Por que ainda não conseguimos conversar com o computador?

Osvaldo N. Oliveira Jr.*, Maria das Graças V. Nunes e Maria Cristina F. Oliveira****

*Instituto de Física de São Carlos, USP

CP 369 - 13560-970 São Carlos, SP - Brasil

FAX: +55 16 2713616; e-mail: chu@ifqsc.sc.usp.br

**Instituto de Ciências Matemáticas de São Carlos, USP

CP 668 - 13560-970 São Carlos, SP - Brasil

FAX: +55 16 2749150; e-mail: [mdgvnune,ferreira]@icmsc.sc.usp.br

Resumo

São discutidas algumas das muitas dificuldades do **p**rocessamento de **l**inguagem **n**atural que poderá um dia levar a uma comunicação com o computador (ou robôs) na linguagem natural dos humanos. Tais dificuldades foram subestimadas pelos autores de ficção científica das décadas de 60 e 70, de maneira que a comunicação possível com a máquina está hoje muito menos evoluída do que o previsto. Apesar das limitações, os estudos do PLN são extremamente úteis para várias aplicações do computador na vida cotidiana. Apresentamos, por exemplo, resultados do desenvolvimento de um Revisor Gramatical para o português do Brasil, que também servem para ilustrar os desafios do PLN para as próximas décadas.

I. Introdução

Talvez os maiores enganos da ficção científica das décadas de 60 e 70, com relação à robótica e à informática em geral, consistiram nas previsões quanto à comunicação da máquina com o homem. Foram previsões hoje consideradas extremamente otimistas acerca da evolução de áreas como visão artificial e **p**rocessamento de **l**inguagem **n**atural. A título de ilustração, vale lembrar que no filme "2001: Uma odisséia no espaço", que fez sucesso nos anos 70, uma personagem essencial era um computador capaz de ver com grande precisão e riqueza de detalhes, interpretar linguagem falada e escrita, expressar-se por estes mesmos veículos, e o que hoje parece mais incrível, ter a capacidade da leitura labial. Sabemos hoje que o reconhecimento de padrões é ainda um grande desafio para pesquisadores de visão artificial, o mesmo ocorrendo com a compreensão da linguagem humana por uma máquina. Queremos tratar, aqui, somente do problema do processamento de linguagem natural, conhecido como PLN nos meios computacionais. PLN é uma área da inteligência artificial voltada aos estudos e desenvolvimento de sistemas que permitam interpretar e gerar linguagem natural, ou seja, a linguagem utilizada pelos humanos (em oposição às linguagens computacionais, ditas artificiais), quer seja o português, espanhol, inglês ou outra língua qualquer. Ao abordar as dificuldades do PLN, tentaremos responder à pergunta do título e mostrar que a comunicação com a máquina é muito mais complexa do que se imaginava.

Embora a compreensão e geração automática de linguagem natural ainda estejam longe de se tornar realidade, e não possamos nem mesmo prever quando isso vá ocorrer, já existem várias aplicações originárias de pesquisas em PLN que, além de sua importância intrínseca enquanto aplicações, servem como passos iniciais para a comunicação com o computador. Dentre as aplicações mais sofisticadas, do ponto de vista da compreensão e geração automática, destacam-se os sistemas para domínios específicos, ou seja, sistemas que geram texto em linguagem natural sobre um determinado assunto em uma certa área do conhecimento. Pertencem a essa classe os sistemas especialistas que fornecem diagnósticos médicos (na forma de texto) automaticamente, com base em informações obtidas sobre o paciente. Ressalte-se que estes textos não são simplesmente extraídos de mensagens pré-armazenadas (*canned text*), como é usual nas comunicações do computador com os seus usuários. Por exemplo, suponha que um usuário deseje localizar dois arquivos em um dado diretório. Se os arquivos não forem encontrados, receberá uma mensagem do tipo "Os arquivos não foram encontrados", que já estaria armazenada na memória do computador para ser exibida nestas situações. Entretanto, é possível que os arquivos estivessem em outros diretórios. Se o computador pudesse realmente compreender o que estava sendo solicitado, poderia procurar pelos arquivos, e a resposta poderia ser algo como "Os arquivos estão no diretório X" ou "O arquivo A está no diretório X e o arquivo B está no diretório Y". Perceba que, para fornecer tais mensagens, o computador, além de ter a informação sobre a localização dos arquivos nos diretórios, deveria ser capaz de "montar" as sentenças a serem exibidas, inclusive com conhecimento da gramática da língua para respeitar as regras de concordância nominal e verbal, por exemplo.

Aqui é bom fazer uma ressalva sobre notícias freqüentemente veiculadas nos meios de comunicação tratando de sistemas computacionais inteligentes capazes de reconhecer a voz humana e obedecer a comandos falados. Esses sistemas são, geralmente, restritos a um conjunto bastante pequeno de comandos e têm seu valor muito mais pelo processamento do som, no reconhecimento dos fonemas de um certo usuário falando numa determinada língua, do que pelo processamento da linguagem natural. Esse mito de que a solução para a comunicação em linguagem natural está próxima também se estende à tradução automática, área investigada há várias décadas, que consiste na tradução de textos de uma língua fonte para outra. Ressalte-se, novamente, que os tradutores automáticos existentes também atuam em domínios limitados. Por domínio limitado entenda-se uma área do conhecimento particular com um número finito (e pequeno) de palavras, geralmente utilizadas em contextos únicos, sem dar margem a ambigüidades, que são comuns no uso diário da língua. Tais ambigüidades são facilmente resolvidas em uma conversação entre humanos, em geral através de uma análise (talvez inconsciente) do contexto, mas constituem um enorme problema para o tratamento automático da língua. Na expressão "o homem viu a mulher na montanha de binóculos", por exemplo, não há dúvida para um humano sobre quem estava usando os binóculos, mas certamente a máquina precisaria ter muito conhecimento para inferir que era o homem que usava os binóculos.

II. Aplicações do Processamento de Linguagem Natural

O objetivo final do tratamento automático de linguagem natural espelha o desejo de que usuários possam se comunicar com o computador através da linguagem do dia-a-dia, e não mais através de linguagens limitadas que requerem treinamento específico - as chamadas linguagens de programação. Se isso fosse possível, qualquer pessoa poderia delegar tarefas aos computadores ou trabalhar em parceria com eles, sem necessidade de treinamento prévio, assim como supunham os autores de ficção científica de algumas décadas atrás. Há, entretanto, muitas outras possíveis aplicações do computador na vida cotidiana de usuários leigos que podem se beneficiar do PLN. Uma das mais importantes é talvez a crescente utilização do computador na educação. Esse tema tem sido abordado notadamente nos países mais desenvolvidos, principalmente devido à preocupação em encontrar formas eficazes de ajudar no aprendizado. Críticas veementes têm sido feitas às abordagens adotadas, principalmente porque consistem em apenas transferir material didático de livros, filmes, etc. para um meio eletrônico. De fato, se quisermos tornar o computador uma ferramenta eficiente de apoio ao ensino, há que se ir além de produzir interfaces agradáveis e bonitas, e mergulhar no universo dos processos de aquisição de conhecimento dos usuários aprendizes. Auxiliar alunos no processo de alfabetização, no aprendizado de uma língua estrangeira, ou no próprio processo de escrita de textos vários, é uma tarefa bastante complexa.

Um contexto de utilização mais específico dos computadores associado ao processamento de textos em linguagem natural diz respeito às ferramentas de auxílio à escrita. Ferramentas de software vêm sendo criadas para auxiliar na escrita de diversos tipos de textos, desde relatórios a artigos científicos. Em geral, elas podem ser classificadas em três categorias: ferramentas de pré-processamento do texto, através das quais o computador oferece suporte para um escritor organizar suas idéias relativas ao que vai escrever (atividade algumas vezes denominada em inglês de "brain storming"); ferramentas de pós-processamento de texto, que oferecem ajuda na correção e/ou revisão de um texto pré-preparado; e os chamados ambientes de auxílio à escrita (conhecidos como "Computer-Aided Writing Environments"). Estes últimos são mais sofisticados, e geralmente englobam ferramentas de pré- e pós-processamento, além de ferramentas adicionais capazes de auxiliar o escritor a transformar suas idéias iniciais em texto.

Uma outra aplicação está no auxílio à escrita de um texto em uma língua estrangeira. A necessidade que os cientistas têm de escrever em inglês, por exemplo, gerou um projeto na Universidade de São Paulo, em São Carlos, Brasil [1], para o desenvolvimento de ferramentas que auxiliem escritores, cuja língua-mãe não é o inglês, a escrever artigos científicos nessa língua. Nestas ferramentas, procura-se oferecer trechos de texto, contextualizados, para evitar que o cientista faça uma "tradução" do conteúdo do seu artigo para o inglês sob a forte influência da sua língua materna, que no caso do português apresenta características distintas com relação à estruturação de sentenças, entre outras diferenças. Por exemplo, a expressão em português "Foi observado um aumento na população" poderia ser traduzida para "It was observed an increase in the population", mantendo a estrutura da sentença original. Embora correta gramaticalmente, esta estrutura dificilmente seria empregada por um falante da língua inglesa, que provavelmente preferiria "An increase was observed in the population".

Dentre as ferramentas de software de auxílio à escrita, certamente as mais desenvolvidas são as de pós-processamento, ou seja, os revisores ortográficos e gramaticais, muitos dos quais já estão no mercado há vários anos, principalmente para textos escritos em inglês. Estes revisores detectam, em primeira instância, erros ortográficos oriundos principalmente de falhas na digitação. Esse tipo de revisão é conceitualmente bastante simples, pois basta comparar as palavras do texto com as de um banco de palavras mantido na memória do computador, verificando sua existência. A única dificuldade estaria, então, na necessidade de se armazenar um grande número de palavras e realizar consultas ao banco de dados de maneira rápida. Tais problemas foram superados e os revisores ortográficos disponíveis são bastante eficientes, inclusive oferecendo sugestões de palavras alternativas àquela não encontrada, geralmente através de uma análise de similaridade entre a palavra procurada e palavras que estão disponíveis no banco de dados. Além de revisão ortográfica, existem revisores que detectam erros de malformação de sentenças e de concordância, entre outros. Para a detecção desses últimos tipos de erros, há que se embutir conhecimento no sistema computacional, que deve ser capaz não só de reconhecer as palavras do texto, mas também de categorizá-las. Ou seja, dada uma palavra, é necessário verificar se ela é um substantivo, adjetivo ou verbo; se for substantivo se é masculino, singular, etc. O revisor precisa também ter conhecimento das regras gramaticais da língua em questão. Apesar dos enormes avanços já alcançados, há muito por fazer para que a revisão automática se aproxime daquela que pode ser feita por um humano competente no uso da língua. Mais adiante abordaremos em detalhe os problemas e soluções adotadas no desenvolvimento de um revisor para o português do Brasil, com ilustração das suas limitações.

Apesar de ser recente o interesse no desenvolvimento de revisores gramaticais do ponto de vista do mercado de software, o processamento de linguagem natural da língua portuguesa foi iniciado há muitos anos, quase sempre com trabalhos de cunho acadêmico. Por essa razão, os tópicos abordados, em geral, traziam alto grau de complexidade, pois atacavam fenômenos lingüísticos de difícil tratamento, mesmo para humanos. São exemplos os tratamentos de ambigüidades, problemas de ordem semântica e omissões de palavras que podem ser resgatadas pelo contexto. Considere o exemplo: *"O homem voltou da guerra. Estava magro."* Embora não esteja explicitado na segunda oração, é fácil (para um humano) perceber que o sujeito da segunda oração é o mesmo "homem" mencionado na primeira. Entretanto, essa conclusão está longe de ser óbvia para um sistema automático de tratamento de linguagem natural, e ainda hoje coloca sérios problemas para um sistema de correção. Uma dificuldade adicional até alguns anos atrás eram os restritos recursos computacionais, em termos de hardware, capacidade de memória, etc., muito mais limitados do que os atuais. Isso significa que problemas facilmente solúveis no momento poderiam ser intratáveis há 5 ou 10 anos. Algoritmos de busca extremamente eficientes, antes indispensáveis em certas situações, hoje podem não ser tão relevantes para estas mesmas situações. Para se ter uma idéia da evolução da capacidade computacional, ferramentas como o Revisor Gramatical a ser apresentado a seguir hoje rodam em microcomputadores tipo PC, ao passo que os primeiros revisores da década de 1980 requeriam computadores "mainframes" para serem executados, o que restringia severamente a sua utilização em contextos reais.

Numa comparação com o inglês, em princípio deve-se esperar dificuldades maiores no processamento do português. À semelhança da língua espanhola e outras línguas latinas, a língua portuguesa utiliza mais tempos verbais e flexões dos substantivos e adjetivos do que a língua inglesa. Outro fator potencialmente complicador está nas múltiplas possibilidades de estruturação de uma dada sentença. Entre outros exemplos, pode-se citar a alternância de posições entre substantivo e adjetivo, que pode até acarretar mudança de sentido. Esse é o caso da expressão "homem grande" que tem significado bastante diverso de "grande homem".

Uma outra dificuldade, agora comum a muitas línguas, é a existência da polissemia (uma mesma palavra pode ter vários significados diferentes), e da homografia (palavras diferentes podem ser escritas da mesma maneira). O fenômeno conhecido como polissemia pode ser ilustrado pela palavra "cadeiras", que tanto pode ser o plural de cadeira (silla em espanhol) como uma parte do corpo humano (parte inferior das costas). Quanto à homografia, pode-se exemplificar através das palavras "governo", substantivo masculino e "governo", a primeira pessoa do indicativo do verbo governar. Esses dois fenômenos geram um grande número de ambigüidades, muitas vezes só esclarecidas com o conhecimento do contexto. Uma análise do contexto, entretanto, é quase sempre impraticável para os sistemas de tratamento de linguagem natural, pelo menos em seu estágio atual de desenvolvimento. Isso porque tal análise exigiria um conhecimento do "mundo" que não é facilmente embutido em um sistema computacional a não ser, como já mencionado, em contextos e domínios limitados.

As várias dificuldades mencionadas apontam para a necessidade do desenvolvimento de um grande número de estratégias lingüísticas e computacionais para que o processamento de linguagem natural seja eficaz. Em resumo, o computador precisaria ser capaz de: i) reconhecer as palavras de uma língua, ii) saber seu significado, iii) saber as regras da gramática daquela língua e iv) ter informações adicionais que pudessem ser utilizadas para eliminar ambigüidades causadas principalmente pela polissemia e homografia. Projetos criados com essa finalidade precisam, portanto, contar com uma equipe multidisciplinar de pesquisadores, notadamente lingüistas e cientistas da computação, e em alguns casos estar preparados para fazer estudos exaustivos de tópicos da língua em questão. Vem daí a importância de contar com parceiros comerciais, face aos grandes recursos humanos e de tempo requeridos.

III. Revisor Gramatical - Uma aplicação de PLN para a língua portuguesa do Brasil

Numa parceria entre a indústria de informática Itautec-Philco S/A e a Universidade de São Paulo (USP), de São Carlos, Brasil, está sendo desenvolvido um Revisor Gramatical para o português do Brasil [2], cuja primeira versão já está disponível comercialmente, acoplado ao Editor de Textos **REDATOR**. O Revisor é fruto de um projeto de pesquisa de mais de 3 anos, envolvendo especialistas em computação e lingüística. A seleção dos erros prioritariamente tratados baseou-se em um levantamento dos erros mais freqüentes cometidos por pessoas de nível médio de escolaridade, isto é, em nível de segundo grau, que formam o público-alvo para o Revisor. O desenvolvimento envolveu um detalhado estudo da língua em uso, com testes em textos reais. Para tanto foi construído um *corpus* (banco de textos) que contém tanto textos já corrigidos e editados que servem como referência do uso corrente da língua escrita, quanto textos escritos por algumas classes de usuários comuns, sem correção. Atualmente, o banco de textos conta com aproximadamente 35 milhões de palavras.

3.1 Características do Revisor Gramatical

O Revisor Gramatical detecta inicialmente erros ortográficos, para o qual utiliza tecnologia desenvolvida e implementada por uma equipe da Universidade Estadual de Campinas, Unicamp (Brasil) [3]. Além disso, detecta também erros mecânicos, que são facilmente identificáveis, embora não sejam normalmente detectados por revisores ortográficos. São exemplos de erros mecânicos a presença de palavras e símbolos de pontuação repetidos e o uso não balanceado de símbolos delimitadores.

O módulo central do Revisor, entretanto, é o Módulo Lingüístico, responsável pela detecção de erros gramaticais. Dois tipos de estratégias foram adotados para a implementação de regras de detecção automática dos erros. Na primeira estratégia, denominada *"error driven"*, procura-se por padrões indicativos de erros freqüentes. Essa estratégia apresenta vários benefícios do ponto de vista da implementação computacional: agilidade, especificidade, rapidez, portabilidade, e disponibilidade de memória. Entretanto, seu escopo de atuação é muito limitado pois não são detectados problemas envolvendo itens lexicais não contíguos e estruturas recursivas. Para superar essa limitação, uma segunda estratégia foi adotada, incluindo análise sintática automática das sentenças, antes de operar a revisão propriamente dita. Tal análise obviamente requer que todos os itens lexicais estejam categorizados apropriadamente, o que demandou um trabalho enorme de categorização para os mais de 1.5 milhões de itens lexicais do banco de palavras.

A implementação das regras de correção foi sempre acompanhada de um estudo sobre a ocorrência de **falsos erros** do sistema. Por **falso erro** subentende-se uma intervenção do sistema que induzirá o usuário a um erro gramatical, através da modificação de uma estrutura lingüística originalmente correta. O desempenho do Revisor, quanto a tempo de execução, pode ser considerado como ótimo, uma vez que as mensagens de erro são apresentadas ao usuário, praticamente instantaneamente. As limitações do Revisor, entretanto, estão localizadas nos falsos erros que ainda comete e nos erros não detectados. A maior parte destes problemas advém da impossibilidade de se prever todas as estruturas sintáticas desviantes que possam ser empregadas por usuários médios. Embora o sistema conte, hoje, com mais de 600 produções, ainda aparecem estruturas para as quais nenhum casamento (matching) é obtido. Além disso, as dificuldades advindas de pluricategorização de alguns itens lexicais, principalmente nos casos de homografia, precisarão ser tratadas em casos especiais, o que certamente demandará grandes esforços de pesquisa.

3.2 Exemplos de erros detectados pelo Revisor

Nos tipos de erros a serem detectados tentou-se cobrir o maior número de casos freqüentes. Nesta Seção apresentamos alguns exemplos dos tipos de erros detectados:

a) Crase[1]

Após os erros ortográficos, são os erros de crase os mais freqüentes. Por exemplo, na sentença *"Chegaram **à** conclusões diferentes"* não deveria haver crase no **a** porque a palavra seguinte está no plural.

b) Pontuação

Os erros de pontuação podem ser graves, pois o sentido da sentença pode ser completamente alterado. Na oração *"Outra conclusão importante, foi que a temperatura aumentou"*, não deveria haver vírgula.

c) Concordância Nominal e Verbal

---

[1] Crase é a contração do artigo **a** com a preposição **a**, como no exemplo: Ela foi à (a+a) escola.

Erros de concordância nominal são muito freqüentes para uma língua como o português, em que substantivos e adjetivos precisam concordar em gênero e número. Para correção de erros de concordância, é necessária a análise sintática. Por exemplo, na oração *"**Os fogos de artifício** da noite de São João são **lindo**"*, somente com análise sintática é possível verificar que "lindo" deveria concordar com "fogos de artifício". Esse é o caso, também, do erro de concordância verbal da oração *"**As guerras** entre os deputados pelas possíveis alianças **começou**"*, em que o sujeito deveria concordar com o verbo da oração.

Deve-se ressaltar que a exigência de análise sintática automática traz uma série de limitações. Se houver algum erro na análise, talvez até oriundo de um descuido ortográfico (por exemplo, a troca da palavra "está" (verbo) por "esta" (pronome)), o sistema de análise automática é induzido a um erro que se reflete na aplicação das regras de correção.

d) Uso inadequado de palavras e expressões

Boa parte dos erros cometidos por usuários comuns advém do uso inadequado de palavras, expressões e também da influência da linguagem falada (oralidade). Num sistema automático, só é possível detectar erros em contextos limitados. Esse é o caso do emprego dos verbos "Fazer" e "Haver" junto a marcadores de tempo, situações nas quais deveriam aparecer no singular. Alguns exemplos de erros são apresentados abaixo, com a correção entre parênteses:

Ele chegou **a** (há) dois anos.

Ela chegará daqui **há** (a) dois anos.

Fazem (Faz) dois meses que não tomo cerveja.

Uma outra dificuldade importante no Português do Brasil está no uso dos prefixos. É muito difícil lembrar de todas as regras de composição de palavras. Para o computador, por outro lado, não há nenhuma dificuldade em armazenar as regras, que são fixas. Assim, o Revisor corrige com segurança erros cometidos na formação de palavras como: *neo* + *clássico* = *neoclássico*; *super* + *rentável* = *super-rentável*; *auto* + *afirmação* = *auto-afirmação*.

3.3 A legibilidade de um texto

A discussão da adequação de um texto ao nível de escolaridade do leitor é bastante antiga. Alguns trabalhos científicos já demonstraram que flutuações da tiragem de um jornal ou revista de grande circulação podem depender do índice de legibilidade do texto. Ou seja, à medida que o texto exige um nível maior de escolaridade, ou mesmo de atenção, do que o público-alvo está acostumado as tiragens caem. Não é fácil verificar se um dado texto é adequado aos leitores a que se destina, pois pressupõe informações sobre o conhecimento prévio desses leitores. Uma alternativa está na utilização de métricas que forneçam um índice de legibilidade, dependente exclusivamente de parâmetros físicos de um texto, como o tamanho das palavras e das sentenças. Desde o trabalho pioneiro de Rudolf Flesch [4] para a língua inglesa, várias fórmulas têm sido criadas para "medir" a legibilidade de um texto, inclusive para outras línguas. Através destes índices, pode-se determinar a escolaridade mínima necessária de um leitor para compreender o texto. Nas fórmulas, pressupõe-se que textos com sentenças mais longas e palavras maiores serão mais difíceis de ler.

Num trabalho inédito para o português, pudemos identificar as faixas de índices para diferentes graus de dificuldade, como mostra a Tabela 1. Estas faixas compreendem escores de Flesch modificados (adaptados para o português) maiores que 75 para textos chamados **fáceis**, que correspondem a 1a. a 4a. série do 1o. grau, de 50 a 75 para textos **médios** (de 5a. a 8a. série) e de 25 a 50 para textos **difíceis** (2o. grau e universitário). Existe ainda uma outra classe de textos, considerados muito difíceis, que englobam textos acadêmicos (ou de áreas muito específicas) com características marcantes. É o caso de textos de áreas médico-biológicos, em que a grande incidência de termos técnicos, como aminoácido, celulósica, com palavras enormes leva a índices de Flesch bastante baixos, ou seja, extremamente difíceis.

Índice de legibilidade (Flesch modificado)

pontuação do texto	**Grau de dificuldade**
75 a 100:	muito fácil
50 a 75:	fácil
25 a 50:	difícil
0 a 25:	muito difícil

Obviamente não se pode afirmar que o emprego de um índice como o Flesch, baseado apenas em parâmetros físicos do texto, garanta que o grau de inteligibilidade previsto seja adequado. Portanto, a identificação de faixas de dificuldade associadas à escolaridade **brasileira** foi baseada apenas nos resultados obtidos com textos reais. Saliente-se que os índices apontam para os níveis de escolarização dos possíveis leitores do texto, e não, necessariamente, para o próprio nível do escritor do texto.

IV. Conclusões

Neste artigo tentamos demonstrar as grandes dificuldades para o processamento automático de linguagem natural, que poderá no futuro permitir a comunicação do homem com a máquina numa língua natural, como inglês, português ou espanhol. Essas dificuldades advêm de ambigüidades lexicais, semânticas e principalmente do contexto. Para resolvê-las, além de informações lingüísticas, uma enorme quantidade de conhecimento do "mundo" precisa ser embutida no sistema computacional. Apesar de essas metas não poderem ser atingidas no futuro próximo, pesquisas em PLN já geraram vários produtos com bom desempenho em domínios restritos, tais como revisores gramaticais, tradutores automáticos e geração de linguagem natural em sistemas especialistas. Como ilustração de uma aplicação do PLN, foi descrito o desenvolvimento de um revisor gramatical para o Português, numa parceria entre a Universidade de São Paulo, em São Carlos, e a Itautec-Philco S/A, do Brasil.

Agradecimentos

Os autores são pesquisadores apoiados com bolsas de produtividade do CNPq. O projeto do Revisor Gramatical foi patrocinado pela Itautec/Philco S.A. Os autores agradecem à equipe de desenvolvimento do Revisor.

Referências

[1] ALUISIO, S.M.; OLIVEIRA JR., O.N.; A case-based approach for developing writing tools aimed at non-native English users, Lecture Notes in Artificial Intelligence, 1010, 121 (1995).

[2] NUNES, M.G.V.; GHIRALDELO, C.M.; MONTILHA, G.; TURINE, M.A.S.; DE OLIVEIRA, M.C.F.; HASEGAWA, R.; MARTINS, R.T.; OLIVEIRA JR., O.N.; Desenvolvimento de um sistema de revisão gramatical automática para o português do Brasil. In: Simpósio Brasileiro de Inteligência Artificial, 13. Encontro para o Processamento Computacional do Português Escrito e Falado, 2., Curitiba, Outubro, 1996.

[3] KOWALTOWSKI, T.; LUCCHESI, C.L.; Appications of finite automata representing large vocabularies, Software Practice and Experience, 23, 15, (1993).

[4] FLESCH, R.F.; A new readability yardstick, J. Appl. Psychology, 32, 221 (1948).