Instituto de Ciências Matemáticas e de Computação

ISSN - 0103-2585

O Processamento de Línguas Naturais: para quê e para quem?

Maria das Graças Volpe Nunes

Nº 73

NOTAS DIDÁTICAS DO ICMC

São Carlos

MAIO/2008

ÍNDICE

O Processamento de Línguas Naturais: para quê e para quem?*

Maria das Graças Volpe Nunes

NILC-ICMC-USP

1. Introdução

Processamento de Línguas Naturais (PLN) é o nome que se dá à área de pesquisa que se dedica a investigar, propor e desenvolver formalismos, modelos, técnicas, métodos e sistemas computacionais que têm a língua natural como objeto primário. Essa caracterização abrangente dificulta muitas vezes distinguir essa área de outras correlatas, como a Lingüística Computacional (LC), a Lingüística Aplicada, a Lingüística de Corpus (na grande área de Lingüística), e mesmo com outras da Inteligência Artificial, como Text Mining (Mineração de Textos). No exterior, grandes conferências denominadas de Lingüística Computacional abrangem de fato os estudos de PLN. Por outro lado, não é raro haver distinção entre LC e PLN com base apenas nas diferentes perspectivas sobre um mesmo objeto: se lingüística ou computacional, respectivamente.

A fim de deixar mais claro o que entendemos, de fato, por PLN, tentamos esclarecer o que se faz nessa área. De modo geral, em PLN buscam-se soluções para problemas computacionais, ou seja, tarefas, sistemas, aplicações ou programas, que requerem o tratamento computacional de uma língua natural (português, inglês, etc.), quer seja escrita (texto) ou falada (fala). Línguas naturais alternativas, como a linguagem de sinais para os deficientes auditivos, têm igualmente sido alvo crescente de estudos para alguma forma de automatização. No caso da fala, por questões tecnológicas, os trabalhos nessa linha costumam ser tratados em fóruns próprios ou mesmo ligados à Engenharia Elétrica, mais precisamente, à área de Processamento de Sinais. Os principais problemas nessa área são relacionados a questões da produção (síntese) e recepção (reconhecimento) do som, e menos a questões lingüísticas, fazendo com que estas sejam tratadas de forma independente. Por esse motivo, PLN é quase sinônimo de processamento de língua escrita. Evidentemente, o fato de tratar língua escrita não confere à área qualquer facilidade. Escrita e fala compartilham diversos problemas, e, ao contrário da fala, a escrita não pode contar com recursos importantes para a solução de problemas, como a entonação, o volume da voz, entre outros.

* Texto-base para palestra proferida na 1ª. Escola Brasileira de Lingüística Computacional (EBLC) FFCLCH-USP, São Paulo, 3 a 6 de setembro de 2007

A primeira grande – e eterna – tarefa de PLN foi a Tradução Automática (TA). Ao contrário do que ocorre em muitas áreas, em PLN não se iniciou pelo começo, mas sim pelo fim. A TA (aqui me refiro à completa, ideal, similar à humana) requer que um texto seja tratado em todos os níveis, de forma completa e sofisticada. Basta dizer que idealmente o texto na língua fonte deve ser completamente compreendido e a sua tradução corretamente gerada na língua alvo. Evidentemente isso requer um sofisticado tratamento (a) do texto fonte, no caso da interpretação: desde o reconhecimento das unidades lexicais até o mapeamento sintático e semântico, passando pela categorização morfossintática, a estruturação sintática e o tratamento semântico; (b) do texto alvo, no caso da geração: a escolha lexical não ambígua, a linearização superficial, a adequação ao estilo, entre outros critérios; e (c) eventualmente da transferência entre uma representação e outra.

Com esse desafio inicial, não é de se surpreender com o fato da TA ter sido relegada a um segundo plano por várias décadas, até ser retomada recentemente, quando os avanços tecnológicos e técnicos tornam mais factíveis os resultados esperados. Soma-se a isso o fato de, após décadas de convívio mais intenso com o computador, seu usuário ser mais realista quanto ao que pode esperar.

Logo os envolvidos se deram conta da complexidade da tarefa, mais especificamente, da necessidade de se tratar cada etapa separadamente, investigando a fundo os fenômenos característicos de cada uma. Dessa forma, tem sido possível determinar o papel (entenda-se a complexidade, os recursos, a contribuição) de cada etapa de processamento lingüístico, para diferentes tarefas ou aplicações.

Atualmente é mais comum o estudo/proposta para o tratamento de problemas bem pontuais (p.ex. como reconhecer e/ou classificar um sintagma ou uma sentença, como rotular uma ocorrência textual com sua categoria gramatical no contexto, como relacionar retoricamente duas porções textuais, como escolher o sentido correto de uma palavra ambígua, como transformar uma estrutura sintática hierárquica numa sentença linear, etc.). A idéia é fragmentar o problema e solucioná-lo por meio da combinação das soluções das partes. Como, via de regra, as soluções não são completas (a ambigüidade da língua natural é apenas uma das razões para isso), a avaliação da solução completa é mais um problema da área: quais das soluções intermediárias mais contribuíram para o (a falta de) desempenho geral? Embora seja natural recorrer ao falante humano para a tarefa de avaliar sistemas de PLN, esse procedimento recebe muitas críticas pela subjetividade envolvida e, principalmente, por impedir uma replicação de resultados – fatores essenciais para tornar o processo científico. Nem sempre é possível, no entanto, abrir mão de avaliações humanas. Nesses casos, para mantê-las cientificamente coerentes, medidas estatísticas são usadas para avaliar a credibilidade dos julgamentos. Por essas razões, a avaliação de sistemas de PLN tem sido um tópico de pesquisa por si só.

É nesse cenário que o PLN se dedica a **tarefas básicas** tais como:

- No pré-processamento de textos: subdividir o texto em unidades fonéticas, lexicais, gramaticais, semânticas ou discursivas, de acordo com o objetivo da tarefa em questão;

- Classificar (Etiquetar) automaticamente as unidades do texto, segundo classes pertinentes à tarefa: morfossintáticas (PoS-tagger), sintáticas (Parser), semânticas (Parser Semântico ou Interpretador), discursivas (Parser discursivo). Em cada caso, é necessário definir linguagens de anotação, usadas para representar as classes: etiquetas, relações, estruturas (p.ex. árvore sintática).

- Mapear representações: da LN para uma representação sintática, semântica ou discursiva; e dessas para LN – Interpretação e Geração de LN.

Qualquer aplicação de PLN requer uma ou mais tarefas básicas. Por aplicação entende-se qualquer sistema computacional que, a partir de uma entrada lingüística, realiza um processamento que a transforma de tal maneira a produzir um determinado resultado. Devido a esse caráter genérico, uma aplicação pode ser parte de outra aplicação maior, e nesse caso, ela assume o papel de um dos módulos da aplicação maior. Como **exemplos de aplicação** que envolve ou pode envolver PLN, citamos:

- Sistemas de TA: não só os completos (dado texto, retorna a tradução), mas variações, como os sistemas que consideram alguma forma de edição humana, seja ela feita previamente, durante a tradução, ou posteriormente - *Human-Aided Machine Translation* ou, simplesmente, HAMT. Quando servem de auxílio à tradução humana, são chamados *Machine-Aided Human Translation* ou, simplesmente, MAHT. Esses últimos incluem ferramentas de acesso a dicionários e enciclopédias, recursos de processamento de textos, verificação ortográfica e gramatical, entre outras.

- Sistemas de Sumarização Automática (SA): geram extratos (justaposição de porções do texto fonte) ou resumos (texto gerado a partir de um plano de resumo) de um ou mais textos, de acordo com uma taxa de compressão ou outros critérios, como o objetivo do usuário, o domínio da aplicação, etc.

- Sistemas de Categorização de Textos: classificam textos de acordo com alguma taxonomia (domínio, gênero, estilo, retórica, etc.). Uma variação dessa aplicação são os Sistemas de Identificação de Autoria ou de Idioma. As técnicas empregadas vão desde as mais simples, guiadas pela freqüência de palavras-chave, até as estatísticas, em que probabilidades são calculadas a partir de um corpus de referência, passando por métodos baseados em conhecimento lingüístico, superficial (freqüência de padrões) ou profundo (baseado em teorias lingüísticas ou literárias).

- Sistemas de Recuperação de Informação (RI): esses sistemas se destinam a recuperar documentos relevantes a uma dada consulta, entre uma coleção de documentos. Com a Web, essa aplicação ganhou muito destaque, tendo o

sistema de consulta Google como seu mais conhecido representante. Se antes a consulta deveria obedecer a linguagens rígidas de busca (como o SQL), hoje em dia o usuário é livre para formalizar sua consulta em LN, podendo até parametrizá-la de modo a aperfeiçoar a busca. Quando a RI trata a consulta como um padrão qualquer, dizemos que não se trata de PLN. Mas quando a consulta é vista e processada como um fenômeno lingüístico, requerendo qualquer tipo de processamento básico, então dizemos que se trata de uma RI linguisticamente motivada, e, portanto, de uma aplicação de PLN. Sistemas de RI *cross-language,* em que consulta e documentos são em línguas distintas, têm merecido a atenção dos pesquisadores.

- Sistemas de Extração de Informação (EI): ao contrário da RI, a EI busca uma resposta a uma pergunta de entrada em um ou mais documentos. Essa resposta pode ser eventualmente identificada por meio de um simples casamento de padrão da pergunta, mas, para ser precisa e robusta, requer um processamento mais complexo tanto da pergunta quanto dos documentos inspecionados, geralmente envolvendo processos de PLN, como etiquetação, consulta a dicionários e ontologias, aplicação de heurísticas, etc.

- Sistemas de Diálogos: englobam os sistemas de interpretação de diálogos (conseguiriam, por exemplo, responder perguntas sobre o que foi veiculado pelos participantes do diálogo) e os sistemas que participam de um diálogo, geralmente travado com o usuário (p.ex. um sistema de reserva/compra de passagens, de informações turísticas, etc.).

- Sistemas de Auxílio à Escrita: desde os editores de texto (revisor ortográfico, gramatical, estilístico) até sofisticados ambientes de auxílio à produção de texto, onde o usuário pode encontrar recursos para construir textos bem estruturados, de um gênero e/ou domínio específicos (p.ex. Scipo, Scipo-Farmácia[2]).

Para tanto, são necessários **recursos lingüístico-computacionais**, cujo planejamento e construção constituem tarefas por si só. Exemplos desses recursos são corpora, léxicos ou dicionários, ontologias e gramáticas.

- **Corpora** de textos são úteis para o levantamento de conhecimento lingüístico (léxico, estilo, desvios, padrões, etc.), consumido por pesquisadores da língua ou por uma aplicação computacional. A extração desse conhecimento exige que (a) a quantidade de texto (exemplos) seja grande, variada e representativa; (b) os textos estejam em formato adequado para que a extração se faça de forma automática e eficiente. Sistemas de manipulação de corpora permitem consultas otimizadas de ocorrências no corpus, estatísticas e, em alguns casos, a definição de padrões por meio de formalismos especiais (e.g. WordSmith[1], Lacio-Web[2],

---

1 http://www.lexically.net/wordsmith/

2 http://www.nilc.icmc.usp.br/lacioweb/

IMS-Corpus Workbench[3]). A construção de corpus, por outro lado, exige um estudo detalhado de fatores apontados na Lingüística de Corpus. Encontrar textos disponíveis em tipo e número adequados para o balanceamento do corpus, obter sua autorização de uso, disponibilizá-los de modo que sua manipulação seja eficiente, qualquer que seja o tipo de acesso, não são tarefas triviais. Se somarmos a isso a necessidade de etiquetar milhares de unidades de texto para ter informações mais sofisticadas, teremos um conjunto de fatores que fazem dessa tarefa uma empreitada de difícil realização. Quando o corpus é usado por outros programas para obtenção de conhecimento não para consumo humano, mas computacional, como os algoritmos de aprendizado, os corpora devem ser bastante volumosos, uma vez que os métodos mais promissores são os estatísticos. A partir de informações de freqüência de ocorrência de unidades em contexto, esses métodos podem gerar representações formais do conhecimento, ou ainda modelos probabilísticos (conjunto de probabilidades associadas a padrões), que podem então ser usados por aplicações variadas.

- **Léxicos computacionais** são estruturas de dados, em formato digital e adequado para consultas eficientes, contendo informações sobre o léxico (conjunto de unidades lexicais) de uma LN. A granularidade das unidades lexicais pode variar, porém é usual que tais unidades constituam palavras (seus lemas) e, eventualmente, multipalavras. Uma lista simples das palavras de uma LN não é suficiente para a maioria das aplicações. Informações fonéticas, morfológicas, sintáticas, semânticas e outras, dependendo de sua categoria, completam o conjunto de atributos de cada entrada de um léxico computacional genérico. Para uma estrutura que pode ser consultada muito freqüentemente (p.ex. num editor de texto), é fundamental que ela seja representada computacionalmente de forma a (a) ocupar pouco espaço, pois se trata de parte de um sistema maior; (b) permitir acesso em tempo ótimo. A tecnologia que permite isso é a dos Autômatos Finitos Mínimos, que são estruturas que armazenam palavras de forma que prefixos comuns são representados uma única vez, minimizando o espaço utilizado. Além disso, durante a busca, o tempo gasto é pequeno e praticamente constante para todas as palavras. Por outro lado, o preenchimento dos campos do léxico pode ser bem mais trabalhoso. Classes abertas de palavras não são passíveis de classificação automática, requerendo verificação manual após a aplicação de algoritmos. Informações de freqüência de palavras em categorias gramaticais distintas podem exigir seu cálculo em corpora (como é o caso do Português brasileiro, para o qual não se dispõe de estudos publicados sobre freqüência). Léxicos para apoio a processamento de corpora, para diferentes línguas, podem ser encontrados na web, como os do projeto Unitex[4] .

- **Ontologias** de domínio em formato digital têm hoje um papel muito importante para aplicações Web. Seu uso tem aumentado o poder de processamento

[3] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/
[4] http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/index.html

lingüístico em ambientes controlados. Por exemplo, em sites de venda de produtos eletrônicos, com possibilidade de alguma entrada lingüística, uma ontologia desse domínio permite que a intenção do usuário seja percebida com mais facilidade. Ao saber que uma TV é um "equipamento", por exemplo, a ocorrência desse termo após a menção à TV, resolve um problema de referência anafórica de difícil solução se não houvesse essa relação entre ambos. Métodos automáticos de extração de terminologia a partir de corpora, bem como ambientes de criação e manipulação de terminologias são hoje bem mais populares do que há poucos anos atrás. Ontologias de línguas naturais têm na Wordnet de Princeton[5], do inglês, seu mais importante exemplo. Trata-se de um conjunto de classes de sinônimos (*synsets*), formando uma rede de relações semânticas, e representada computacionalmente para permitir consultas por usuários humanos e por outras aplicações computacionais.

- **Gramáticas Computacionais** são formalismos para a representação de regras de formação de unidades da LN. O que as gramáticas definem pode variar desde sintagmas (nominais ou verbais) até sentenças. No primeiro caso, são úteis para a construção de *chunk parsers,* que detectam e muitas vezes classificam os sintagmas de um texto. Se desejamos uma análise mais completa, necessitamos de *parsers*; é o caso da revisão gramatical em um editor de texto, por exemplo. No contexto da geração de texto, uma gramática define sentenças superficiais a partir de formalismos de representação semântica.

Embora seja possível planejar a construção de recursos de forma independente das aplicações, como o trabalho envolvido é complexo e volumoso, é razoável que eles sejam adaptados à tarefa em questão. Sendo assim, para o processamento de textos, é comum não haver informações fonético-fonológicas no léxico. Não obstante, são vários os projetos de construção de recursos que pretendem servir a vários fins. Para o português brasileiro, por exemplo, registra-se o projeto PLN-BR[6], cujo propósito é exatamente a construção de um grande corpus do português, cujas composição, formatação, ferramentas de acesso e manuseio possam ser disponibilizadas na Web e compartilhadas por todos. A meta é fazer com que muito esforço e custos sejam recompensados pelo amplo proveito que for feito dos recursos construídos.

2. PLN para quê?

O interesse no processamento computacional da LN surgiu concomitantemente com o computador, no contexto de uma aplicação prática: a tradução automática. Na era pré-computador pessoal, já se cogitava o uso da LN na comunicação com a máquina, ainda que apenas para simplificar a vida dos poucos programadores e técnicos que tinham contato direto com ela. Nesse cenário, linguagem natural era sinônimo de

[5] www.wordnet.princeton.edu
[6] www.nilc.icmc.usp.br/plnbr/

mnemônicos, cujo processamento pouco desafio suscitava. Quando os computadores pessoais se popularizaram, o uso da LN pelos usuários foi posto no topo da pilha dos problemas urgentes. Por várias décadas, até o final dos anos de 1980, as pesquisas de PLN concentravam-se principalmente nas interfaces de aplicações diversas (que presumiam um usuário humano interagindo como fonte de entrada ou receptor da saída do sistema). Quando surgiu a internet e sua interface Web, as possibilidades de uso da LN tornaram-se ilimitadas. De outro lado, a disponibilidade de grande quantidade de texto na Web tornou-se uma fonte de conhecimento lingüístico de valor inestimável.

Em relação à demanda atual, são inúmeros os cenários que requerem ou se beneficiam do PLN. A começar pelos estudiosos da língua, que se beneficiam do processamento digital, capaz de disponibilizar milhões de textos para consulta rápida e eficiente. Sofisticados sistemas de manipulação de corpora possibilitam estudos das mais variadas naturezas: sobre freqüência, padrões, desvios, neologismos, semântica, etc.

No campo dos usuários leigos, a Web ocupa lugar de destaque. Além dos *browsers*, como Google, cuja função é buscar documentos e informações na internet, várias aplicações requerem a interação com o usuário, e qual a melhor forma humana de interagir, senão por meio da LN? A resposta a essa pergunta parece óbvia, mas a história do PLN tem mostrado o contrário. Durante as décadas de 1970 e 1980, pré-Web, muito se investiu nas pesquisas em interfaces em LN. Desde cedo se constatou a complexidade envolvida no tratamento robusto da LN, mesmo considerando que o usuário fazia apenas requisições limitadas, como perguntas sobre a previsão do tempo, por exemplo. No mesmo período, interfaces a base de ícones e menus, propostas pela Apple, ganharam popularidade e, de certa forma, contemplaram os anseios de uma interface mais "natural". O Windows da MS massificou esse conceito e aos poucos a interface em LN deixou de ser uma necessidade urgente. Dessa forma, já há algumas décadas, o conceito de eficácia na comunicação com o computador (ou seus sistemas) varia desde a linguagem de comandos, para especialistas, até a LN falada, passando por várias outras formas, visuais, icônicas, sub-linguagens naturais e, naturalmente, a LN escrita.

Dessa forma, deve ficar claro que a LN satisfaz vários quesitos para a comunicação que envolve humanos, porém não satisfaz outros quando um dos participantes é uma máquina. Porém, quando, além da interface, a LN aparece como elemento da própria aplicação, então o PLN se torna mais importante.

Na Web, onde usuários e aplicações se multiplicam sem fim, temos situações peculiares. A grande tarefa na Web é a busca por documentos e informações. Assim, o principal problema de PLN na Web se resume a uma forma lingüística cuja interpretação é a de uma pergunta ou consulta, mas cuja forma nem sempre obedece a esses padrões. Isso porque o usuário de hoje já se familiarizou com o uso de palavras e expressões que disparam os mecanismos internos de busca na Web. Acontece que para que a busca se dê de forma eficaz (altas precisão e cobertura) e eficiente (quase

instantaneamente), muitos fatores têm que ser considerados. O primeiro deles é que a Web é muito grande e, para satisfazer o usuário, é necessário separar muito bem o joio do trigo, ou seja, não basta retornar todos os documentos que tenham essa ou aquela palavra; é preciso distinguir os relevantes dentre eles. Isso é tarefa da área de Recuperação da Informação, mas hoje em dia já são muitos os trabalhos que propõem tratamento lingüístico da consulta – e conseqüentemente dos documentos – para melhorar o desempenho desses sistemas. Um exemplo desse tratamento é a consideração de sintagmas nominais (ao invés de palavras) na consulta e nos documentos, ampliando a noção de termos significativos. A idéia é estender o tratamento de palavras-chave da consulta em contexto. Outro recurso usado são listas de sinônimos e antologias, visando aumentar a cobertura. O desafio é manter muito baixo o tempo computacional, portanto, processamento muito complexo é proibitivo.

Como fonte de conhecimento, a Web tem feito parte de inúmeros experimentos com a LN, no âmbito do PLN e também da Lingüística de Corpus. Nunca tanto texto esteve disponível aos pesquisadores! Praticamente todos os gêneros estão amplamente representados em número, vários domínios e várias línguas. Esse cenário fortaleceu a área de Lingüística de Corpus ao mesmo tempo em que criou outras demandas. O próprio PLN sofreu mudanças metodológicas definitivas. Os métodos estatísticos, que necessitam de grandes amostras e de robustez computacional, voltaram com carga total e hoje são imbatíveis em algumas tarefas. Novos algoritmos de aprendizagem de máquina surgem constantemente e, aos poucos, aquela tarefa de construção manual de recursos lingüísticos (léxicos, gramáticas, etc.) tem sido desempenhada parcial ou totalmente pelo computador graças principalmente ao conhecimento implícito na Web. No entanto, a qualidade do conhecimento extraído depende diretamente da preparação ou pré-processamento do texto em estado bruto. E aí entra em cena o PLN. Informações morfossintáticas e semânticas associadas ao texto incrementam consideravelmente a extração do conhecimento, em número e em qualidade. Por exemplo, além de descobrir que X é mais freqüente que Y, é possível inferir que X ocorre mais no contexto C1 do que no contexto C2, sendo que, na forma sintática S, a acepção mais freqüente é A. Todas as tarefas básicas citadas anteriormente podem ser utilizadas para esse pré-processamento.

Em suma, não se faz PLN visando necessariamente à comunicação com o computador, como se almejava há algumas décadas. Hoje se faz PLN para vários propósitos, sendo que, para alguns, ele ocupa papel secundário e complementar, para outros, ele é fundamental. Decidir se e como usar PLN merece uma análise que transcende questões simples de tempo e recursos.

3. PLN para quem?

É digno de nota um fenômeno típico de um usuário de computador: quando percebe que pode usar a LN, passa a usá-la de forma bastante coloquial e mais próxima aos padrões da fala do que da escrita – o que torna o processamento mais complexo, ainda que isso possa soar contraditório. Por outro lado, quando percebe que o computador é incapaz de entender alguma requisição em LN, ele passa a ter uma atitude de descrença e desconfiança – o que contribui para o fracasso de qualquer interação. O fato é que o usuário comum não sabe (e nem deveria saber) das dificuldades envolvidas no PLN. O cenário ideal para o projetista do sistema é que o usuário percebesse as limitações da interface e fosse cooperativo a ponto de fornecer a entrada em LN exatamente da maneira ideal para o sistema, ou seja, usando o padrão culto, sem cometer erro algum, de forma não-ambígua e objetiva. Evidentemente isso é utópico e o que temos no comando do teclado, na realidade, é um usuário que geralmente sequer sabe muito bem o que quer.

Se o projetista de um sistema de PLN não tiver uma idéia absolutamente clara a respeito do usuário de seu sistema, ele corre dois tipos de risco: um, mais provável, de subestimar o usuário, e descobrir muito cedo que o usuário-alvo não existe; outro, menos provável, de superestimar o usuário, e descobrir que investiu demasiado no tratamento robusto de uma LN, que na verdade raramente ocorre na interação, ou seja, o usuário daquele sistema é pouco exigente em relação à língua e um subconjunto mais controlado da LN daria conta do recado. Diagnosticar esse último cenário é mais difícil, pois mesmo sem necessitar de todo poder da LN, o usuário se adapta a usá-la.

Usuários experientes são usuários exigentes. Experiência implica adaptação, habilidade, portanto, entre oferecer soluções paliativas em LN e oferecer ícones e menus, opte pelo último: o usuário já está adaptado a eles e saberá o que fazer.

Usuários iniciantes apresentam outro tipo de desafio: se não sabemos sobre eles, como trabalhar com hipóteses razoáveis? Preferências pessoais são muito importantes. Vários sistemas oferecem a possibilidade de configuração personalizada. Usuários ingênuos ou eventuais nem chegam a saber que isso é possível.

Há usuários especiais, como os lingüistas, que são os usuários dos ambientes de processamento de corpus, por exemplo. São as interfaces desses ambientes próprias para usuários com esse perfil? Se o são, porque tantos lingüistas não iniciados sentem-se como estranhos nesse mundo?

Questões sobre interface mereceram um fórum de discussão e pesquisa próprio, e não é intenção estendê-las aqui. No entanto, ao pensar na LN como elemento de interface, e a considerar o custo que isso envolve, há que se questionar sobre esses e outros fatores.

O paradigma da LN será definitivamente adotado quando muitos, senão todos, dos problemas ainda em aberto do PLN forem solucionados. E o que isso significa na prática? Significa robustez sintática e semântica. Significa tratar a LN em toda sua variedade sintática e ser capaz de atribuir significado correto quaisquer que sejam o contexto, o domínio e outras variáveis possíveis.

4. Conclusões e Perspectivas

A história do PLN tem nos revelado várias contradições, das quais devemos tirar algumas lições. Enfrentar o PLN com ingenuidade, achando que nossa facilidade em aprender e compreender a LN era igualmente transferível à máquina, não levou muito longe os anseios iniciais. Ignorar todo o conhecimento construído por lingüistas, simplificando demasiadamente o processo, tentando adaptá-lo à tecnologia vigente, também nos fez encontrar logo severas limitações. A tecnologia tem sido fundamental para o PLN. A demanda por soluções robustas, no entanto, tem pressa e não espera: se conhecimento lingüístico é essencial, mas não é possível obtê-lo de modo eficiente, então de alguma forma ele tem sido substituído, parcial ou totalmente, em nome da corrida tecnológica. Soluções parciais ou quase ótimas são preferíveis à ausência de solução. Unir conhecimento lingüístico e tecnologia, tirando proveito dos dois mundos, é hoje o grande desafio dos pesquisadores de PLN.

Em paralelo à história do PLN segue a história do usuário de computadores. E, a partir de um certo ponto, essas histórias se cruzam e uma afeta a outra de modo definitivo. O PLN foi afetado, a partir do momento em que o usuário vislumbrou a possibilidade de se comunicar em LN com o computador, por muitos fatores inerentes ao usuário: seus objetivos, suas expectativas, seu comportamento perante o computador, entre outros. O usuário, por sua vez, passou a ter uma atitude mais agressiva, exigindo mais, apontando alternativas para a comunicação com o computador. Disso resultou não uma LN típica desse contexto, mas várias facetas da LN, uma para cada cenário, entre tantos já detectados e outros inimagináveis.

A conseqüência para os pesquisadores e profissionais de PLN é que os problemas de PLN devem ser tratados da mesma forma com que interpretamos a LN: em contexto. Desde a construção de recursos básicos, como dicionários e gramáticas, de ferramentas de auxílio à escrita, até o tratamento de diálogos ou a tradução automática, em qualquer caso, não se pode perder de vista o contexto da aplicação, o usuário alvo e a tecnologia disponível.