho do software que lhe dá acesso.

### Questões Legais

Gravar língua falada é potencialmente uma invasão da privacidade, razão pela qual cada país tem emitido legislação variada para controlar esta actividade. Podendo a operacionalidade da seguinte afirmação ser desmentida por leis nacionais, gravações sub-reptícias não devem conter nada que possa identificar os seus participantes (excepto a sua própria voz, o que é inalienável). Aos voluntários deve ser pedido que assinem um acto de aceitação de participação.

### Notas

\* Tradução Paula M. Neto.

<sup>1</sup> Em inglês *tabloid*, «jornal de pouca qualidade».

## ANEXOS

### Como ler os quadros

Os quadros que se seguem foram retirados, a traços largos, do relatório NERC (Calzolari et al, 1995), relatório este no qual se analisaram vários projectos de corpus linguístico, com vista a determinar qual o mais praticável desenho de corpus. Entretanto, os corpora constantes deste relatório sofreram processos de actualização, e projectos de corpus que não faziam parte do NERC inicial foram nele posteriormente integrados (caso do British National Corpus e do Survey of English). Os quadros pretendem ilustrar as características comuns aos diferentes sistemas de classificação utilizados em corpora europeus. As categorias apresentadas são as do relatório NERC, em que se interpretaram terminologias provenientes de diferentes projectos de corpus, de modo a reunir as características comuns a todos os corpora. As categorias constantes dos quadros apresentam um terreno comum a muitos ou à maioria dos corpora recenseados. Os quadros contêm categorias «de aplicação geral, e não específicas de determinados corpora».

Tal como sucede no relatório NERC, «+» significa «Explicitamente incluído como tipo de texto», «-» significa «Explicitamente rejeitado como tipo de texto» e «A» refere-se às características «dados administrativos» (por oposição a tipo de texto). Casos a suscitar dúvida são marcados com «?».

# 1. Tipologia Textual

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
<b>gênero literário</b>																														
poesia	-																													
narrativa																														
(auto)biografia																														
romance conto																														
histórico																														
ficção científica																														
humor																														
teatro: drama																														
<b>tema</b>																														
tema																														
<b>meio de difusão</b>																														
livros																														
cartas: corresp.																														
jornais																														
brochuras: folhetos																														
<b>ficção/não-ficção</b>																														
ficção																														
não-ficção																														
<b>estilo</b>																														
distante																														
popular: solene																														
especializado: leigo																														
(=técnico)																														
<b>outros</b>																														
manuais: estudo																														
traduções																														

1. Bou Escrito + Oral	10. Allen Escrito	17. Morales Escrito	25. Biber Escrito + Oral
2. Hoz unspéc.	11. Altenberg Oral	18. Collins et al Escrito + Oral	26. Malaga Escrito
3. Svartik et al Escrito + Oral	12. Birmingham Collection of English	19. Summers (componente selectivo)	27. Malaga Oral
4. Juilland et al Escrito	13. Texts Escrito	Escrito	28. Bank of English Escrito and Oral
5. Kucera et al Escrito	14. Birmingham Collection of English	20. Crowley Oral	29. British National Corpus
6. de Vriendt-de Man Oral	15. Texts Oral	21. Bindi et al Escrito	30. Survey of English
7. Uit den Bogaart Escrito + Oral	16. Gonzalez et al Escrito	22. Martin et al Escrito + Oral	
8. de Jong Oral	17. Staphorsius Escrito	23. Werkgroep Taalbank Escrito + Oral	
9. Lara Escrito + Oral	18. Feldweg Oral	24. Atkins et al Escrito + Oral	



## 2. Tipologia de Temas

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20*
Religião	+											+	+		-	+	+	-		+
Técnica (-ologia)				+	+	-		+	+	+		+		-	+	+		+		
Direito		-	+		+			+	+	+	+	+		-	+	+		+		
Desporto		+			+			+	+	+		+		-	+	+	+	+		+
Arte				+			+	+				+	+		+	+		+		+
Política								+		+		+		-	+	+		+		
História				+	+			+		+		+		+	+	+		+		
Medicina					+			+		+		+		+	+	+		+		
Filosofia					+			+		+		+		+	+	+		+		+
Economia					+			+		+		+		+	+	+		+		
Educação					+			+		+		+		+	+	+		+		
Psicologia					+			+		+		+		+	+	+		+		
Ciência		+	-	+	+		+	+		+		+		+	+	+		+		+
Sociologia					+			+		+		+		+	+	+		+		+
Tempos Livres								+		+		+		+	+	+		+		+
Civilização					+			+		+		+		+	+	+		+		+
Física					+			+		+		+		+	+	+		+		+
Biologia					+			+		+		+		+	+	+		+		+
Matemática					+			+		+		+		+	+	+		+		+
o Lar					+			+		+		+		+	+	+		+		+
Viagens								+		+		+		+	+	+		+		+
Antropologia								+		+		+		+	+	+		+		+
Assuntos Militares								+		+		+		+	+	+		+		+
Meios de Comunicação								+		+		+		+	+	+		+		+
Língua								+		+		+		+	+	+		+		+
Literatura								+		+		+		+	+	+		+		+
Arquitectura								+		+		+		+	+	+		+		+
Moda/Indumentária								+		+		+		+	+	+		+		+
Informática								+		+		+		+	+	+		+		+
Agricultura								+		+		+		+	+	+		+		+
Geografia								+		+		+		+	+	+		+		+
Ecologia/Ambiente								+		+		+		+	+	+		+		+
Trafego/Transportes								+		+		+		+	+	+		+		+
Química								+		+		+		+	+	+		+		+
Finança								+		+		+		+	+	+		+		+

- |                    |                          |                        |   |
|--------------------|--------------------------|------------------------|---|
| 1. Bou             | 6. Lara                  | 11. Morales            | 16. Werkgroep Taalbank  |
| 2. Svartik et al   | 7. Altenberg             | 12. Summers            | 17. Biber   |
| 3. Juillard et al  | 8. Birmingham Collection | (componente selectivo) | 18. Malaga  |
| 4. Kucera et al    | of English Texts         | 13. Crowdy             | 19. Bank of English   |
| 5. Uit den Bogaart | 9. Gonzalez et al        | 14. Bindi et al        | 20. British National Corpus. Apenas se classificam tematicamente as obras não-ficcionais. |
|                    | 10. Staphorius           | 15. Martin et al       |   |

### 3. Tipologia do Oral

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
<b>Situação de Comunicação</b>																
cara a cara	+			+	+	+			+		+		+	+		+
por telefone	+			+	+				+		+		+	+		+
<b>Número de Intervenientes</b>								A			+			+	+	
monólogo	+								+					+	+	
diálogo	+					+			+			+		+	+	+
<b>Monólogo</b>																
conferência				+	+				+		+		+	+	+	
comentário	+		+	+	+		+		+				+	+	+	
discurso				+			+		+		+			+	+	
sermão				+					+					+	+	
apresentação					+				+					+	+	
narração					+	+							+			
<b>Diálogo</b>																
conversa				+	+	+	+		+				+	+	+	+
entrevista		+	+		+	+			+				+	+	+	
discussão/debate				+	+				+				+	+		+
consulta									+					+	+	
organização									+					+	+	
talk-show									+						+	
reunião (profissional ou não)									+				+		+	
<b>Local</b>								A								
educação				+	+				+		+		+		+	
comércio/negócio		+							+				+		+	
rádio		+		+	+				+	+	+		+		+	
televisão					+				+	+	+		+		+	
org. política ou social				+					+				+			
privado/pessoal		+				+			+		+	+				
trabalho									+			+				
institucional									+			+			+	
<b>Espontaneidade</b>											+			+		
com guião								A				+		+		+
sem guião/pré-planeado	+					+		A	+			+	+	+		+
espontâneo/não-planeado	+	+						A		+		+	+	+		+
<b>Tema</b>	+			+	A	+		A	+		+	+	?	+		
<b>Estilo</b>																
distante	+		+		+			A				+				
especializado/leigo (técnico)									+			+				
<b>Intervenientes</b>																
sexo			+		?	+		+	+		?			+	+	
idade			+		?	+		+			?			+	+	
grupo étnico					?	+	A		A							
região			+		?	+		+	+		?			+	+	

A representa 'dados administrativos'.

1. Svartik et al
2. de Vriendt de Man
3. de Jong
4. Altenberg

5. Birmingham Collection of English Texts Oral
6. Feldweg
7. Collins et al
8. Crowdy demographic

9. Crowdy context governed
10. Martin et al
11. Atkins et al
12. Biber

13. Malaga
14. Bank of English
15. British National Corpus
16. Survey of English

#### 4. Aspectos Regionais e Temporais

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
Aspectos Regionais	+	+	+	+	+	+		+	+		+	+	+	+	+	+	+	+	+	+		+	+	+		+			+
Aspectos Temporais			+	+	+	+	+	+				+	+		+	+	+		+	+	+	+		+		+			+

1. Bou Escrito + Oral	9. Lara Escrito and Oral	15. Staphorsius Escrito	23. Werkgroup Taalbank Escrito and Oral
2. Hoz. unspec.	10. Allen Escrito	16. Feldweg Oral	24. Atkins et al Escrito and Oral
3. Svartvik et al Escrito and Oral	11. Altenberg Oral	17. Morales Escrito	25. Biber Escrito and Oral
4. Juilland et al Escrito	12. Birmingham Collection of English Texts Escrito	18. Collins et al Escrito and Oral	26. Malaga Escrito
5. Kucera et al Escrito	13. Birmingham Collection of English Texts Oral	19. Summers Escrito	27. Malaga Oral
6. de Vriendt-de Man Oral	14. Gonzalez et al Escrito	20. Crowley Oral	28. Bank of English
7. Uit den Bogaart Escrito and Oral		21. Bindi et al	29. British National Corpus
8. de Jong Oral		22. Martin et al Escrito and Oral	

## Referências

- ATKINS S., CLEAR J., OSTLER N. (1992) *Corpus Design Criteria*, Journal of Literary and Linguistic Computing, Vol 7, No.1.
- BHATIA V.K (1993) *Analysing Genre: Language Use in Professional Settings*, London and New York: Longman.
- BIBER Douglas (1988) *Variation across speech and writing*, Cambridge: CUP. (1989) *A Typology of English Texts*, Amsterdam: Mouton de Gruyter. (1993) *Representativeness in Corpus Design* in Journal of Literary and Linguistic Computing, Vol 8, No.4, Oxford: OUP.
- BIBER and FINNEGAN (1988) Adverbial stance types in English. *Discourse Processes* 11:1-34.
- CALZOLARI N., BAKER M. and KRUYT P.G. (eds) (1995) *Towards a Network of European Reference Corpora*. Report to the NERC Consortium, Feasability Study, Co-ordinated by A.Zampolli, Pisa: Giardini.
- HALLIDAY M.A.K., MCINTOSH A. and STREVEENS P. (1964) *The Linguistic Sciences and Language Teaching*, London: Longman.
- HALLIDAY M.A.K., J.R. MARTIN (1993) *Writing Science, Literary and Discursive Power* London & Washington: Falmer.
- FIRTH J.R. (1957) A Synopsis of Linguistic Theory, 1930-1955 in *Studies in Linguistic Analysis, Special Volume of the Philological Society*; Oxford. Reprinted in Palmer FR (ed) *Selected Papers of J R Firth 1952-9*, Longman 1968.
- HAZADIAH (1993) Topic as a Dynamic Element in Oral Discourse in *Text and Technology*, Baker M., Francis G., Tognini-Bonelli E. (eds) Amsterdam: John Benjamin.
- HOEY M. (1991) *Patterns of Lexis in Text*, Oxford: Oxford University Press
- HUNSTON S. (1989) *Evaluation in Experimental Research* Article, unpublished PhD thesis, University of Birmingham.
- HUI-ZHONG Y. (1986) A new technique for identifying scientific/technical terms and describing science texts. *Literary and Linguistic Computing* 1, no 2: 93-103.
- JOOS M. (1961) *The Five Clocks*, New York, Harcourt: Brace and World
- LEECH G., MYERS G., THOMAS J. (1995) (eds) *Oral English on Computer: Transcription, Mark-up Application* London: Longman.
- Lyons J (1977) *Semantics* Cambridge: Cambridge University Press.
- Macchi (ed) (1988) *The Collins Sansoni Italian Dictionary*, Firenze: Sansoni.

- NAKAMURA J. (1986) Classification of English Texts by Means of Hayashi's Quantification Method Type III in *Journal of Cultural and Social Science*, College of General Education, University of Tokushima 21:71-86. [(1987)] Notes on the Use of Hayashi's Quantification Method Type III for Classifying English Texts in *Journal of Cultural and Social Science*, College of General Education, University of Tokushima 22:127-145. (1992) *Hayashi's Quantification Method Type III: A Tool for Determining Text Typology in Large Corpora*. An Annex to a general report on Annotation Tools of the NERC Report. Unpublished manuscript. (1993) Statistical Methods and Large Corpora: A new tool for describing text types. *Text and Technology*, Baker M., Francis G. and Tognini-Bonelli E. (eds) Amsterdam: John Benjamin.
- NAKAMURA J. and SINCLAIR J. (1995) The World of Woman in the Bank of English, in *Journal of Literary and Linguistic Computing*. Oxford: OUP.
- NORLING-CHRISTENSEN O. (1995) Design and Composition of reusable Harmonised Written Language Reference corpora for European Languages, draft report for PAROLE project.
- PHILLIPS, M.K. (1983) *Lexical Macrostructure in Science Text*, Ph.D. thesis, University of Birmingham.
- RENOUF A. (1992) Report on AVIATOR's Clustering Methodology and Software, unpublished paper.
- ROGET P.M. (1962) *Roget's Thesaurus of English words & phrases*, London: Longmans.
- SAGER Juan C., DAVID Dungworth, PETER McDonald (1980) *English Special Languages, Principles and practice in science and technology*, Wiesbaden.
- SINCLAIR J. (1994) Corpus Typology report prepared for EAGLES project.
- SPERBERG-MCQUEEN C M and BURNARD L. (eds) (1993) *Guidelines for Electronic Text Encoding and Interchange*.
- SUMMERS et al (1993) *Longman Language Activator*, Essex: Longman.
- Wales, K. (1989) *A Dictionary of Stylistics*, London and New York: Longman.
- WOOLF et al (eds) (1976) *Webster's New Collegiate Dictionary*, Massachusetts: Merriam.