A LÍNGUA PORTUGUESA FACE AO CHOQUE TECNOLÓGICO DIGITAL

António Branco

Sobre o autor

Universidade de Lisboa
Contato: Antonio.Branco@di.fc.ul.pt

Resumo

No presente artigo, apresento uma breve análise da situação da língua portuguesa face ao choque tecnológico digital, os riscos e as oportunidades que neste contexto se colocam, e indico as ações imediatas que surgem como necessárias para se recuperar atrasos e assegurar para o português a sua posição de língua internacional de comunicação com projeção global.

Palavras-chave

Choque tecnológico digital. Tecnologia da linguagem. Língua portuguesa.

1 - Introdução

Tal como qualquer outro dos cerca de 6 000 idiomas existentes no planeta, a língua portuguesa é uma janela para o mundo que nos rodeia. É através da linguagem, e do seu uso em todas as situações do quotidiano, que comunicamos, aprendemos, nos entreajudamos, planeamos o futuro, e fabulamos ou nos comprazemos com uma história ou um poema.

Por paradoxal que à primeira vista pareça, sendo uma janela para o mundo que nos rodeia, a linguagem humana é também e cada vez mais um dos últimos obstáculos comunicacionais com que nos deparamos na era digital e num mundo globalizado.

As novas tecnologias da informação e da comunicação colocam ao nosso alcance pessoas de todo o mundo com quem será fácil interagir, assim como um acervo infindável de informação a que será possível aceder. Contudo, mesmo quando se encontram asseguradas exaustivamente as mais avançadas condições técnicas de acesso, este novo e ilimitado universo de possibilidades continua na sua esmagadora maioria inacessível, encerrado que está dentro das barreiras invisíveis das línguas que o dividem.

De igual modo, as novas tecnologias digitais colocam ao nosso alcance novos e cada vez mais poderosos dispositivos, desde eletrodomésticos até robôs pessoais, que alargarão as capacidades de cada um de nós para limites ainda por explorar. Porém, a utilização generalizada destes dispositivos continua em grande medida condicionada por interfaces idiossincráticas que restringem a sua utilização e limitam de forma drástica a concretização do seu potencial.

No presente artigo, apresento uma breve análise da situação da língua portuguesa face ao choque tecnológico digital, tendo como enfoque os riscos e as oportunidades que neste contexto se colocam.

No presente artigo, apresento uma breve análise da situação da língua portuguesa face ao choque tecnológico digital, tendo como enfoque os riscos e as oportunidades que neste contexto se colocam.

Uma discussão mais alongada e devidamente circunstanciada e referenciada do assunto abordado neste artigo pode ser encontrada em Branco *et al.,* 2012[2]

2 - Linguagem e tecnologia

Tal como em outras áreas da existência humana, a evolução científica e tecnológica tem alterado as condições de utilização, e da própria existência, das línguas naturais ao longo da história. Em alguns casos, estas novas condições de utilização das linguagens resultaram daquilo a que se pode chamar de choques tecnológicos, que estiveram na origem de profundas revoluções civilizacionais.

[1] Por conveniência da leitura, a referência completa a esta publicação é repetida aqui: BRANCO, António, Amália MENDES, Sílvia PEREIRA, Paulo HENRIQUES, Thomas PELLEGRINI, Hugo MEINEDO, Isabel TRANCOSO, Paulo QUARESMA, Vera Lúcia Strube de LIMA & Fernanda BACELAR, 2012, *A Língua Portuguesa na Era Digital / The Portuguese Language in the Digital Age, Coleção Livros Brancos, Berlim, Springer. Este é o volume dedicado à língua portuguesa, de uma Coleção de Livros Brancos sobre as línguas europeias na era digital, que conta com 30 volumes, correspondendo a outros tantos idiomas.*

Um dos primeiros choques tecnológicos envolvendo a linguagem humana de que há registo consistiu no advento da escrita, há cerca de seis mil anos. Esta inovação passou a permitir que os interlocutores comunicassem em linguagem natural de forma assíncrona, sem terem de estar na presença um do outro em simultâneo.

Com a escrita, pode-se dizer que se quebrou a barreira do tempo na utilização da linguagem. Por sua vez, com o advento da imprensa mecânica, há cinco séculos atrás, quebrou-se a barreira social no acesso à informação escrita. As publicações generalizaram-se e deixaram de estar acessíveis apenas para um pequeno grupo de leitores.

Há algumas décadas atrás um outro choque tecnológico para a linguagem natural teve lugar com o advento das telecomunicações. A barreira do espaço na utilização da linguagem foi quebrada, passando então a ser possível aos interlocutores comunicarem de forma síncrona apesar de não se encontrarem presentes no mesmo local.

Todos estes choques tecnológicos na utilização da linguagem natural tiveram impactos civilizacionais enormes, sobejamente assinalados pelos historiadores. O que tem sido porém muito menos assinalado são os seus impactos nas condições de existência das próprias línguas naturais. Estas mudanças tecnológicas têm proporcionado uma evolução tremenda nas condições de comunicação entre indivíduos: possibilitaram alargar as virtualidades da linguagem humana muito para além do que esta permite se restringida apenas a situações de conversa oral face a face. Mas a par dessa evolução, induziram também fortes mecanismos de involução ao nível das línguas.

A cada choque tecnológico, novas condições de utilização da linguagem têm colocado novas e mais estritas condições de existência para os diferentes idiomas. Um exemplo bem conhecido é o das línguas sem sistema de escrita, que foram desaparecendo, e continuam a desaparecer, perante a vantagem de se usar outras línguas, com escrita, que permitam tirar partido dos benefícios dessa superioridade tecnológica.

O progresso tecnológico parece assim exibir uma natureza mefistofélica no que toca à linguagem humana. Ao mesmo tempo que permite alargar a capacidade de comunicação entre indivíduos, cria também fatores de redução da diversidade linguística e do multilinguismo, e dessa forma de redução do património cultural e da pluralidade de mundividências que as diferentes línguas sustentam. Dos cerca de 6 000 idiomas existentes hoje em dia, estima-se que cerca de 2 500 correm o risco de se extinguirem nas próximas décadas, como acontece a cada ano que passa com as várias línguas que morrem com o desaparecimento do seu último falante.

Neste enquadramento, cabe atentarmos no mais recente choque tecnológico para a utilização da linguagem humana, que se encontra em curso nos dias de hoje. Este choque é provocado pela utilização das novas tecnologias digitais e resulta na expansão das condições de utilização das línguas para um novo patamar sem precedentes.

3 - Tecnologia da linguagem

A aplicação das novas tecnologias à linguagem natural, e em concreto o processamento computacional das línguas naturais, está a dar origem a uma nova área de investigação, desenvolvimento e inovação conhecida por tecnologia da linguagem. De um ponto vista genérico, esta tecnologia pode ser caracterizada como se ocupando com a obtenção da representação do significado a partir do processamento computacional de expressões

linguísticas, e vice-versa, da obtenção de expressões linguísticas a partir do processamento da representação do seu significado.

A tecnologia da linguagem engloba duas subáreas, conhecidas por processamento da linguagem e por processamento da fala. O processamento da fala ocupa-se com a obtenção de uma representação discreta a partir de um sinal analógico correspondente a enunciados orais, e vice-versa. O processamento da linguagem, por sua vez, ocupa-se em mapear entre uma sequência discreta de símbolos linguísticos e a representação do seu significado.

A tecnologia da linguagem está a abrir o leque inaudito de novas condições e de novas oportunidades para a utilização da linguagem natural:

- Ajudará as pessoas a comunicar entre si ainda que não falem uma língua comum;

- Apoiará uma nova geração de interfaces naturais e intuitivas, baseadas em linguagem natural, com todo o tipo de dispositivos, desde eletrodomésticos até robôs pessoais.

a. Aplicações

As novas oportunidades propiciadas pela tecnologia da linguagem resultam da sua exploração para o desenvolvimento de diversas aplicações computacionais, que podem afetar de modo diverso as condições de utilização da linguagem natural, e que se encontram atualmente em diferentes estados de maturidade tecnológica e de comercialização.

Algumas destas aplicações são usadas em contextos de utilização profissional especializados, como é o caso, por exemplo, dos detetores de plágio, para o trabalho de avaliação dos estudantes por parte dos professores, ou dos ambientes de apoio automatizado à tradução, para o trabalho dos tradutores.

Algumas outras aplicações são utilizadas de forma mais generalizada, fazendo parte do nosso quotidiano, como por exemplo, os corretores ortográficos.

Certas aplicações da tecnologia da linguagem, ainda que cruciais para o desempenho do sistema global em que se inserem, não são apercebidas pelos utilizadores, como é o caso, por exemplo, dos detetores de linguagem nos motores de busca.

Algumas delas, sendo de interesse para todo o tipo de utilizadores, são de especial relevo para pessoas portadoras de deficiências, como é o caso dos sintetizadores de voz no apoio à leitura por invisuais.

Há aplicações que servem propósitos circunscritos de utilização, como no caso dos sistemas de produção de documentação com base em linguagem controlada, usados na produção de manuais técnicos de manutenção de aeronaves. Outras aplicações servem uma leque mais amplo de situações de utilização, como é o caso dos sistemas de reconhecimento de fala, que podem ser integrados em dispositivos de controlo num automóvel ou suportar sistemas de legendagem automática de vídeos, entre inúmeras outras utilizações.

Face à limitação de espaço do presente artigo e à natureza do seu objetivo, não caberia fazer aqui uma listagem sistemática nem uma apresentação exaustiva de cada uma das possíveis aplicações da tecnologia da linguagem, das suas funcionalidades ou das situações da sua utilização. Para efeitos de alargar a ilustração apresentada até aqui, listam-se em seguida alguns casos mais notórios de aplicações e aqueles cuja designação será em grande medida autosexplicativa das suas funcionalidades. Para ajudar a exposição, estes exemplos serão

reunidos nos seguintes grupos temáticos:

Interação homem-máquina:
- Interfaces com dispositivos e agentes artificiais
- Deteção de linguagem, autor, domínio,…
- Classificação de textos
- Agrupamento de textos
- Busca de documentos
- Extração de informação
- Levantamento de opiniões
- Interfaces com bases de dados
- Resposta a perguntas
- Reconhecimento de fala
- Síntese de fala
- ...

Interação multilingue
- Tradução automática
- Agentes conversacionais
- Publicação multilingue
- **...**

Produção e verificação de linguagem
- Correção ortográfica
- Correção gramatical
- Deteção de plágio
- Linguagens controladas e sistemas de produção de documentação
- Localização de software
- Legendagem automática
- Sistemas de ditado
- Sumarização
- Geração de relatórios
- Ambientes de apoio à tradução
- Simplificação de textos
- ...

Aprendizagem de linguagens:
- Formação
- Avaliação de competências
- …

Web:
- Anotação de metadados
- Busca web avançada
- Gestão de ontologias
- …

Algumas destas aplicações já encontraram não só a sua maturidade tecnológica mas também os modelos de negócio apropriados que permitem a sua ampla difusão e utilização, como acontece com os populares motores de busca de documentos.

Outras aplicações encontram-se ainda em fase de investigação científica ou de protótipo, como é o caso, por exemplo, dos sumarizadores.

Muitas outras aplicações encontrar-se-ão certamente ainda por ser desenhadas e imaginadas consoante a tecnologia da linguagem, que lhes servirá de base, vier a ser

desenvolvida, amadurecida e explorada em novas soluções para os utilizadores.

b. Tecnologia nuclear

A tecnologia da linguagem é assim uma tecnologia facilitadora e para algumas línguas que são objeto de esforços intensos e concertados de investigação, como é o caso notório da língua inglesa, encontra-se em acelerado ritmo de progresso e desenvolvimento.

Quer integrada em sistemas mais amplos quer suportando aplicações autónomas, esta tecnologia é crucial para a revolução tecnológica em curso: ajudará a ultrapassar as últimas fronteiras comunicacionais tanto no que toca à comunicação entre os seres humanos e os agentes e dispositivos artificiais, como no que toca à própria comunicação dos seres humanos entre si.

A investigação em tecnologia da linguagem pode ser vista como se desenrolando em duas dimensões principais inter-relacionadas. Por um lado, envolve o desenvolvimento das chamadas ferramentas de processamento computacional e, por outro, de aquilo que é usual designar-se por recursos linguísticos.

Os recursos linguísticos para uma dada língua são os conjuntos de dados, de diferentes tipos, que são cruciais para apoiar quer a investigação científica sobre essa língua, quer o desenvolvimento e a avaliação de ferramentas de processamento para a mesma.

Como exemplo de um tipo de conjunto de dados dos mais simples, pode-se mencionar os corpora de texto corrido, que consistem em coleções de textos, de domínios diversos, tal como eles foram publicados. Mas mais comummente, os recursos linguísticos são conjuntos de dados altamente complexos que são laboriosamente produzidos de modo a registar e a compilar os aspetos mais sofisticados dos fenómenos linguísticos tal como estes ocorrem ou são instanciados em enunciados usados pelos falantes. Por exemplo, no caso dos léxicos, cada expressão pode ser pormenorizadamente classificada de acordo com as suas várias características linguísticas, desde os aspetos fonéticos até aos semânticos. No caso dos chamados *treebanks*, num outro exemplo, são as frases que são alvo de caracterização exaustiva em termos das relações sintáticas que se encontram instanciadas entre as suas palavras.

Exemplos de recursos linguísticos:

- Corpora anotados
- Corpora multilingues paralelos e alinhados
- Bases de dados de fala
- Listas de palavras
- Abreviaturas, nomes próprios, palavras funcionais,…
- Vocabulários
- Léxicos
- Ontologias lexicais
- Terminologias
- Treebanks
- Propbanks
- DeepBanks
- ...

As ferramentas de processamento para uma dada língua, por sua vez, realizam um leque de diferentes tarefas que, quando encadeadas, contribuem para executar o mapeamento entre forma e significado nessa língua. Essa tarefas incluem desde a funcionalidade mais básica de determinar o início e o fim de frases, por exemplo, até funcionalidades altamente complexas e sofisticadas, como acontece nas gramáticas de processamento linguístico profundo, que permitem obter a representação semântica em forma lógica da frase de entrada.

A título de ilustração, pode-se também mencionar os etiquetadores morfossintácticos — que anotam as expressões num texto com a categoria morfossintáctica (verbo, nome, advérbio, etc) que lhe cabe em cada uma das suas ocorrências —, os lematizadores — que associam a cada palavra a sua forma flexionada canónica, e.g. a forma infinitiva no caso dos verbos —, ou os analisadores de dependências gramaticais — que identificam as funções gramaticais (Sujeito, Objecto Direto, etc) entre as expressões constituintes de uma frase—, entre vários outros exemplos de diferentes tipos e funcionalidades de ferramentas para o processamento linguístico, que lidam com a estrutura, a ambiguidade e a criatividade das línguas naturais.[2]

Exemplos de ferramentas de processamento:

- Separador de frases
- Separador de palavras
- Etiquetador morfossintáctico
- Lematizador
- Analisador morfológico
- Reconhecedor de nomes de entidades
- Desambiguador de aceções de palavras
- Analisador de constituência sintática
- Analisador de dependências gramaticais
- Etiquetador de papéis semânticos
- Gramática para processamento linguístico profundo (análise semântica)
- …

4 - Desafios

Depois das barreiras do tempo e do espaço e da barreira social terem sido quebradas em anteriores choques tecnológicos, é a própria barreira da linguagem, e da diversidade linguística, que se encontra agora a ser ultrapassada com a revolução resultante da tecnologia da linguagem.

Tal como nos choques tecnológicos anteriores, o impacto civilizacional será tremendo e de uma dimensão ainda difícil de antever na sua totalidade. De modo análogo, a par das novas oportunidades abertas, também novos desafios e riscos se colocam agora aos idiomas e à sua existência.

No passado, línguas que em consequência do seu contexto económico e histórico particular, não foram estudadas e, por exemplo, para as quais não foi desenvolvido um sistema de escrita viram-se votadas ao beco sem saída da extinção.

[2] Para experimentar e melhor compreender a funcionalidade de algumas destas ferramentas, é possível recorrer o LX-Center, que apresenta vários demonstradores e serviços linguísticos online gratuitos para a língua portuguesa, em http://lxcenter.di.fc.ul.pt.

Hoje em dia, esse desafio joga-se ao nível da ciência e da tecnologia da linguagem. Línguas que, em consequência do seu contexto económico e histórico, não venham a ser alv¬o de estudo científico e para as quais não sejam desenvolvidas as soluções tecnológicas apropriadas, são línguas que a prazo arriscam a sua progressiva irrelevância, em detrimento de outras melhor preparadas, e eventualmente a sua própria extinção.

Neste enquadramento, importa analisar a actual situação no que diz respeito à língua portuguesa.

Existem 236 milhões de falantes do português em quatro continentes. Este é um número que crescerá para cerca de 335 milhões em 2050 e que faz desta língua a quinta com maior número de falantes no mundo, depois do chinês, castelhano, inglês e árabe. Entre os falantes das suas diferentes variantes existe compreensão mútua generalizada, e se se atender ao critério da compreensão mútua entre falantes para a circunscrição de um idioma, isso pode fazer do português a terceira língua com maior número de falantes no mundo, depois do inglês e do castelhano. Cabe assinalar também que o português é atualmente um dos idiomas de trabalho de 27 organizações internacionais.

A língua portuguesa reúne assim condições ímpares para assegurar o estatuto de língua de comunicação internacional com projeção global.

Esta oportunidade encontra-se porém em preocupante contraste com os riscos que decorrem do baixo nível da preparação do português para a era digital em termos de ciência e tecnologia da linguagem.

No quadro da rede científica europeia de excelência META-NET,[3] foi publicada uma Coleção de Livros Brancos cujos volumes analisam a situação de diferentes línguas na era digital, tendo em atenção a sua preparação em termos de tecnologia da linguagem. Cada volume, elaborado por um grupo de especialistas, é dedicado a uma de 30 línguas europeias. No âmbito desta iniciativa, foi elaborado um estudo comparativo do estado dessas diferentes línguas, tendo sido feita uma classificação apoiada numa escala de cinco níveis, nomadamente "Apoio excelente", "Apoio bom", "Apoio médio", "Apoio fragmentário" e "Pouco/nenhum apoio".

Em resultado desse estudo comparativo, em termos de processamento da linguagem, a língua portuguesa surge classificada como tendo "Apoio fragmentário": surge em situação mais vantajosa que apenas 8 das outras 29 línguas, e em situação igual ou pior que as restantes, entre as quais se inclui o Alemão, o Castelhano, o Francês, o Italiano e o Neerlandês, avaliada como gozando de "Apoio médio", e o Inglês, com "Apoio bom" (Branco *et al.,* 2012, p.36).

Para enfrentar o choque tecnológico digital, a língua portuguesa encontra-se muito menos preparada tecnologicamente que as línguas com que compete por protagonismo no sistema mundial — e inclusive muito menos preparada que outras línguas com muito menor projeção.

Face a estes desafios, são necessárias ações imediatas para que se possam obter progressos importantes para recuperar atrasos e assegurar para o português a sua posição de língua internacional de comunicação com projeção global. Neste enquadramento, cabe assinalar que há uma boa comunidade de centros de investigação, em Portugal e no Brasil,

[3] http://www.meta-net.eu

que cooperam ativamente entre si e que, de momento, têm capacidade instalada para fazer avançar a tecnologia da linguagem para a língua portuguesa. É porém necessário garantir o incremento estratégico do esforço aplicado nesta área para segurar e incrementar esta capacidade e se vir a alcançar um patamar de progresso sustentado.

De modo similar ao que tem vindo a ser feito nas últimas décadas para outras línguas, e em particular e de forma notória para a língua inglesa, este esforço deve ser articulado de acordo com as seguintes medidas estratégicas:

- estabelecer programas de estímulo à **investigação especificamente inter e multidisciplinar na área do processamento computacional da língua portuguesa,** em particular, e da ciência e tecnologia da linguagem, em geral, nas Universidades dos países de língua portuguesa;

- estabelecer programas de investigação e desenvolvimento de longo alcance que fomentem a **construção de recursos linguísticos e de ferramentas para o processamento computacional do português,** e de **aplicações da tecnologia da linguagem capacitadas especificamente para a língua portuguesa;**

- fomentar programas de **cooperação entre os países de língua portuguesa** que promovam a partilha e a transferência de conhecimento e contribuam para um estado de desenvolvimento tecnológico das diferentes variedades do português mais equilibrado;

- promover a **adesão a infra-estruturas de investigação internacionais dedicadas especificamente à ciência e tecnologia da linguagem natural,** como é o caso do CLARIN, a primeira infra-estrutura internacional para a área, criada em fevereiro de 2012.

Referências

BRANCO, António, Amália MENDES, Sílvia PEREIRA, Paulo HENRIQUES, Thomas PELLEGRINI, Hugo MEINEDO, Isabel TRANCOSO, Paulo QUARESMA, Vera Lúcia Strube de LIMA & Fernanda BACELAR. 2012. *A Língua Portuguesa na Era Digital / The Portuguese Language in the Digital Age,* Coleção Livros Brancos, Berlim, Springer, ISBN 978-3-642-29592-8.