A Criação de um Corpus de Sentenças Através de Gramáticas Livres de Contexto

Tiago Martins da Cunha[1], **Paulo Bruno Lopes da Silva** [2]

[1]Instituto de Humanidades e Letras – Universidade da Integração Internacional da Lusofonia Afro-Brasileira (UNILAB)

[2]Grupos de Redes de Computadores Engenharia de Software e Sistemas (GREat)

tiagotmc@unilab.edu.br, paulobruno@great.ufc.br

Abstract. *This work presents a new view towards linguistic data collection. In this paper, we propose to alter the direction in which the linguistic analysis is carried out. This lato sensu acknowledgement of linguistic information storage focus in reduce the ammount of space of storage and increase productivity for specific linguistic domains in corpus analysis. Therefore we propose the creation of specific grammar to generate possible sentences to compose a corpus. We present the methodology we used to compose our corpus of sentences and the tools required in the process. Using the creation of grammars to sentences generation, we produced over 10 thousand valid sentences per day. This sort of methodology showed itself very reliable and extremely productive towards specific domains.*

Resumo. *Este trabalho apresenta uma nova visão para com a coleta de dados linguística. Neste trabalho , propomos alterar a direção na qual a análise lingüística é realizada . Este reconhecimento lato sensu sobre o armazenamento de informação linguística foca em reduzir a quantidade de espaço de armazenamento e aumentar a produtividade em análise de corpus para domínios linguísticos específicos. Por isso, propomos a criação de gramáticas específicas para gerar possíveis sentenças para compor um corpus. Nós apresentamos a metodologia que usamos para compor nosso corpo de sentenças e as ferramentas necessárias no processo. Usando a criação de gramáticas para geração de sentenças,produzimos mais de 10 mil sentenças válidas por dia. Este tipo de metodologia se mostrou muito confiável e extremamente produtivo em relação a domínios específicos.*

1. Introdução

Os avanços na área de Processamento de Linguagem Natural (PLN) têm cada vez mais requerido recursos que possam servir como modelos de língua. Estes recursos de dados linguísticos têm expandido sua quantidade na busca incessante para tornarem-se representativos. Não somente o tamanho desses corpora, mas a sua riqueza de anotações são elementos que favorecem a variedade de possíveis estudos e aplicações sobre esses dados.

No entanto, a coleta de amostras da língua para contextos específicos seguindo os protocolos especificados na literatura da linguística de corpus [Sardinha 2004] pode não

ser suficiente para prover os dados necessários para análises satisfatórias. As mineração de dados disponíveis e o levantamento de dados através de entrevistas podem ser dispendiosos e até mesmo ineficazes na representação necessária.

Dessa forma, propomos a criação de corpus, em seu lato sensu, não partindo da coleta de sentenças amostrais, mas na geração dessas sentenças. Segundo Alencar [Alencar and de Ávila Othero 2012], toda língua regular pode ser representada por gramáticas independente de contexto.

Assim, propomos a criação de gramáticas, a partir da intuição de falantes nativos, que representem o contexto específico desejado. Essa criação de gramáticas deve seguir aspectos já elencados na literatura sobre a engenharia da gramática [Bender et al. 2008].

Na seção seguinte explicaremos o contexto específico que causou a busca frustrante seguindo a metodologia contemporânea da Linguística de Corpus. Também apresentaremos os princípios que seguimos na criação de um banco de dados a partir do formalismo da gramática livre de contexto (CFG).

2. Necessidade de dados

Os dados linguísticos mostram-se cada vez mais necessários para a interpretação do comportamento humano, assim como a possível interação humana com as máquinas. Em nossa pesquisa precisávamos elencar um conjunto de comandos úteis para a realização de atividades vinculadas a um recurso móvel (e.g. um telefone móvel).

Esse recurso, ao interpretar corretamente o comando, é capaz de realizá-lo de acordo com o pedido do usuário. No entanto, o sistema desse aparelho deve ser capaz de compreender os possíveis comandos para a realização de ações, e mesmo que alguma requisição não cadastrada previamente no sistema seja solicitada, esse sistema deve ser capaz de inferir o comando desejado pelo usuário e executá-lo.

O cadastramento e as inferências linguísticas devem ser elencados de acordo com um representativo banco de dados para cada domínio de ações realizáveis pelo aparelho móvel. Assim, deu-se início à incessante busca para a saturação das possibilidades de construções linguísticas para cada domínio.

A metodologia proposta pela Linguística de Corpus, em seu stricto sensu, propõe a coleta de amostras gerada em situações autênticas de uso da língua. Mas nesse contexto, o uso autêntico não se aplica devido ao fato de que esse tipo de interação até muito pouco tempo só existia no gênero da ficção científica.

Realizamos incessantes buscas por amostras em diferentes corpora que representassem comandos, instruções ou sentenças imperativas. Mas essa busca foi frustrada em ser representativa. Logo, decidimos que a única forma de saturar essas possibilidades seria por meio da construção intuitiva de sentenças.

A metodologia de construção de sentenças para o nosso corpus de comandos será apresentada na seção seguinte. Também serão apresentados os princípio que governaram o processo de criação de dicionários de Entidades Nomeadas no processo de construção das gramáticas e produção de sentenças.

3. Geração de Sentenças

A criação de sentenças para um contexto tão específico, como o estipulado em nossa pesquisa, tornou-se algo desafiador devido a seu caráter de inovação. Inicialmente, determinamos as ações a serem realizadas pelo aparelho celular. Ao todo, elencamos 98 ações a serem executadas no aparelho móvel que foram classificadas em 30 domínios. Como exemplo disso, obtivemos um domínio "mobile" englobando as tarefas relativas às funcionalidades de telefonia do equipamento.

O passo seguinte consistiu na produção manual de sentenças prototípicas para cada ação selecionada. Tal produção visava a criação de sentenças contendo padrões diversificados de estruturas sintáticas que atingissem o objetivo proposto pela ação.

Logo percebemos que, além das estruturas sintáticas, também havia a necessidade de trabalharmos destacando o reconhecimento de Entidades Nomeadas (NEs, i.e. *Named Entities*)[Wang et al. 2012]. Em geral, essa tarefa se encarrega de identificar expressões como nomes de pessoas, organizações, locais e assim por diante. No contexto da telefonia, as NEs das ações determinam os parâmetros que devem completar a funcionalidade a ser realizada pelo aparelho.

A terceira etapa para a construção de um corpus foi iniciada a partir da criação de gramáticas fazendo uso de um parser sintático construído com ferramentas oriundas do NLTK (*Natural Language Toolkit)*, biblioteca de ferramentas para o processamento de linguagem natural na linguagem de programação Python [Bird et al. 2009].

O programa desenvolvido para a geração de sentenças foi construído para receber uma gramática seguindo o formalismo CFG e consultar dicionários para efetuar a alimentação do léxico na geração de sentenças. Nessa etapa do processo, surgiram algumas preocupações e observações relacionadas à quantidade de sentenças e à estruturação do corpus gerado. Isto vai ser aprofundado nas próximas sessões.

3.1. Criação de Gramáticas

Antes de começarmos a falar sobre a complexidade das linguagens, faz-se necessário comentar a noção de gramática. Intuitivamente, pode-se afirmar que a gramática é um conjunto de regras que manipulam símbolos, isto é, a gramática é tida como um aparato que manipula um outro conjunto de símbolos com a intenção de transformá-los em cadeias, denominadas *strings*, de uma língua formal[Clark et al. 2013].

Caracterizada pelo pareamento entre nós terminais e não-terminais, o modelo de regras na CFG pode ser sintetizado pela notação $X \rightarrow Y$. No caso das nossas gramáticas, essas relações podem ser demonstradas pelo seguinte exemplo:

Tabela 1. Pareamento de nós terminais

V $\rightarrow$ telefonar
Prp $\rightarrow$ para
Det $\rightarrow$ o
N $\rightarrow$ Carlos

A regra anterior é lida não somente com símbolos, mas com palavras da língua portuguesa. A saída S, ou seja, a sentença gerada, será uma das possíveis combinações

sintáticas produzidas com esses elementos lexicais. Porém, no cuidado da produção, há sempre a preocupação de que o resultado seja uma sentença coerente em Língua Portuguesa.

Tabela 2. Geração de sentença com CFG

Regra	Sentença
S → V Prp Det N	telefonar para o Carlos

No caso de múltiplas gerações usando o formalismo da CFG para a construção de um corpus de estruturas de comandos, ampliamos o número de possibilidades com a variação de termos dentro das regras criadas na gramática. Como exemplo, aumentamos a variação lexical de verbos utilizados.

$$V \rightarrow \text{telefonar} \mid \text{ligar} \mid \text{chamar}$$

O novo resultado gerado a partir da regra anterior será ampliado. Isso ocorre devido à multiplicação do número de elementos lexicais.

Tabela 3. Geração de sentenças ampliada com variação verbal

Regra	Sentença
S → V Prp Det N	telefonar para o Carlos
	ligar para o Carlos
	chamar para o Carlos

Além disso, a concepção inicial das gramáticas deve lidar com outras características inerentes à linguagem natural, como as flexões de gênero e número para substantivos, adjetivos, artigos etc., bem como modo, tempo, gênero e número para os verbos do Português Brasileiro.

Essa problematização nos levou a criar subcategorias baseadas nas distinções de traços flexionais para tais classes gramaticais, como **Detms**(determinante masculino singular) e **Detfs**(determinante feminino singular), além de **Vinf**(verbo no infinitivo) e **Vimp**(verbo no imperativo). Por meio dessas subdivisões, evitamos a produção de sentenças sintaticamente agramaticais.

Tabela 4. Validade do uso de traços flexionais em determinantes e substantivos

Regra	Sentença	Validade
S → V Prp Det N	telefonar para o Carlos	Gramatical
	telefonar para a Carlos	Agramatical
	telefonar para o Carla	Agramatical
	telefonar para a Carla	Gramatical
S → V Prp Detms Nms	telefonar para o Carlos	Gramatical
S → V Prp Detfs Nfs	telefonar para a Carla	Gramatical

A atribuição de traços a nossa gramática exigiu uma mais criteriosa subdivisão quanto aos elementos lexicais. Estes elementos foram agrupado em dicionários.

3.2. Criação de dicionários

O passo seguinte para nossa produção de corpus a partir de gramáticas é a preocupação com as NEs do domínio móvel. Para essa etapa, tivemos de organizar sub-tarefas que atendessem aos requisitos de reconhecimento e classificação de elementos-chave para o reconhecimento da ação.

De forma geral, o Reconhecimento de Entidades Nomeadas consiste em identificar as NEs a partir de textos e classificá-las de acordo com determinadas categorias pré-definidas [Amaral and Vieira 2013]. Por exemplo, nomes próprios são divididos em pessoas, lugares, organização, entre outros elementos que nos remetam a referentes bem definidos. Em um sistema baseado em regras, por meio dos padrões linguísticos presentes no corpus, é possível identificar e classificar tais entidades.

Dentro do contexto de telefonia, as NEs foram extraídas e classificadas em 20 dicionários distintos, tais como contatos, lugares e estabelecimentos, tempo e data, por exemplo. Tais dicionários são essenciais, inicialmente, na identificação e classificação das sentenças que serão produzidas pelas gramáticas.

Dada, primeiramente, a ação de telefonar, retomamos à sentença "telefonar para o Carlos", gerada por meio da gramática. Nesse caso, inicialmente, foi possível identificar o padrão para o nome próprio "Carlos" por meio da estrutura estabelecida: **V Prp Det N**. O sistema de classificação baseado em regras atribui a essa palavra uma etiqueta referente à NE que a representa.

Tabela 5. Identificação de Entidades Nomeadas

Regra	**Sentença**	**Entidade Nomeada**
V $\rightarrow$ V Prp Det N	telefonar para o Carlos	telefonar para o [Carlos:n_contact]

Além disso, os sistemas de etiquetagem de NEs não classificam somente os nomes próprios, como exemplificado anteriormente, mas também se aplica a outros elementos lexicais e outras classes gramaticais, como numerais, adjetivos ou locuções.

Tabela 6. Identificação de NEs

Regra	**Sentença**	**Entidade Nomeada**
V $\rightarrow$ V Prp Num	telefonar para o 99999999	telefonar para o [99999999:num_contact]
V $\rightarrow$ V Adv	telefonar de novo	telefonar para o [de novo:tm_repeat]

Em um segundo momento, os dicionários não ajudaram somente no processo de classificação e categorização as NEs, mas também na população das sentenças. Fazendo o processo inverso, no decorrer da produção das gramáticas usando o formalismo CFG, substituímos alguns elementos chave pela própria *tag*, que por sua vez, apresentava uma lista de termos relacionados a ela. A partir de então, o acesso aos elementos dos dicionários permitem gerar sentenças populadas com diferentes entidades.

Da mesma forma como na geração anterior à utilização das NEs, a partir de então também mostrou-se necessária a subcatergorização de entidades nomeadas devido à falta de traços flexionais das entidades. Novamente houve a geração de sentenças gramaticalmente inválidas, tais como "liga para o Carla". A partir disso, assim como realizado com

Tabela 7. My caption

Regra	Sentença Gerada	Sentença Populada
V → V Prp Det Nm	ligue para o [:n_contact]	ligue para o [Carlos:n_contact]
		liga para o [:Carla:n_contact]
V → V Adv	liga [:tm_repeat]	liga novamente
		liga de novo
V → V Prp Num	telefonar para o [:num_contact]	telefonar para o [99999999:num_contact]

os nós da gramática usado em CFG, criamos subentidades que também refletem os traços de flexão de gênero e número nas gramáticas a fim de eliminar as incoerências.

Tal uso de subcategorias pautadas nas distinções de traços flexionais na produção das gramáticas refletem um direcionamento a uma implementação usando não simplesmente o formalismo CFG, mas tendendo uma construção que poderia ser melhor utilizada com a FCFG, modelo relativamente simples, de estruturas e traços não tipadas, o qual não dispões de metavariáveis e cujo único operador de expressões regulares é a disjunção lógica "|" [de ALENCAR 2012].

Entretanto, também é válido ressaltar que a subcatergorização de NEs e a construção dos dicionários requerem uma enorme quantidade de esforço humano além da dificuldade em ter uma boa cobertura de todos os tipos de entidades nomeadas [Neelakantan and Collins 2015].

4. Discussão

A criação de um corpus de gramáticas ou até mesmo de sentenças geradas através de gramáticas implica em uma atualização da literatura sobre a Linguística de Corpus. A criação e a gama de utilizações de um corpus desse tipo é governada por princípios que ainda não estão especificados na literatura tradicional.

Esse tipo de abordagem de criação de corpus agrega conhecimento e princípios de diversas disciplinas linguísticas e suas aplicações e teorias computacionais. Em nossa pesquisa, percebemos que o conhecimento sobre a engenharia da gramática foi de extrema importância para a organização e padronização dos dados a serem submetidos ao banco pelos colaboradores.

A formação e treinamento dos colaboradores para a criação desse tipo de corpus foi realizado no período de um mês, uma vez que os nossos 10 colaboradores, apesar de serem alunos ou graduados do curso de Letras, não dominavam ainda os princípios da teoria da gramática gerativa [Chomsky 1995].

O nosso processo de criação das gramáticas durou um mês para esgotar a necessidade da aplicação. O corpus gerado por essas gramáticas totalizou 450 mil sentenças. Ainda vale ressaltar que esse corpus esgotou as necessidades da aplicação que fará uso dessas sentenças, e não as possibilidades de cada domínio de ação que fora estipulado.

A atividade de criação de gramáticas e a geração das sentenças passou por vários momentos de reformulação de sua metodologia. Certas necessidades específicas, quanto à importância dos traços flexionais e escolha das NEs para cada domínio contemplado pela gramática, foram alguns percalços encontrados durante esta atividade.

Houve um consenso entre os participantes dessa atividade, sejam colaboradores ou professores, na necessidade de uma melhor padronização dos termos a serem utilizados nas gramáticas e uma maior dedicação no momento de elencar as NEs. Esse tipo de reflexão gerou algumas revisões para a padronização das gramáticas que compõem o corpus.

5. Conclusão

Percebemos que os métodos atuais de coleta de dados linguísticos propostos pela literatura não compreendem as tão recentes aplicações dentro do universo da PLN. Acreditamos que a metodologia proposta atende a diversas requisições de dados linguísticos que estejam restritos a domínios específicos de uso da língua.

A produtividade desse método, em relação a geração de sentenças, deve ser criteriosamente analisada na criação de suas gramáticas e defendemos que ela deve ser produzida de forma supervisionada. O processo de validação deve ser feito por falantes nativos que sejam contextualizados em relação ao uso da sentença no banco de dados.

Essa metodologia da criação de gramáticas pode favorecer e se valer de estudos descritivos da gramática. Da mesma forma, os dicionários utilizados podem ser favorecidos com a implementação de estudos lexicais como WordNet e ConceptNet.

Por fim, o estudo e a metodologia se mostram importantes para a exploração de novos contextos e fenômenos linguísticos, apresentando contribuições que beneficiem outras áreas de conhecimento, além de ajudar a renovação da literatura tradicional, e viabilizando uma nova concepção sobre a criação de um corpus.

Referências

Alencar, L. F. and de Ávila Othero, G. (2012). *Abordagens computacionais: da teoria da gramática*. Mercado de Letras.

Amaral, D. O. F. and Vieira, R. (2013). O reconhecimento de entidades nomeadas por meio do conditional random fields para a língua portuguesa. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, pages 59–68.

Bender, E. M., Flickinger, D., and Oepen, S. (2008). Grammar engineering for linguistic hypothesis testing. In *Proceedings of the Texas Linguistics Society X conference: Computational linguistics for less-studied languages*, pages 16–36.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python*. "O'Reilly Media, Inc.".

Chomsky, N. (1995). *The minimalist program*, volume 1765. Cambridge Univ Press.

Clark, A., Fox, C., and Lappin, S. (2013). *The handbook of computational linguistics and natural language processing*. John Wiley & Sons.

de ALENCAR, L. F. (2012). Donatus: uma interface amigável para o estudo da sintaxe formal utilizando a biblioteca em python do nltk. *ALFA: Revista de Linguística*, 56(2).

Neelakantan, A. and Collins, M. (2015). Learning dictionaries for named entity recognition using minimal supervision. *arXiv preprint arXiv:1504.06650*.

Sardinha, T. B. (2004). *Lingüística de corpus*. Editora Manole Ltda.

Wang, J., Liu, Z., and Zhao, H. (2012). Named entity recognition based on a machine learning model. In *Research Journal of Applied Sciences, Engineering and Technology*, pages 3973–3980.