O TRABALHO DE ROBERTO BUSA, SJ: ESPAÇOS ABERTOS ENTRE A COMPUTAÇÃO E A HERMENÊUTICA.

Giancarlo Bolognesi (†) – Università Cattolica del Sacro Cuore, Milano.
Luigi Dadda – Politecnico di Milano.
Adriano De Maio – Libera Università degli Studi Sociali, Roma.
Tullio Gregory – Università di Roma "La Sapienza".

Resumo: Um retrospecto das realizações de Pe. Busa durante os 60 anos de seu trabalho na área da linguística computacional: hipertextos internos, a *sistematização* de alografias, lematização, homografias e tipologias: o *sistema léxico*; as leis de economia para grafemas, para tipologias semânticas, para heterogeneidade entre os termos, e dos dois hemisférios lexicais. Por fim, o projeto de *disciplinas linguísticas* é mencionado, uma resposta ao desafio linguístico proposto pela globalização informatizada.
Palavras-chave: Roberto Busa, Index Thomisticus, Linguística computacional, Linguagem.

Abstract: A review of the achievements of Fr. Busa over the course of his 60 years of work in the area of computational linguistics: internal hypertexts, the *systematization* of allographs, lemmatization, homographs and typologies; the *lexical system*; the laws of economy for graphemes, for semantic typology, for heterogeneity among terms, and of two lexical hemispheres. Finally, the project of *disciplined languages* is mentioned, a response to the linguistic challenge resulting from informational globalization.

Keywords: Roberto Busa, Index Thomisticus, Computational Linguistics, Language.

1. INTRODUÇÃO

O aspecto do trabalho de Pe. Busa que mais tem sido enfatizado pela mídia é que ele foi o primeiro a usar computadores para processar palavras e textos e não apenas números. Esse feito de Pe. Busa seria impressionante mesmo se tomássemos em consideração apenas a quantidade e a dimensão de seu trabalho nos sessenta anos de sua carreira. Por exemplo, ele analisou e classificou pelo computador algo em torno de onze milhões de palavras em latim e, ao mesmo tempo, uma quantidade semelhante em vinte outras línguas: Albanês, árabe, aramaico, armênio, tcheco, catalão, hebraico, finlandês, francês, gaélico, georgiano, grego clássico, inglês antigo, italiano, nabateu, português, russo, espanhol e alemão; e o fez em oito alfabetos: árabe, armênio, cirílico, hebraico, fonético (IPA), georgiano, gótico, grego clássico e latim.

Somado a isso, ele participou de mais de cem congressos internacionais em quatro continentes, além de ter organizado um em Tubinga (Alemanha) em 1960. Ele fundou dois departamentos de linguística computacional, um na Universidade Católica de Milão e o outro na Pontifícia Universidade Gregoriana em Roma. E, nos últimos seis anos, tem sido convidado pela Politécnica de Milão para ministrar aulas de filosofia e psicologia aplicadas à inteligência artificial e robótica.

Não obstante, as descobertas e conquistas alcançadas com tamanho empenho não têm sido justamente celebradas, porque muitas delas foram realizadas nas áreas obscuras da pesquisa linguística.

2. QUANDO A IDÉIA NASCEU

1. A idéia de automatizar a análise linguística ocorreu a Pe. Busa entre os anos de 1942 e 1945, os quais foram tempos de guerra, e para ele, o tempo de se preparar para lecionar filosofia na Pontifícia Universidade Gregoriana. Ele não superestima sua descoberta: "Se a idéia jamais tivesse ocorrido a mim, teria ocorrido a outra pessoa pouco tempo depois. *Causalidade* nada mais é do que Providência. No máximo, o mérito vem a seguir, por causa da perseverança.

3. MÉTODOS PIONEIROS E TERMINOLOGIA

2. Ele tinha que criar tanto os métodos quanto as terminologias. Não podia procurá-los em bibliografias, ou em suas leituras, já que se tratava de idéias totalmente novas. No entanto, examinou muitas centenas de *concordâncias* em várias línguas em bibliotecas de Roma, Milão, Munique, Paris, Londres e Nova Iorque. Notando o que era preciso para produzi-las em latim – e logo também em grego e hebraico – extraiu delas nomenclaturas e métodos precisos.

Permitiu-se ser guiado pela verdade das coisas, acolhendo o conselho do Aquinate: "studium philosophiae non est ad hoc quod sciatur quid homines senserint, sed qualiter se habeat veritas rerum"[1].

4. "HIPERTEXTOS", AINDA ANTES DE EXISTIR O TERMO

3. "Quasi ab ipsa veritate coactus",[2] desde o começo, e antes de a terminologia corrente existir (*hipertexto,* SGML, HTML, TEI, XML *etc*. , Pe. Busa

[1] *In De caelo,* bk. 1 1.22 n. 8. Para Santo Tomás, a filosofia é a investigação racional da síntese universal de nossa situação de vida.

uniu três centenas de códigos distintos a cada uma das onze milhões de palavras – incluindo *et* e *non* – contidas no *corpus* de Tomás de Aquino; esses códigos foram codificados em cento e trinta *bytes*, que especificavam diversos valores na área da morfologia.

Agora, no final de sua vida, o projeto que pretende pôr em movimento – o Léxico Tomístico Bicultural, ou LTB – somar-se-á aos códigos já feitos, introduzindo outros que definirão a sintaxe de cada palavra.

5. "SISTEMATIZANDO" PRIMEIRO AS ALOGRAFIAS, DEPOIS A LEMATIZAÇÃO, AS HOMOGRAFIAS E POR FIM AS TIPOLOGIAS.

4. Desde o começo, a enorme quantidade de arquivos forçou o Pe. Busa a *sistematizar* – uma palavra muito freqüente em seus escritos – três situações textuais: "alografias" – i.e., variações na grafia de uma única palavra – lematização e tipologias de discurso.

No que concerne às alografias, ele distinguiu e revisou a diferença entre aquelas variantes que eram apenas gráficas, e aquelas que eram *formais* ou estilísticas.

5. Em relação à lematização, Pe. Busa foi um dos primeiros a recolocar em circulação o termo "lemma".[3] Este termo está agora incluído em dicionários com o significado de primeira palavra de um verbete, que serve como uma chamada, representando as várias sub-formas e definições moduladas. Pe. Busa sistematizou os procedimentos para lematização, distinguindo claramente entre aquela que era apenas morfológica e aquela que era sintática.

A lematização morfológica, que se aplica às várias formas de uma palavra conforme seu contexto – no *corpus* tomístico há 150.000 formas diferentes de palavras – está organizada de modo *tripartite* (palavras invariáveis, palavras declináveis, e verbos conjugáveis), e mostrou-se a mais prática para a computadorização de textos extensos.

A lematização sintática mais tarde foi aplicada aos onze milhões de sentenças em contexto, uma por uma, classificando cada palavra de acordo com oitenta aspectos do discurso.

---

[2] *Contra Gentiles*, bk. 1 ch. 43 n. 16, e em outras dez passagens.

[3] A palavra grega *lemma* só foi recebida pelo Latim no período pós-clássico e continua presente hoje em termos como *dilema* (nota original). Segundo SOUSA, F. *Novo Dicionário latino-português*. Porto: Lello e Irmão, 1992, p. 539, *lemma, atis* poderia ser considerada a premissa menor do silogismo. Para o *Novo Michaelis: Dicionário Ilustrado*. Volume I: Inglês-Português. 22 ed. São Paulo: Melhoramentos, 1977, p. 575, a palavra inglesa *lemma* corresponderia à "proposição que prepara a demonstração de outra". Por falta de correspondência exata, recorreu-se ao neologismo "lematização" (nota dos tradutores).

Pe. Busa sempre foi cético quanto à lematização automática, mas não o é com relação a um processo semi-automático, desde que a primeira parte importante tenha sido feita manualmente. Não obstante isso, ele reconhece que a primeira tem o valor metodológico de lidar sincronicamente com a formalização das estruturas presentes na superfície de nossa expressão. De fato, é tremendamente importante para ele distinguir, por exemplo, os dois sistemas de forças interiores, i.e. entendimento e expressão.

6. O processo de lematização forçou-o a encarar de imediato o fenômeno linguístico da *homografia*, que não foi sistematizada antes do advento do computador. Na verdade, nós todos falamos e lemos por frases, o que quase sempre impede a homografia de ser percebida. Pe. Busa começou a estudá-la pelas suas causas, tipos e causalidade: ele descobriu (mesmo sem mencionar aquelas que ocorrem entre partes do discurso, ou entre palavras de línguas diversas) que ao menos metade das onze milhões de palavras nos textos latinos do Aquinate se revelavam homografias por uma razão ou outra. Assim, ele tinha que individualizar todas as *formas* homográficas dentro de limites bem definidos, para descobrir como avaliar sua probabilidade de ocorrência no *corpus* tomístico, como criar um repertório de todas elas – até onde elas eram *possíveis*, ao menos – de modo a ser capaz de distinguir as mais importantes, deixando as outras para o futuro. De fato, ele fez essa diferenciação em 600.00 contextos. A necessidade de tal sistematização é óbvia para a validade de qualquer elaboração computadorizada de textos, dado que o computador só pode trabalhar na forma física de signos significantes.

7. As tipologias de discurso revelaram-se numerosas nos gêneros literários de amostras de trabalhos analisados em 20 línguas diferentes: resumos científicos, peças de teatro, cartas, literatura, edições manuscritas...

No *Index Thomisticus*, Pe. Busa marcou cada termo com pelo menos dois dos seguintes códigos contextuais: 1. o discurso do próprio autor; 2. uma citação literária; 3. uma citação *ad sensum*; 4. um apanhado resumido de palavras iniciais (*incipit*); 5. uma referência a outro texto; 6. uma referência ao próprio texto corrente; 7. numa digressão; 8. seu peso na fluência do discurso, tanto central quanto periférico.

8. Essas longas e cuidadosas preparações, e especialmente a lematização, tiveram sua recompensa ao permitir e acelerar outras pesquisas mais avançadas.

6. O PRIMEIRO "SISTEMA LEXICOLÓGICO" DE UM AUTOR

O *sistema lexicológico* do *corpus* tomístico é o primeiro e, ainda hoje, o único existente, se consideramos tal sistema como o resultado final da análise e

síntese posterior de um universo linguístico fechado, de acordo com todos os seus elementos de morfologia, sintaxe e léxico. Esse é um novo tipo de documento linguístico: um quantitativo integral e classificação estatística dos principais, mais importantes e fundamentais elementos expressivos de um sistema linguístico.

As 294 páginas do *Tratado de Lexicologia* do Pe. Busa[4] fornecem uma suma de um sistema geral e de três subsistemas (homografia, tipologia e quantidade) das quarenta tabelas dos nono e décimo volumes do *Index Thomisticus*[5], que resume em 2.470 páginas as informações detalhadas encontradas nas 8.022 páginas dos oito volumes anteriores[6].

Entendido desse modo, o *sistema lexicológico*, graças ao computador, iniciou uma nova disciplina, senão nominalmente, ao menos *de facto*. Realmente, uma lexicologia entendida dessa forma corresponderia a uma linguística computacional considerada em sentido pleno em seu objetivo final, i. e. fornecer uma síntese lingüística integral, classificada e estatística obtida de um crescente número de textos – quer dizer, de universos linguísticos fechados – como uma base de documentos. É evidente que tal lexicologia contribuiria grandemente para uma metodologia saudável para a pesquisa científica, inclusive para a ciência humana da linguística.

7. A DESCOBERTA DAS QUATRO LEIS (OU QUASE)

10. Na base deste sistema lexicológico e com milhares de horas de trabalho em equipe, Pe. Busa foi capaz de fazer por si mesmo[7] quatro *descobertas*, no sentido etimológico do termo: alguma cognoscível que somente agora foi trazida à luz, uma invenção no sentido de um encontro, passagem daquilo que estava implícito, mas escondido, para aquilo que é explicitamente conhecido.

---

[4] BUSA, R. *Il libro dei metodi*, vol. 6: *Tratado di Lessicologia*, CAEL, Gallarate, 2001, 264 pp.

[5] BUSA, R. *Index Thomisticus. Sancti Thomae Aquinatis operum omnium indices et concordantiae*, vol. 1: *Sectio prima. Indices*, t. 9: *Systemata lexici, I*: *Systema lexicologicum: Tabula* 1: *Systema lemmatum. Tabula* 2: *Systema formarum A-O* (Frommann-Holzboog, Stuttgart, 1980) XVI, 1257 pp.; *IBIDEM*, t. 10: *Systemata lexici*, I: *Systema lexicologicum: Tabulae* 2 *(finis)*-5. II: *Systema homographiae: Tabulae* 6-12. III: *Systema typologicum: Tabulae* 13-26. IV: *Systema quantitatum: Tabulae* 27-38 (Frommann-Holzboog, Stuttgart, 1980) XII, 1210 pp.

[6] *IBIDEM*, vol. 2: *Sectio secunda. Concordantiae operum thomisticorum. Concordantia altera, t.* 1-8 (Frommann-Holzboog, Stuttgart, 1979) XVIII+1286, 1282, 1286, 1293, 1287, 1300, 1270, 1297 pp. De fato, ele já calculou o número total e percentuais das categorias de onze milhões de palavras. Primeiro sobre 150.000 formas de palavras e depois sobre 20.000 *lemmas* sintéticos (cada um correspondendo, em média, a quatro em dicionários latinos usuais).

[7] Durante a sua longa vida, Pe. Busa observou como, desde os primórdios, aquilo que é novo se dissemina vagarosamente, por conta dos atritos e obstáculos trazidos pelo conhecimento estabelecido.

11. A primeira descoberta foi um tipo de lei de economia na relação entre o número de palavras e o das variadas cadeias de caracteres que as contêm. Especificamente, ele dividiu cada palavra – ou seja, cada *lemma*, depois de separá-lo de seus morfemas de declinação – dividindo aquilo que é constante em no máximo três segmentos (não morfemas!): inicial, central e final. Ele aplicou o nome *string*[8] para cada uma das sequências iguais de fileiras que foram encontradas em diferentes palavras, combinada com outras fileiras e ignorando seus significados.

Verificou-se que 1.500 cadeias de caracteres (depois capazes de serem reduzidos) entre uma e doze letras, combinados juntos, foram capazes de produzir todas as onze milhões de palavras (salvo 4.000, que foram identificadas) do *corpus* latino que foi analisado. Isso é o fato documentado, embora possa se supor que poderia ser válido também para outras línguas, ao menos de tipo análogo. Não sabemos se este padrão foi testado e provado existir em outras línguas. De qualquer modo, seria interessante para a compreensão e transmissão eletrônica dos textos.

12. Padre Busa distinguiu o registro da heterogeneidade das palavras do registro de seu *tipo semântico*, entendendo este último como uma relação entre signo e conhecimento (*significante-significado*) dentro de um arco operativo bidirecional, do conhecimento à expressão e vice-versa.

13. O que se segue é um resumo esquemático desses tipos semânticos, omitindo os códigos decimais:

> 1% é de explícitas palavras diretas (ou demonstrativas?) (distintas daquelas que são sempre implícitas nas declinações das primeira e segunda pessoas do singular e do plural dos termos latinos) que são uma parte dos pronomes pessoais e dos pronomes demonstrativos e advérbios. Elas não expressam imagens mentais, mas *conhecimento* sobre uma presença (seja o que for).
>
> 2% são de nomes próprios. Estes são palavras-etiquetas que significam um indivíduo singular por vez, embora por vezes também possam significar coletivos.
>
> 6% são de substantivos comuns, que denominam tipos específicos de *objetos* e *coisas*. Por exemplo, planta, cavalo, carro, sanduíche *etc.*
>
> 46% são daqueles adjetivos e verbos que especificam os *aspectos* das coisas ou objetos: atividade, passividade, qualidade, dimensões, figuras, odores, sabores *etc*.
>
> 35% são partículas, preposições, conjunções *etc*. que significam direção, relação, correlação *etc*.

[8] "Fileira" em vernáculo (nota dos tradutores).

8% são palavras vicárias que apontam para outras palavras, conceitos ou coisas. São pronomes ou pronominais (em latim não há artigos).

1% é de palavras que significam pessoas ou inteligências *além da física*, que podemos denominar *invisíveis*.

14. Muitos se recordarão das analogias ou correspondências entre essa caracterização e aquela das categorias supremas da realidade de Aristóteles e de Kant.

15. Baseando-se nessa classificação, Pe. Busa extraiu duas conseqüências: primeiro a de que em qualquer léxico as palavras são heterogêneas. E isso é verdade ao ponto de atribuir os resultados escassos das estatísticas concernentes às freqüências de palavras em textos naturais ao fato de que as palavras são normalmente contadas como se fossem homogêneas, como números no mesmo cálculo. Para cada um dos sete grupos apontados acima, um deve efetuar um cálculo de estatísticas e somente depois unir os resultados distintos em um resultado estatístico superior.

16. A segunda conseqüência, descoberta ou redescoberta, era a de que em todo léxico podem ser achados dois hemisférios. Um, que expressa a lógica interna do discurso, consiste em poucas, normalmente palavras breves, que são repetidas frequentemente e que estão igualmente presentes em todos os tipos de discurso. Algumas vezes essas são chamadas *termos gramaticais* ou palavras-função. O segundo hemisfério, que especifica a mensagem a ser comunicada, consiste em muitas palavras diferentes, frequentemente longas, que variam de acordo com o conteúdo do discurso (também chamadas palavras-conteúdo), e cujas freqüências são sempre inferiores àquelas do primeiro hemisfério.

No caso de Santo Tomás, as palavras dêiticas, palavras relacionais e palavras vicárias chegam a 44% do *corpus* total. Nomes próprios, termos aspectos e objetos invisíveis perfazem os restantes 56%. Além disso, vários adjetivos e verbos universais de alta freqüência devem ser atribuídos ao primeiro hemisfério. De fato, ordenando as 150.000 formas de palavras distintas do *Index Thomisticus* por sua freqüência, descobrimos que as oitenta palavras mais frequentes perfazem 41% do corpus e que as oitocentas mais frequentes perfazem 68%.

17. Padre Busa acredita que um progresso substancial deve ser esperado no domínio da linguística computacional através do emprego de todas as informações mencionadas até aqui.

8. O DESAFIO LINGUÍSTICO DA TRADUÇÃO AUTOMÁTICA.

18. Entre os anos de 1950 a 1965, Pe. Busa exerceu um ativo papel nos esforços para desenvolver tecnologias para a tradução automática, pesquisa que foi sustentada por financiamento do Pentágono. Este auxílio econômico repentinamente parou em 1965, porque as ciências linguísticas não estavam fornecendo dados precisos para um programa de computador que traduziria textos para uma outra língua. Quarenta anos depois, o mesmo desafio apareceu, com outros nomes e outros fatores de motivação, devido às redes de comunicação da globalização.

9. A PERSPECTIVA DA *"DISCIPLINE LANGUAGES"*: UMA PROPOSTA PARA A UNIÃO EUROPÉIA

19. Durante um congresso oficial de linguística da União Européia ocorrido em Estrasburgo em 2002, Pe. Busa formulou uma proposta estratégica, que ele chamou de *disciplines languages*; esse conceito foi o fruto de sua pesquisa anterior nas profundezas da expressão latina e de tudo o que ele viu e viveu durante sessenta anos trabalhando com linguística computacional.

20. Muitas décadas atrás, Pe. Busa enfatizou a natureza fragmentada do trabalho que ele produziu por meio de muitas expressões vivazes – no estilo de comandos heroicamente audaciosos durante uma batalha – no que diz respeito ao seu enfoque em textos literários, três dos quais estão incluídos aqui:

- "Uma milha de algoritmos construída no topo de uma polegada de texto",
- "Somente o segundo andar, sem o primeiro",
- "Dez pessoas construindo a primeira milha de uma auto-estrada, através da mesma floresta e na mesma direção, sem ninguém construindo a segunda milha, a terceira, a quarta *etc*.".

21. Em Estrasburgo, Pe. Busa se perguntou e questionou os outros assistentes se o que se segue (esquematicamente resumido aqui) seria audácia ou um pensamento utópico:

- Uma iniciativa comunitária, globalizada e sincronizada, em três fases:

Primeira fase:

- Em todas as principais línguas,
- Baseados nos livros-textos universitários em cada uma das principais disciplinas acadêmicas, transcritos em formato eletrônico,
- O *sistema lexicológico* deveria ser extraído – no sentido descrito pelo Pe. Busa – para cada uma das disciplinas acadêmicas selecionadas,

- Com vistas a combiná-las depois em um sistema para cada linguagem que poderia especificar e quantificar as convergências e divergências em léxico, morfologia e sintaxe.

Segundo estágio:

- Ao mesmo tempo, os sistemas para cada linguagem individual deveriam ser fundidos em um único sistema lexicológico interlinguístico que deveria conter, em formato computadorizado, o mapa geográfico das correlações de convergência e divergência. Esse deveria ser um repositório detalhado de "discipline-specific languages", com percentagens e ligações entre as correspondências e entre as divergências no léxico, morfologia e sintaxe de cada linguagem com vistas aos outros.

Finalmente,

- Em cada linguagem, um manual da "*discipline language*" deveria ser definido e publicado, com léxico, morfologia e sintaxe, para que seja empregado como o material nas mensagens de rede.
- Na produção final, a mensagem recebedora deveria ser capaz de requisitar do servidor central uma tradução para uma língua-alvo.