

CAIO ESDRAS DE BRITO BEGOTTI

**O LATIM CLÁSSICO DE CÍCERO À LUZ DA LINGUÍSTICA
DE CORPUS: DESCRIÇÃO E IMPLEMENTAÇÃO DE
MÉTODO COMPUTACIONAL**

Monografia apresentada à disciplina Orientação Monográfica II do Curso de Letras (Bacharelado em Estudos Linguísticos), do Setor de Ciências Humanas, Letras e Artes da Universidade Federal do Paraná.

Orientador: Alessandro Rolim de Moura

CURITIBA
JULHO DE 2012

Sumário

Lista de Siglas	v
1 Introdução	1
1.1 Motivação	2
1.2 Por que Cícero?	3
1.3 Diferenciais tecnológicos	5
1.4 Objetivos	6
2 Fundamentação Teórica	9
2.1 Linguística Aplicada e Computacional	9
2.2 Linguística de Corpus	10
2.2.1 Histórico	10
2.2.2 Casos de uso	11
3 Metodologia	13
3.1 Corpora	13
3.1.1 Especificação	13
3.1.2 Escopo	14
3.1.2.1 Discursos	14
3.1.2.2 Retórica	16
3.1.2.3 Poesia	17
3.1.2.4 Filosofia	17
3.1.2.5 Cartas	18

3.1.2.6	Outros	18
3.2	Stopwords	19
3.2.1	Conjunções	20
3.2.2	Preposições	20
3.2.3	Pronomes	20
3.3	NLTK	21
3.3.1	Análise de Frequência	21
3.3.2	Concordâncias	22
3.4	Python	22
4	Análises	23
5	Resultados	24
6	Propostas de continuidade	25
	Referências	26
	Apêndice A – Códigos	28
A.1	Scrapper de dados e montador dos corpora	28
A.2	Filtros dos corpora	30
A.3	Parser e filtro de prenomes latinos	31
A.4	Analisador de frequências	32
A.5	Gerador de concordâncias	32
A.6	Protótipo de algoritmo de stemmer	32
	Apêndice B – Cronograma	35
B.1	Definição	35
B.2	Coleta	35
B.3	Montagem	35

B.4 Filtragem	36
B.5 Análise	36
Apêndice C – Anotações	37

Lista de Siglas

NLTK	Natural Language Toolkit
LC	Linguística de Corpus
API	Application Programming Interface
PHI	Packard Humanities Institute
LA	Linguística Aplicada
PLN	Processamento de Linguagem Natural
NLP	Natural Language Processing
BLEU	Bilingual Evaluation Understudy

1 Introdução

O projeto de monografia que aqui se apresenta foca na construção de *corpora* linguísticos de textos do período clássico do latim e em uma análise computacional desses *corpora*. O propósito por trás dessas duas tarefas é o de sugerir melhoras em métodos de ensino do latim e permitir uma melhor compreensão dessa língua de forma mais empírica, ou seja, uma compreensão baseada no uso que um dia foi dado à língua através da análise de textos reais em latim clássico.

Para tal, esse trabalho se baseará no projeto NLTK (Natural Language Toolkit), que consiste em um conjunto de rotinas de programação de computador, em forma de kit de ferramentas¹, para processamento de linguagem natural. Com a ajuda do NLTK e à luz da Linguística de Corpus (LC) procurarei evidenciar os usos mais comuns do latim clássico por Cícero e irei sugerir uma abordagem para uso disso em sala de aula.

Marcus Tullius Cicero, daqui para frente simplesmente Cícero, será o autor escolhido para a montagem dos *corpora*. Cícero é um dos autores que mais representam o período clássico do latim — tanto pela grande produção e variedade de estilos de seus textos quanto pelos seus discursos —, e devido a sua importância após a Renascença nada mais natural do que usá-lo nas análises desse trabalho.

Espera-se ainda, com esse trabalho, contribuir para o ainda pequeno grupo de linguistas computacionais no Brasil, especialmente os que apreciam e trabalham com o latim. A exemplo do que já foi feito com o inglês utilizando métodos semelhantes aos que aqui serão apresentados, pode-se ganhar bastante em sala de aula e na pesquisa com tal abordagem: aproximando a linguística de corpus à linguística computacional e às línguas naturais (SARDINHA, 2000b).

¹A descrição pode soar estranha em português e para não iniciados em programação; em bibliografias de computação ambos os termos aparecem simplesmente como API e toolkit, em inglês.

1.1 Motivação

É ponto pacífico que atualmente o ensino de latim se dá de forma bastante limitada em escolas e universidades brasileiras. Todavia, o latim clássico parece estar em uma espécie de retomada em países como Inglaterra (que vale dizer sempre foi um centro de referência em estudos clássicos) e Estados Unidos. Nos Estados Unidos, por exemplo, relatórios mostram o latim entre as oito línguas mais estudadas hoje no país, na frente do russo e do português, com crescimento de mais de 20% em matrículas nos últimos anos (FURMAN; GOLDBERG; LUSIN, 2010). Embora parte dessas matrículas ainda seja para cursos introdutórios, nos quais os alunos pouco aprendem sobre a língua e dependem exclusivamente dos professores, esse número é notável.

Muitas vezes o latim acaba sendo ensinado somente de cima para baixo, ou seja, dos professores aos alunos, pois os primeiros são uma fonte natural de conhecimento e autoridade. Assim, alunos de latim com frequência simplesmente tem que confiar na capacidade de julgamento dos mestres na escolha do melhor material a ser estudado. O que de fato faz sentido, pois os professores sabem avaliar a melhor linha de aprendizado. Mas embora os professores possam fazer boas escolhas quanto aos textos e ao vocabulário a serem estudados, tais escolhas serão sempre parciais ou limitadas, de modo que a criação de materiais e métodos que fundamentem tais escolhas de forma objetiva é sempre bem-vinda.

Por si só isso não representa um grande problema. Entretanto, por se tratar de uma língua sem usuários nativos — logo, poucos podem julgar se o que está sendo estudado de fato representaria a realidade dos usuários de uma língua de dois mil anos atrás —, ninguém na realidade sabe dizer com total certeza se, por exemplo, a seleção de textos e a escolha de determinado vocabulário de estudo são ou não adequados, ou ainda se eles são suficientemente representativos para os alunos ganharem proficiência mais rapidamente.

Isso é até aceitável em línguas estrangeiras modernas, pois o aluno tem mais contato com elas no seu dia-a-dia e pode sair de um beco sem saída com mais facilidade e sem recorrer à ajuda dos professores, mas com línguas clássicas como o latim e o grego, o prejuízo pode se tornar maior para o aluno que se dedicou por tanto tempo e então percebe que tomou um caminho de estudo não muito bom e perdeu uma melhor oportunidade de aprendizado. É claro que se espera dos alunos que eles possam compensar eventuais lacunas desse aprendizado através de estudo e leituras próprias, mas infelizmente não se

pode contar com isso sempre.

Em uma situação limite, um estudante de latim clássico poderia estar aprendendo um vocabulário que outrora pertenceu a um texto menor e que não corresponde à realidade linguística que ele buscava, ou que também não facilita a assimilação de idéias do estilo que procurava entender. Naturalmente, deve-se considerar aqui que o ensino de latim costumava focar em aspectos altamente literários e clássicos. Por exemplo, alunos de latim comumente leem Cícero após um ou dois anos de estudo, mas Cícero representa apenas uma fatia do todo que foi o latim em sua época. Deixa-se então o latim vulgar, falado pelo povo, para aulas de linguística românica, e textos religiosos para aulas sobre latim medieval ou eclesiástico.

Esse trabalho, obviamente, não tentará esgotar as análises linguísticas que se pode fazer em *corpora* em latim. Seu escopo será delimitado em torno do latim clássico por razão do tempo limitado para pesquisa, pela complexidade da tarefa de montar *corpora* maiores sem ajuda de terceiros e pela importância do latim clássico de Cícero. Também restarão, de qualquer forma, os métodos e as conclusões desse trabalho para que no futuro outros o levem adiante. Realizando-se as montagens e análises sobre os *corpora* de Cícero, já se terá conseguido avançar um bom caminho.

1.2 Por que Cícero?

Provavelmente se conhece mais a respeito da vida de Cícero do que de qualquer outra pessoa do mundo antigo, ainda assim poucas pessoas hoje saberiam dizer quem foi ele. Dessa maneira, cabe aqui uma breve biografia de Cícero, como narrada em detalhes por Petersson (2005) e Everitt (2001).

Nascido fora da cidade de Roma, Cícero foi levado ainda pequeno para lá. Foi em Roma onde estudou e cresceu até entrar no mundo político da cidade, quando começou a acompanhar, ainda bastante jovem, casos públicos no fórum. Aos trinta e poucos anos Cícero já era uma figura importante no meio político e jurídico da cidade. Por volta dos quarenta anos Cícero vira cônsul, o título civil mais alto que alguém poderia ter. Não é exagerado dizer que Cícero foi possivelmente o maior e melhor advogado e orador de Roma.

Embora personagem secundário na história geral do império, Cícero esteve ligado a acontecimentos marcantes como a conspiração de Catilina contra a república e a oposição à César quando este subiu ao poder. Foi Cícero quem trouxe parte do pensa-

mento filosófico grego para o mundo latino, e sua importância para o pensamento ocidental foi tamanha que a igreja católica o declarou um bom pagão, fato decisivo para que tantos dos seus escritos fossem preservados até os dias atuais. Cícero entrou para a história política e jurídica de Roma não por simplesmente vencer casos, mas por ser brilhante em suas exposições e pela sua capacidade de transformar o ato de falar em público em uma arma poderosa.

A influência e a importância de Cícero como pensador jamais serão suficientemente demonstradas, porém é evidente que o latim como língua deve muito a Cícero. Ele popularizou estilos de discurso, cunhou novas palavras (e limou da língua tantas outras), criou parâmetros de expressão em latim. De certa forma, pode-se dizer que Cícero está até os dias atuais impregnado no modo como pessoas públicas se expressam, por exemplo políticos e juristas. Obviamente Cícero não exerce tanto fascínio simplesmente porque foi Cícero, mas pela qualidade da sua expressão e seus estilos (FISHWICK, 2007).

É importante lembrar, entretanto, que Cícero não pode ser tomado como um autor que representa toda a grandeza do latim clássico sozinho. É necessário considerar isso por todo esse trabalho. O que será analisado pode muito bem nem mesmo representar toda a produção de Cícero de forma justa, vale dizer.

Neste trabalho será estudada uma das grandes características de Cícero, a chamada *elegantia*. *Elegantia* se traduz não simplesmente por elegância como seria de se esperar pelo cognato, mas como um refinamento lexical e uma perfeita escolha de palavras para expressar uma idéia. *Elegantia* é saber utilizar a palavra certa, no momento certo, de maneira adequada (PALMER, 1988). Kennedy e Wooten (2001) também lembram elogios do próprio Cícero a César pelo uso exemplar de vocabulário que César demonstra em seu *De Analogia*. Cícero declara que a escolha precisa de palavras ao falar e escrever é um dos vários requisitos para alguém ser um verdadeiro cidadão romano. Portanto, nada mais adequado do que tentar enxergar como isso de dava utilizando a LC.

Ajuda entender isto algumas passagens em um de seus textos mais conhecidos, *De Oratore*. Podemos ver a preocupação de Cícero com o modo rústico de se expressar de alguns romanos — tanto no sotaque falado quanto nas palavras grosseiras utilizadas por alguns —, como isso afetava a latinidade deles segundo Cícero e como isso devia ser expurgado da língua². É particularmente interessante o trecho em que Cícero destaca a importância de ser claro em um discurso, a necessidade de se evitar ambiguidades e conseguir passar o sentido literal do que se tem em mente empregando o uso exato de

²*De Oratore*, livro 3, parágrafo 42.

palavras³. É nesse texto que Cícero nos dá embasamento para esse trabalho: segundo ele, ao resumir os bons valores de oratória, um orador precisa ter cultura abrangente, que de uma forma ou de outra se mostra em seu texto ou discurso pela riqueza de seu vocabulário, pois não é possível esgotar todos os temas com um universo lexical reduzido⁴.

Embora já se tenha visto *corpora* de latim clássico — alguns, como o do Packard Humanities Institute (PHI), cobrindo todos os textos e discursos de Cícero —, pesquisadores e alunos até então somente formularam conclusões a respeito da estilística desses *corpora*, e não parecem ter desenvolvido pesquisas computacionais além disso; ou seja, processamento de linguagem natural que conseguisse chegar a conclusões a partir de análises estatísticas dos textos e que fosse útil, por exemplo, em sala de aula. Vale notar ainda que, segundo May e Craig (2002) a respeito dos famosos bancos de dados de latim do PHI⁵, “citar o PHI se tornou banal em qualquer artigo trabalhando com o uso de palavras ciceriano. Talvez a facilidade em obter dados do PHI faça com que conclusões retiradas deles sejam menos valorizadas, e ainda não se viu a publicação de trabalhos sobre o vocabulário ciceriano, e seu uso, feitos especificamente com a ajuda de computadores”⁶.

1.3 Diferenciais tecnológicos

Parece que um dos motivos pelos quais isso ocorre é a falta de um catálogo linguístico cientificamente montado com as palavras de fato mais usadas da língua. Já houve tentativas de se fazer isso para o latim, porém em uma época cujos poderes da análise linguística feita pelo computador ainda não eram conhecidos (DIEDERICH, 1939), ou foram tentativas focadas em outras línguas, como o inglês (ALMEIDA, 1997) e o espanhol (JACOBI, 2001). Outros autores, como Pellegrino (2009), no máximo se limitaram a criar listas de palavras bastante usadas por Cícero a serem estudadas para exames de latim de escolas americanas, sem uma maior racionalização sobre os *corpora* de Cícero e sem também elaborar em seus métodos. Um trabalho bastante superficial e parecendo-se mais com um manual de prova. Embora os resultados de todas essas experiências tenham sido utilizados de maneira distinta da que aqui se propõe, a motivação é similar: utilizar artefatos reais da língua para o aprendizado dela mesma.

³*De Oratore*, livro 3, parágrafo 49.

⁴*De Oratore*, livro 3, a partir do parágrafo 74.

⁵O mesmo parece se aplicar aos *corpora* de latim clássico do projeto Perseus, da Tufts University.

⁶Tradução livre do autor. No original, em inglês “[...] citation of this resource has now become commonplace in any article touching upon Ciceronian word usage. Perhaps because the ease of retrieval of these data makes conclusions drawn from them less prized, publication of specifically computer-based work on Ciceronian vocabulary and usage is still in the future”.

Criando-se hoje *corpora* linguísticos do latim em um formato reconhecido por linguistas computacionais, utilizando-se métodos documentados e comprovadamente eficientes para análise linguística, é possível melhorarmos o ensino de latim em escolas e universidades, além de permitir ao aluno — independente ou aconselhado — que tome o ensino da língua nas suas próprias mãos. Como bem diz Jacobi (2001), “essa forma de trabalho enfatiza o desenvolvimento da habilidade de descoberta nos alunos. A aprendizagem movida a dados posiciona o aluno no papel de descobridor ou de pesquisador, e o professor passa a ter como função primordial propiciar meios para que os alunos desenvolvam estratégias de descoberta”. Para tal, o volume de dados a ser analisado e utilizado nem mesmo precisa ser muito grande para se notar bons ganhos no aprendizado (WILLIS, 1990).

Ainda que existam diversos modelos, bases de dados e programas de computador para a análise de *corpora*, muitos possuem seus códigos fechados ou requerem licenciamentos, ou seja, não permitem o uso irrestrito deles por parte dos usuários e pesquisadores caso estes desejem, por exemplo, aplicar melhorias nesses softwares ou corrigir falhas nas análises por conta própria. Além disso, alguns são bastante caros — o que porém é compreensível dada a dificuldade de catalogar e organizar bases enormes —, e fogem da realidade acadêmica brasileira, ainda carente de recursos para latinistas⁷. Logo, uma solução aberta, sem nenhum tipo de restrição e preferivelmente barata se faz necessária. Essa solução, que será detalhada nas próximas seções, deve ainda aproveitar a possibilidade de utilizar textos sem amarras de copyright como no caso de textos clássicos em latim. Portanto, não está sendo proposta aqui uma revolução no jeito de trabalhar o latim com computação, mas simplesmente uma alternativa viável, barata e eficiente para alunos, pesquisadores e entusiastas brasileiros.

1.4 Objetivos

O objetivo primordial desse trabalho é permitir uma maior compreensão de textos clássicos em latim para os estudantes dessa língua através da criação de vocabulários específicos e automatizados por processamento de linguagem natural, cujas fontes serão os textos de Cícero, do período clássico do latim. De forma análoga a Almeida (2003),

⁷Softwares para análise linguística, por mais simples que sejam suas funcionalidades, não custam menos que algumas dezenas de dólares, quando não centenas. Bancos de dados prontos com *corpora* costumam também cobrar licenças de uso temporário que podem em alguns casos chegar a milhares de euros ou, embora “gratuitas”, acabam restringindo o uso e a pesquisa comercial através desses mesmos bancos de dados.

esperamos que alunos de latim consigam se beneficiar do mesmo modelo de aprendizado comprovado para o inglês. Dessa maneira, poderemos encontrar respostas para perguntas que não são facilmente comprováveis sem métodos como os propostos nesse trabalho, como por exemplo:

1. que vocabulário é mais adequado para determinados alunos ou pesquisadores em um contato inicial com o latim clássico?
2. entre os textos clássicos em latim deixados por Cícero, que características lexicais eles apresentam?
3. que tipo de palavreado era mais comum em textos filosóficos e literários, e qual era mais comum em textos informais (como em cartas)?

Através desses resultados formatados, alunos e professores poderão focar no aprendizado dos textos em si, não desperdiçando tempo em decorar vocabulários sem necessidade, assim como já foi feito para outras línguas.

Indiretamente, teremos um objetivo secundário que deriva do propósito anterior: a criação de *corpora* e um catálogo de *stopwords* do latim. Os *corpora* poderão ser utilizados para processamento de linguagem natural por terceiros, da mesma forma que serão utilizados nesse trabalho. O catálogo de *stopwords* poderá também ser utilizado por terceiros, e tentará ser abrangente, pois atualmente não se tem um catálogo completo para o latim assim como se tem para outras línguas.

Além de responder as perguntas acima e de chegar a alguns resultados indiretos já citados, será possível gerar uma lista ordenada de palavras mais comuns do latim clássico de Cícero que possam ser utilizadas em concordâncias. Chama-se de concordâncias os extratos de um ou mais textos, geralmente de uma única linha, contendo uma palavra procurada alinhada verticalmente com todas as suas ocorrências nos *corpora*. O uso disso permite ajudar não só em um ganho de vocabulário por estudantes do latim, mas também para pesquisadores que precisa de uma boa e imediata visualização do uso dessas palavras em textos reais.

A busca pelo domínio da leitura em uma segunda língua é um dos maiores objetivos de qualquer interessado em línguas, e isso não é diferente para o latim. Esse trabalho se propõe a ajudar nisso. A motivação inicial veio através de um projeto similar focado em língua inglesa na UNICAMP. Hoje transformado em livro e guia de estudo, o projeto

começou como um mero catálogo das palavras mais comuns do inglês, e desencadeou novas formas de estudar a língua que poderiam beneficiar também o latim (ALMEIDA, 2003).

A força que instiga esse trabalho a tentar responder essas questões é muito bem explicada por Sapir (1949)⁸, que também acreditava que “o vocabulário é um índice bastante sensível da cultura de um povo, e mudanças de significado, perda de palavras, a criação e o empréstimo de novas outras são todos fatos dependentes das histórias de suas próprias culturas. Línguas diferem grandemente pela natureza de seus vocabulários. Distinções que parecem inevitáveis para nós podem ser solenemente ignoradas em línguas que refletem um tipo de cultura totalmente diferente, ao passo que elas podem também insistir em distinções que são incompreensíveis para nós. Tais diferenças de vocabulário vão muito além de nomes de objetos como a ponta de uma flecha, armadura ou canhoneira. Elas se aplicam também ao mundo mental”.

Enfim, tal abordagem se mostra, além de eficiente, bastante realista, ao passo que aponta para os alunos somente o que vale a pena ser aprendido e aprendido não através de contextos artificiais ou idealizados da língua.

⁸Tradução livre do autor. No original, em inglês “vocabulary is a very sensitive index of the culture of a people and changes of meaning, loss of old words, the creation and borrowing of new ones are all dependent on the history of culture itself. Languages differ widely in the nature of their vocabularies. Distinctions which seem inevitable to us may be utterly ignored in languages which reflect an entirely different type of culture, while these in turn insist on distinctions which are all but unintelligible to us. Such differences of vocabulary go far beyond the names of cultural objects, such as arrow point, coat of armor or gunboat. They apply just as well to the mental world”.

2 Fundamentação Teórica

2.1 Linguística Aplicada e Computacional

Todo esse trabalho se sustenta sobre os princípios da Linguística Aplicada (LA). A LA por sua vez remete ao uso de ferramentas linguísticas destinado à resolução e análise de problemas reais das línguas, ou seja, é o estudo prático e não teórico delas. Particularmente, o subcampo da LA mais em voga atualmente é o da linguística computacional, que de forma simplificada é um campo de estudo que envolve diversas áreas como: ciência da computação, matemática, neurologia e a própria linguística, para citar os mais populares.

A história da linguística computacional está, naturalmente, ligada ao surgimento dos primeiros computadores logo após a II Guerra Mundial. Muitos problemas linguísticos hoje envolvem resolver algoritmos ou modelar dados — com frequência em grande escala — e foi exatamente para estes tipos de tarefas que os computadores foram criados. Dessa forma, não é surpresa alguma que ambas as áreas estejam hoje tão próximas; no mundo privado ao menos, pois na academia brasileira tenho a impressão de que ainda não alcançamos massa crítica de pesquisa. Houve trabalhos que tentaram medir essa ainda pequena produção em LC no Brasil, porém em níveis acadêmicos muito distantes dos estudantes de graduação (LADEIRA, 2010), e buscas em portais de pesquisa ainda mostram um número reduzido de grupos trabalhando nisso, alguns exclusivos e de acesso dificultado (CNPQ, 2012).

Do contato da linguística e da computação, então, chegamos ao Processamento de Linguagem Natural (PLN)⁹, que basicamente foca em problemas como análise automática de discurso, tradução por máquina, análises morfo-sintáticas, reconhecimento ou geração de fala¹⁰, etiquetamento de anunciados¹¹, lematização de palavras¹², gramática categorial,

⁹É digno de nota que ainda encontra-se maior diversidade bibliográfica procurando-se pelo acrônimo e termo em inglês, NLP, de Natural Language Processing.

¹⁰Speech-to-text e text-to-speech.

¹¹Tagging.

¹²Word stemming.

prosódia semântica de textos¹³, análises estatísticas, entre outros. Pelo fato de a linguística computacional e PLN serem áreas que atualmente quase sempre andam juntas, poucos conseguem distinguir diferenças entre elas. A cada dia surgem novas aplicações possíveis além das listadas e convencionou-se chamar tudo isso simplesmente “processamento de linguagem natural”.

2.2 Linguística de Corpus

2.2.1 Histórico

Dentro da linguística computacional encontramos afinal a LC, campo da linguística que estuda a língua como objeto vivo através de modelos estatísticos e amostragens de dados reais da língua, não representações idealizadas dela. É um campo da linguística totalmente empirista.

Para os aderentes a essa linha teórica, a língua deve se expressar por si mesma, pelos textos e discursos que surgem naturalmente de seus usuários, e daí o linguista pode aferir fatos e analisá-la. A LC, dessa maneira, evita trabalhar com textos criados artificialmente, e seus modelos são modelos de dados e não modelos teóricos de representação da língua; podem ser textos escritos ou falas transcritas, organizados em um dado formato e com categorização padronizada, além de poderem ser manualmente mapeados para um processamento posterior.

Antigamente a LC se encontrava em um estado bastante primitivo, muitas vezes sendo trabalhada por formas manuais. Hoje em dia ela é um campo de estudo que anda junto com a linguística computacional, dado o poder de processamento e automação de um computador a serviço do linguista (SINCLAIR, 1991). Ao se comparar a combinação do linguista com a máquina a um linguista solitário com papel e lápis, as vantagens são evidentes. Contudo, não houve falta de críticas à LC no passado (NORVIG, 2011), quando o uso da LC e análises estatísticas ainda eram tão primitivos que não eram inteiramente compreendidos, e a coleta de dados e a medição deles era vista quase como irrelevante cientificamente. Nada mais longe da verdade. A linguística não é uma ciência inteiramente teórica. Fazer ciência é desenvolver teorias sobre fatos, e fatos não surgem espontaneamente. A LC tenta ajudar no surgimento desses fatos, ainda que eles se pareçam muitas vezes com amontados de números e estatísticas, ou que modelos da LC não sejam universais para todas as línguas; não são e nunca serão.

¹³Sentiment analysis.

2.2.2 Casos de uso

Somente nas últimas décadas a LC tomou o palco das ciências das línguas. Por exemplo, foi somente por volta dos anos 70 que os primeiros *corpora* de línguas faladas foram montados e analisados digitalmente¹⁴. Portanto, não é de surpreender que tão pouco a utilizem ainda e, particularmente no Brasil, exista certa falta de tradição nesse ramo da linguística aplicada.

Hoje, virtualmente todas as ferramentas digitais de tradução, análise linguística e construção de dicionários utilizam a LC como base. Grandes empresas, como o Google e a IBM, utilizam seu poder computacional e infra-estruturas gigantescas para, através de trilhões de fontes de textos disponíveis na Internet — um amontoado de *corpora* de domínio público esperando para ser analisado —, construir conversores de texto para fala, analisadores sintáticos, tradutores automáticos etc.

No caso dos tradutores automáticos, por exemplo, é graças à LC que recentemente o serviço Google Translate passou a suportar a tradução de textos em diversas línguas do e para o latim, já que seria quase impossível implementar de forma fixa todas as regras gramaticais do latim em um tradutor automático. Embora os resultados de ferramentas como o Google Translate ainda não pareçam muito profissionais, o fato é que traduções baseadas em estatísticas de *corpora* tendem a apresentar melhores resultados que as baseadas em regras linguísticas fixas (OCH, 2005). A qualidade dos resultados está diretamente relacionada ao volume de dados a serem processados. É por isso que muitas vezes textos de serviços como o do Google ainda mostram ter uma sintaxe estranha. Quanto maiores e mais abrangentes forem os *corpora*, melhores serão as traduções¹⁵.

O uso da LC se mostrou também bastante útil para filólogos e para a linguística histórica. Devido principalmente ao poder de processamento dos computadores atuais, e a criação de novos algoritmos para análise de LC, linguistas trabalhando na construção de árvores de famílias linguísticas conseguem testar hipóteses a respeito da origem comuns das línguas, do indo-europeu e finalmente solidificar uma base científica à filogenia (CLACKSON, 2007). A filogenia, portanto, hoje depende sobremaneira da pesquisa em LC e análise matemática de dados linguísticos.

Entretanto, os maiores casos de uso da LC ainda são em língua inglesa. Ocorre

¹⁴Os dicionários *Collins COBUILD English Language Dictionary* e *The American Heritage Dictionary of the English Language* foram os primeiros a utilizar a LC em sua construção algumas décadas atrás.

¹⁵Uma das formas mais populares para avaliação de traduções automáticas baseadas em estatísticas de *corpora* é a pontuação BLEU, proposta pela IBM, que indica o grau de inteligibilidade de um texto traduzido por um computador comparando-o a um traduzido por um ser humano, atestando sua qualidade.

que avanços computacionais foram feitos primariamente em inglês nas últimas décadas, e por várias razões políticas e econômicas, a língua inglesa ainda é considerada a língua *de facto* da LC. À princípio isso não é um problema, porém é comum ler e confundir “LC do inglês” simplesmente como sinônimo de “LC”, o que não é verdade¹⁶. Apesar de ser extremamente difícil encontrar pesquisa em LC bem feita com, por exemplo, línguas africanas, orientais, e até mesmo com o latim, isso não significa que é impossível ou que não deveria ser encorajado. O inglês não pode ser um sinônimo exclusivo de LC como é hoje. É justamente essa abordagem que se apresentará nesse trabalho.

¹⁶Mesmo para quem não é iniciado em LC, uma rápida busca no *LRE Resources Map* em <http://www.resourcebook.eu/LreMap/faces/views/resourceMap.xhtml> prova isso falso, por exemplo.

3 Metodologia

A metodologia desse trabalho se sustenta sobre um pressuposto bastante simples, porém muitas vezes renegado em LC: a reprodutibilidade do método para se chegar aos mesmos dados apresentados. Infelizmente ainda se encontra trabalhos de LC e linguística computacional que simplesmente explicam dados sem detalhar como eles foram obtidos, como foram processados e com que critérios foram trabalhados.

Não é fato raro, também, encontrar pesquisas feitas utilizando programas de computador desenvolvidos por terceiros, aos quais nem os pesquisadores que os apresentam tem acesso total (*i.e.* ao código fonte do programa). É um ponto central deste trabalho a crença de que ciência com dados verdadeiramente reproduzíveis é ciência feita com códigos fontes e dados abertos, de uso irrestrito¹⁷. Por isso, todo o código desenvolvido para este trabalho será de domínio público, afim de que outros estudantes e pesquisadores possam aprender com o que foi pesquisado e que possam escrutinar o método proposto aqui.

3.1 Corpora

3.1.1 Especificação

Os *corpora* de Cícero serão construídos de acordo com as regras de formatação e organização de outros *corpora* encontrados no projeto NLTK, que atualmente possui cerca de quarenta *corpora* prontos para uso. Isso facilitará bastante o processamento dos textos e também contribuirá com o projeto, para aumentar — criar, na realidade — sua base de textos em latim. Tentaremos também, se o tempo permitir, iniciar o etiquetamento gramatical dos *corpora*, mesmo que de forma rudimentar.

O escopo dos textos escolhido para compor os *corpora* desse trabalho abrange o

¹⁷É digno de nota um excelente editorial sobre esse tema, originalmente publicado pela revista Nature, reproduzido em <http://arstechnica.com/science/news/2012/02/science-code-should-be-open-source-according-to-editorial.ars> pelo portal Ars Technica. No editorial é possível encontrar uma forte argumentação para o uso de plataformas *open source* de computação para o trabalho científico.

período clássico da literatura latina, da qual Cícero é um dos maiores expoentes, senão o maior. A escolha de Cícero se deu por sua produção ter sido bastante variada, tendo ele escrito desde tratados até cartas pessoais, de textos jurídicos até filosofia. Cícero teve uma grande influência no pensamento moderno, e é ainda um dos autores mais estudados em cursos de latim, o que o torna ideal para uma análise linguística do latim.

Quanto ao tamanho dos *corpora*, que indiretamente determina a utilidade deles, em um teste preliminar chegou-se à marca de 80.000 palavras. No teste, apenas alguns textos de Cícero facilmente encontráveis publicamente foram processados, sem muito método. De acordo com estimativas de Sardinha (2000a), isso já garante a esse trabalho *corpora* de tamanhos razoáveis. O tamanho final dos *corpora* deverá ser muito maior que esse número, porém.

3.1.2 Escopo

Nesta seção encontra-se todo o escopo dos *corpora* para análise deste trabalho.

As listas a seguir foram construídas a partir de catálogos populares de obras de Cícero — como a *Loeb Classical Library*, publicada pela Harvard University Press, e o próprio PHI —, porém se notará a ausência de alguns textos. Entre estes textos está *Pro Acilio*, *Pro Gaio Antonio*, *De Temporibus Suis* e alguns outros que são desnecessários citar no momento. Tais textos são atribuídos indiretamente à Cícero pois estão completamente perdidos e pode-se somente encontrar referências a eles mas não ao seu conteúdo. É comum encontrar estas obras perdidas sob títulos como *Ciceronis Orationum Deperditarum*, e para mais detalhes recomenda-se ver Nisard (1869), Müller (1879) e também Harrison (1990).

É necessário também considerar textos dos quais temos somente fragmentos, como é o caso de *Hortensius*, *De Consulatu Suo* e algumas orações. O fato de serem apenas fragmentos influencia pouco nas análises desse trabalho, uma vez que caso surjam novas partes dos textos elas poderiam ser facilmente incorporadas aos outros *corpora*. Por isso, não há indicações explícitas de fragmentos nas listas a seguir.

3.1.2.1 Discursos

Os discursos a seguir incluem tanto discursos de caráter jurídico quanto discursos políticos e de interesse público. Alguns, como *De Domo Sua* (que trata de uma petição pessoal de Cícero a um colegiado), são de difícil categorização pois podem ser interpretados tanto como discursos jurídicos quanto de outra natureza, ao mesmo tempo. Por essa razão,

procurou-se evitar divisões excessivas entre os textos.

TODO: "kytztler fala em 58 discursos conservados (considerando, obviamente as 4 catilnarias, por exemplo) - bote?"

- De Domo Sua
- De Haruspicum Responsis
- De Lege Agraria
- De Provinciis Consularibus
- Divinatio In Caecilius¹⁸
- In Catilinam
- In Pisonem
- In Vatinius
- In Verrem
- Orationum Deperditarum
- Orationum Incertarum
- Philippicae
- Post Reditum In Quirites¹⁹
- Post Reditum In Senatu
- Pro Aemilio Scauro
- Pro Archia Poeta
- Pro Aulo Caecina
- Pro Aulo Cluentio Habito
- Pro Cornelio Balbo

¹⁸Embora seja um texto de caráter claramente jurídico, no qual há um embate para se decidir entre quem será o acusador no caso contra Verres, Craig (1985) acredita que *Divinatio In Caecilius* pode ser compreendido também como um texto de retórica, principalmente pelo seu gênero de *divinatio* explicitado já no título.

¹⁹Também aparece como *Post Reditum Ad Populum* em algumas compilações.

- Pro Gaio Rabirio
- Pro Gnaeo Plancio
- Pro Lege Manilia
- Pro Ligario
- Pro Lucio Flacco
- Pro Lucio Murena
- Pro Marcello
- Pro Marco Caelio
- Pro Marco Fonteio
- Pro Marco Tullio
- Pro Milone
- Pro Publio Quinctio
- Pro Publio Sulla
- Pro Quinto Roscio Comoedo
- Pro Rabirio Perduellionis Reo
- Pro Rabirio Postumo
- Pro Rege Deiotaro
- Pro Sestio
- Pro Sexto Roscio Amerino

3.1.2.2 Retórica

- Ad Marcum Brutum Orator
- Brutus
- De Inventione
- De Optimo Genere Oratorum

- De Oratore
- De Partitione Oratoria²⁰
- Rhetorica Ad Herennium²¹
- Topica

3.1.2.3 Poesia

- Arati Phaenomena
- Arati Prognostica
- Carmina
- De Consulatu Suo

3.1.2.4 Filosofia

TODO: kytzler: dos 20 conhecidos, possuímos 13

- Academica Priora²²
- Academica Posteriora²³
- Cato Maior De Senectute
- Consolatio
- De Amicitia
- De Divinatione
- De Fato

²⁰Aparece com frequência também como *Partitiones Oratoriae*, no plural.

²¹Autoria disputada há muito tempo mas que certamente não é de Cícero, embora tanto o PHI quanto a Loeb Classical Library possuam o texto em seus catálogos como pertencendo a Cícero. Por completude, e devido ao seu estilo e vocabulário muito próximo ao de Cícero, o texto será considerado. Para mais detalhes, ver a introdução de Harry Caplan à edição da própria Loeb.

²²Também citado como *Catulus*, são fragmentos da segunda parte do primeiro livro que compõe o *Academica* de Cícero.

²³Também citado como *Lucullus*, é a segunda parte do *Academica*, aproximadamente quatro vezes maior do que restou da primeira parte. Para mais detalhes ver a introdução à tradução de Charles Duke Yonge (1875).

- De Finibus Bonorum Et Malorum
- De Legibus
- De Natura Deorum
- De Officiis
- De Re Publica
- Hortensius
- Paradoxa Stoicorum
- Timaeus
- Tusculanae Disputationes²⁴

3.1.2.5 Cartas

- Epistulae Ad Atticum
- Epistulae Ad Brutum
- Epistulae Ad Familiares
- Epistulae Ad Quintum Fratrem

3.1.2.6 Outros

- Commentarii Causarum²⁵
- De Iure Civ. in Artem Redig.²⁵
- Epistula ad Octavianum²⁶
- Epistulae Fragmenta²⁶

²⁴Do gênero *consolatio*, aparece com frequência também como *Tusculanae Quaestiones*. Não confundir o nome do gênero com o texto falsamente atribuído a Cícero. Para mais detalhes ver o excelente *The Pseudo-Ciceronian Consolatio* de Evan Taylor Sage (The University of Chicago Press, 1910).

²⁵Existem apenas pequenos fragmentos, atestados em *S. Hieronymi Presbyteri Opera* de Lardet e em *Iurisprudentiae Anteiustinianae Reliquiae* de Huschke, Seckel e Kübler, respectivamente.

²⁶Embora façam parte de coleções de corpora como a do PHI, são textos espúrios. Entretanto, é curioso que continuem a constar em tais coleções de textos de Cícero se são reconhecidamente falsos. Por exemplo, em *The Text of the Pseudo-Ciceronian Epistula Ad Octavianum*, W. S. Watt — responsável inclusive pelo texto que está no PHI — parece comentar mais sobre a relação entre os manuscritos falsos que levam ao texto que conhecemos do que sobre a autenticidade deste.

- Facete Dicta
- In Sallustium Crispum²⁶
- Incertorum Librorum
- Philosophicorum Librorum

3.2 Stopwords

O catálogo de *stopwords* é uma dependência natural dos *corpora* pois é ele que filtra resultados indesejados ou que poderiam eventualmente poluir os dados. *Stopwords*, do ponto de vista de um computador, nada mais são que palavras ou agrupamentos²⁷ de palavras que interrompem o processamento da linguagem e modificam esse processamento de acordo com alguma regra pré-estabelecida. Toda língua possui uma lista de *stopwords* para processamento, porém o latim ainda não conta com uma verdadeiramente utilizável.

Embora pudessem ter sua criação automatizada por algoritmos de textos, catálogos de *stopwords* geralmente são construídos manualmente por linguistas. Logo, não existe um catálogo definitivo ou verdadeiramente completo para uma determinada língua, somente mais abrangentes e menos abrangentes, a depender do uso que será dado a eles.

Em computação, *stopwords* não são utilizadas somente para filtrar resultados indesejados ou que não são o foco da pesquisa linguística. Muitas vezes são termos repetitivos que levam a uma pior performance de um sistema e que se não forem ignorados diminuem a legibilidade dos resultados. Por exemplo, é bastante comum filtrar-se preposições, conjunções e pronomes, uma vez que eles pouco influenciam na maioria das análises computacionais de um texto e representam um conjunto de termos funcionais — de pouca relevância na maioria das análises — que podem não só poluir os dados do ponto de vista linguístico (SARDINHA, 2000a) mas também dificultar a análise computacional (MAKREHCHI; KAMEL, 2008). Todavia, é claro que tais filtros podem ser desligados se tais termos funcionais da língua forem de fato o alvo da pesquisa.

Para o latim será criado, sob orientação, um catálogo geral de *stopwords* (no modelo clássico, uma lista finita de termos), e catálogos menores onde se encontrarão somente preposições, ou somente pronomes e assim por diante, para que os alunos e pesquisadores possam refinar melhor o uso de *stopwords* em latim. As listas a seguir

²⁷A linguística computacional também se refere a isso como *n-grams*, uma sequência de *n* termos textuais a serem processados conjuntamente e não individualmente.

foram elaboradas pelo autor (LEUMANN; HOFMANN; SZANTYR, 1977) e (KÜHNER, 1912), considerando resultados prévios desse trabalho e comparando os filtros com de outros autores²⁸.

3.2.1 Conjunções

Conjunções a serem filtradas: ac, adque, alii, an, annon, antequam, ast, at, atque, atque, atqui, aut, autem, cum, donec, dum, dummodo, enim, ergo, et, etenim, etiamsi, etsi, igitur, nam, namque, ne, nec, necdum, necne, neque, neve, ni, nisi, nonnisi, postquam, priusquam, prout, quam, quamquam, quantumvis, quasi, que, quia, quid, quo, quod, quominus, quomodo, quoniam, sed, seu, si, sicut, sive, tamen, tametsi, ut, utrum, utrumnam, vel, verumtamen.

3.2.2 Preposições

Preposições a serem filtradas: a, ab, abante, absque, ad, ante, apud, aput, cata, circa, circiter, circum, cis, contra, cum, de, dextra, e, ex, extra, in, infra, inter, intra, iuxta, laude, ob, palam, per, pone, post, prae, praeter, pro, prope, propter, quando, re, sine, sub, subter, super, supra, trans, ultra.

3.2.3 Pronomes

Pronomes a serem filtrados: alicui, alicuius, alicujus, aliquae, aliquam, aliquarum, aliquas, aliquem, aliquibus, aliquid, aliquis, aliquorum, aliquos, chodchod, cuidam, cuius, cuiusdam, cuiusquam, cuiusque, cujus, cujusvis, ea, eae, eam, earum, eas, ecquis, ego, egomet, ei, eis, eius, eo, eorum, eos, eum, eundem, hae, haec, hanc, harum, has, hi, hic, hii, his, horum, hos, huic, huius, hunc, id, ii, illa, illae, illam, illarum, illas, ille, illi, illic, illis, illius, illo, illorum, illos, illum, ipsa, ipsae, ipsam, ipsarum, ipsas, ipse, ipsi, ipsis, ipsius, ipso, ipsorum, ipsos, ipsum, is, ista, istae, istaista, istam, istarum, istas, iste, isti, istius, isto, istorum, istos, istud, istum, me, meamet, memet, meus, mi, mihi, mihimet, nemo, nihil, nobis, nobismet, nos, nosmet, noster, nostra, nostrae, nostram, nostrarum, nostras, nostri, nostris, nostro, nostrorum, nostros, nostrum, qua, quadam, quae, quaedam, qualarum, quale, qualem, quales, quali, qualia, qualibus, qualis, qualium, quam, quandam, quarum, quas, quem, quemcumque, quempiam, quemquam, quendam,

²⁸O projeto Perseus possui um catálogo de stopwords do Latim com cerca de 90 termos, porém os critérios para as escolhas feitas pelos responsáveis pelo projeto são desconhecidos, além de o próprio catálogo parecer incompleto.

qui, quibus, quibusdam, quicquam, quicquid, quicumque, quid, quidam, quidnam, quidquid, quilibet, quinam, quippiam, quis, quisnam, quispiam, quisquam, quisque, quisquis, quivis, quo, quod, quodam, quodcumque, quoddam, quorum, quorundam, quos, quoscumque, quosdam, sapsa, se, semet, semetipse, sese, sibi, suarum, sui, suos, suus, te, tibi, tu, tuae, tuam, tuarum, tuas, tue, tui, tuis, tuo, tuorum, tuos, tute, tutemet, tuus, unaquaeque, unicuique, uniuscuiusque, unumquemque, unusquisque, uterque, utraque, utrique, utrisque, utrumque, vester, vestra, vestrae, vestram, vestrarum, vestras, vestri, vestris, vestro, vestrorum, vestros, vestrum, vobis, vobismet, vos, voster, vostrae, vostram, vostrarum, vestras, vestri, vestris, vostro, vestrorum, vestros.

3.3 NLTK

Nesse trabalho será utilizado o NLTK, um conjunto de ferramentas de programação voltado ao processamento de linguagem natural através de computação. Atualmente o NLTK é um dos projetos de maior sucesso para esse fim, pois, além de ser livre de licenças e custos, é bastante abrangente, cobre diversas línguas oficialmente e é disponível já com diversos *corpora* para testes. O projeto possui livros e publicações sobre seu funcionamento, tem mais de dez anos de desenvolvimento e é utilizado em universidades há algum tempo, tornando-o ideal para uso, uma vez que se provou estar em constante evolução acompanhando novas descobertas no campo de processamento de linguagem natural (BIRD; KLEIN; LOPER, 2009).

3.3.1 Análise de Frequência

Será usado o método de análise de frequência estatística como forma básica de obter os resultados de vocabulários dos *corpora*. Análise de frequência é uma forma de, através de pequenos experimentos automáticos (utilizando-se programação e tentando-se obter combinações de um termo em um texto, por exemplo), descobrir a proporção de uso de uma palavra qualquer em relação ao todo de um corpus específico. Essa é uma das atividades primordiais da LC, muitas vezes chamada simplesmente de estatística lexical por aderentes da linguística quantitativa.

Para chegar ao ponto de usar uma análise de frequência, é preciso antes passar pela montagem dos *corpora*, da criação de filtros e estipular outros parâmetros do processamento. Porém, uma vez que isso é feito, a análise de frequência se mostra relativamente simples e bastante proveitosa. É através da análise de frequência, por exemplo, que se

pode tentar confirmar as Leis de Zipf para o latim clássico de Cícero, que estabelecem relações diretas entre a ordem de palavras em uma lista estatisticamente ordenada e a frequência de suas ocorrências em um corpus suficientemente grande. Em outras palavras, as leis de Zipf indicam que um número baixo de termos da língua corresponde a maior porção do uso efetivo da língua (TESITELOVA, 1992).

3.3.2 Concordâncias

TODO

3.4 Python

Tanto o NLTK quanto ferramentas auxiliares desse trabalho serão feitos utilizando a linguagem de programação Python. A escolha dessa linguagem de programação se dá, entre outros motivos, pelo fato de o NLTK utilizá-la internamente para processar línguas naturais. Também, a facilidade de leitura dos código-fonte da linguagem por pessoas que não são da área de computação é um fator chave (PERKINS, 2010). Pelo fator de Python ser uma linguagem de alto nível, ou seja, de alta abstração ao se programar com ela, o desenvolvimento de programas é mais rápido com ela e permite que ele seja menos críptico para terceiros que precisem analisá-lo.

Utilizar a linguagem de programação Python junto com o NLTK vai permitir uma autonomia muito grande para se criar, analisar e trabalhar de forma geral os *corpora* de latim clássico. Algumas alternativas para análise linguística de *corpora*²⁹ incluem os programas WordSmith Tools, Perseus Hopper (a versão web do projeto Perseus), Diogenes, Wmatrix, LanguageWare da IBM e LIWC. Infelizmente, quase todos esses programas possuem limitações tecnológicas, quando não de preço também. Utilizando a linguagem Python junto com o NLTK se tem uma possibilidade de desenvolvimento, customização e aprendizado muito grande. Digamos que se precise corrigir algum componente interno de um desses softwares, seja porque o erro é aparente, seja porque o pesquisador suspeita do seu comportamento. Isso não é possível com eles, mas é com a linguagem Python e o NLTK, tornando estes até mesmo mais cientificamente confiáveis.

²⁹Em <http://linguistlist.org/sp/GetWRListings.cfm?WRAbbrev=Software> há uma lista mais completa.

4 Análises

TODO: explicacao dos codigos escritos com python e nltk

5 Resultados

Propostas de uso em ambiente acadêmico e para aprendizagem do latim.

6 Propostas de continuidade

Lematização (stemming Porter), etiquetamento básico dos corpora.

Referências

- ALMEIDA, R. *Palavras mais comuns em inglês*. [S.l.]: Dicas-L, 1997. http://www.dicas-l.com.br/arquivo/palavras_mais_comuns_em_ingles.php. Acessado em 5 de novembro de 2011.
- ALMEIDA, R. *As Palavras Mais Comuns da Língua Inglesa*. Campinas: UNICAMP, 2003.
- BIRD, S.; KLEIN, E.; LOPER, E. *Natural Language Processing with Python*. Beijing: O'Reilly Media, 2009.
- CLACKSON, J. *Indo-European Linguistics: An Introduction*. Cambridge: Cambridge University Press, 2007.
- CNPQ. *Diretório dos Grupos de Pesquisa no Brasil*. [S.l.]: Conselho Nacional de Desenvolvimento Científico Tecnológico, 2012. <http://dgp.cnpq.br/buscaoperacional/>. Acessado em 14 de fevereiro de 2012.
- CRAIG, C. P. *Dilemma in Cicero's Divinatio in Caecilius*. Baltimore: The Johns Hopkins University Press, 1985.
- DIEDERICH, P. *The Frequency of Latin Words and Their Endings*. Chicago: University of Chicago Press, 1939.
- EVERITT, A. *Cicero: the life and times of Rome's greatest politician*. New York: Random House, 2001.
- FISHWICK, M. W. *Cicero, classicism, and popular culture*. Binghamton: The Haworth Press, 2007.
- FURMAN, N.; GOLDBERG, D.; LUSIN, N. *Enrollments in Languages Other Than English in United States Institutions of Higher Education, Fall 2009*. New York: Modern Language Association of America, 2010. http://www.mla.org/pdf/2009_enrollment_survey.pdf. Acessado em 22 de novembro de 2011.
- HARRISON, S. J. *Cicero's De Temporibus Suis: The Evidence Reconsidered*. Wiesbaden: Hermes, 1990.
- JACOBI, C. Dissertação de mestrado. *Linguística de Corpus e ensino de espanhol a brasileiros: Descrição de padrões e preparação de atividades didáticas*. São Paulo: PUC/SP, 2001.
- KENNEDY, G.; WOOTEN, C. *The orator in action and theory in Greece and Rome*. Leiden: Brill, 2001. (225).

- KÜHNER, R. *Ausführliche Grammatik der lateinischen Sprache*. Hannover: Hahnsche Buchhandlung, 1912.
- LADEIRA, A. P. *Processamento de linguagem natural: caracterização da produção científica dos pesquisadores brasileiros*. Belo Horizonte: UFMG, 2010. Tese de doutorado.
- LEUMANN, M.; HOFMANN, J.; SZANTYR, A. *Lateinische Grammatik*. Munich: Beck, 1977.
- MAKREHCHI, M.; KAMEL, M. *Automatic Extraction of Domain-Specific Stopwords from Labeled Documents*. Proceedings of ECIR - 30: Advances in information retrieval, 2008.
- MAY, J.; CRAIG, C. *Brill's companion to Cicero, oratory and rhetoric*. Leiden: Brill, 2002.
- MÜLLER, C. F. W. M. *Tulli Ciceronis scripta quae manserunt omnia*. Leipzig: B.G. Teubner, 1879.
- NISARD, D. *Œuvres complètes de Cicéron, avec la traduction en français*. Paris: Institut de France, 1869.
- NORVIG, P. *On Chomsky and the Two Cultures of Statistical Learning*. [S.l.: s.n.], 2011. <http://norvig.com/chomsky.html>. Acessado em 7 de janeiro de 2012.
- OCH, F. J. *Statistical Machine Translation: Foundations and Recent Advances*. Phuket: Google, 2005.
- PALMER, L. *The Latin Language*. Norman: University of Oklahoma Press, 1988.
- PELLEGRINO, D. *Cicero and Horace Vocabulary Frequency Lists for AP Selections*. Wauconda: Bolchazy-Carducci, 2009.
- PERKINS, J. *Python Text Processing with NLTK Cookbook*. Birmingham: Packt Publishing, 2010.
- PETERSSON, T. *Cicero: A Biography*. Berkeley: Kessinger Publishing, 2005.
- SAPIR, E. *Culture, language and personality*. Berkeley: University of California Press, 1949.
- SARDINHA, T. B. *Comparing corpora with WordSmith Tools: How large must the reference corpus be?* Stroudsburg: WCC, 2000.
- SARDINHA, T. B. *Computador, corpus e concordância no ensino da léxico-gramática de língua estrangeira*. Pelotas: Educat/ALAB, 2000.
- SINCLAIR, J. M. *The automatic analysis of corpora*. Estocolmo: Directions in corpus linguistics, 1991.
- TESITELOVA, M. *Quantitative linguistics*. Amsterda: John Benjamins, 1992.
- WILLIS, D. *The Lexical Syllabus, a new approach to language teaching*. London: Collins, 1990.

APÊNDICE A – Códigos

A.1 Scraper de dados e montador dos corpora

```

1  #!/usr/bin/env python
2  # -*- coding: utf-8 -*-
3  # caio begotti <caio1982@gmail.com>
4  # this is under public domain
5
6  # to add some sleep time between fetches
7  # and not hammer down the server
8  import time
9
10 # to avoid charsetting mess with UTF-8 strings
11 import codecs
12 import string
13
14 # to check if a text was already fetched
15 from os.path import exists
16
17 # document parser
18 from lxml import etree
19
20 # output generation
21 from elementtree.SimpleXMLWriter import XMLWriter
22
23 refs = []
24 urls = []
25
26 # the root directory of all latin texts in PHI
27 base = 'http://latin.packhum.org/'
28
29 # marcus tullius cicero entry in PHI
30 browse = base + 'author/474'
31
32 try:
33     page = etree.parse(browse, etree.HTMLParser(encoding='utf-8'))
34 except Exception, err:
35     print 'Browse Error: ' + str(err)
36
37 # gets the list of texts by cicero currently in PHI
38 matches = page.xpath("//span[@class='unam']/text()")
39

```

```

40 counter = 1
41 for entry in matches:
42     # creates a reference list with download addresses for every text
43     refs.append((base + 'dx/text/474/%s/' % str(counter), entry.lower()))
44     counter += 1
45
46 for param in refs:
47     source = param[0]
48     title = param[1]
49
50     # debug
51     print '[%s]' % title
52
53     filename = title.replace(' ', '_')
54     w = XMLWriter(filename + '.xml', encoding='utf-8')
55     xml = w.start("document")
56
57     # metadata entries of the output files
58     w.element("author", 'marcus tullius cicero')
59     w.element("title", title)
60     w.element("source", base)
61
62     # upon checking it no text in PHI attributed to cicero
63     # has more than 500 pages, so this is a safe download limit
64     for x in range(0, 500):
65         lines = []
66         entry = []
67         section = source + str(x)
68         reference = base + 'loc/474/' + str(x) + '/0'
69
70         # debug
71         print '\t<%s>' % section
72
73         # output filename
74         path = 'ready/' + filename + '-' + str(x) + '.txt'
75
76         if not exists(path):
77             # fetches the current page
78             try:
79                 page = etree.parse(section, etree.HTMLParser(encoding='utf-8'))
80             except Exception, err:
81                 print 'Text Error: ' + str(err)
82
83             # parses the page paragraphs
84             try:
85                 entry = page.xpath("//tr/td[1]//text() | //h3//text()")
86             except Exception, err:
87                 print 'Match Error: ' + str(err)
88                 # a priori this is not needed but it is helpful for debugging
89                 f = codecs.open("log.txt", "a", "utf8")
90                 f.write('\nMatch Error: ' + str(err) + ' [missing] ' + section)
91                 f.close()
92
93     # checks if the end of text has been reached

```

```

94         if 'No text' in entry:
95             print 'EOF: ' + str(x)
96             break
97
98         empty = u'\xa0\xa0'
99         if len(entry) > 0:
100             for e in entry:
101                 if e.startswith(empty):
102                     # apparently PHI texts have double blank spaces indicating new paragraphs
103                     lines.append(''.join(e.replace(empty, '')))
104                 else:
105                     lines.append(''.join(e))
106
107             paragraph = ' '.join(lines)
108             y = codecs.open(path, "w", "utf8")
109             y.write(paragraph)
110             y.write
111         else:
112             # if text has been fetched ok, process it
113             paragraph = codecs.open(path, "r", "utf8")
114             strings = paragraph.read()
115
116             # finally writes the new content to the corpus file
117             w.start("page", id=str(x))
118             w.element("paragraph", strings)
119             w.end("page")
120
121             # give the PHI server some time until the next fetch
122             # time.sleep(5)
123         # generates the output file
124         w.close(xml)

```

A.2 Filtros dos corpora

```

1  #!/usr/bin/env python
2  # -*- coding: utf-8 -*-
3  # caio begotti <caio1982@gmail.com>
4  # this is under public domain
5
6  import codecs
7  import glob
8
9  def normalizer():
10     for loop in glob.glob('ready/*.txt'):
11         text = ''
12         with file(loop, 'r') as content:
13             text = content.read().lower()
14
15         # fixes the linebreaking of corpora
16         text = text.replace("- ", "")
17

```

```

18         # fixes unused letters for their real latin ones
19         text = text.replace("v", "u")
20         text = text.replace("j", "i")
21
22         with file(loop, 'w') as content:
23             content.write(text)
24
25 if __name__ == "__main__":
26     normalizer()

```

A.3 Parser e filtro de prenomes latinos

```

1  #!/usr/bin/env python
2  # -*- coding: utf-8 -*-
3  # caio begotti <caio1982@gmail.com>
4  # this is under public domain
5
6  # reference: https://gist.github.com/2307114
7  # double-check: http://en.wiktionary.org/wiki/Appendix:Roman_praenomina
8
9  import codecs
10 import glob
11 import re
12
13 def parser():
14     regex = re.compile("(?:A|Ap|D|C|Cn|K|L|M|Mam|N|O|P|Q|Qu|S|Sp|Ser|Sex|Sec|Seq|Sept|T|Ti|Tit|Vel|Vo) '?\.[A-Z]{0,}\w{0,}"
15     praeenomina = []
16
17     for loop in glob.glob('ready/*.txt'):
18         with file(loop, 'r') as content:
19             text = content.read()
20             for entry in regex.findall(text):
21                 praeenomina.append(entry)
22
23     return sorted(set(praeenomina))
24
25 def replacer():
26     list = parser()
27     regex = re.compile("^(.*)\.[ ]")
28     for loop in glob.glob('ready/*.txt'):
29         with file(loop, 'r') as content:
30             text = content.read()
31             replaced = ''
32             for entry in list:
33                 r = regex.search(entry)
34                 match = r.group(1)
35                 name = re.sub('^\.' + match, '(' + match + ')', entry)
36                 name = name.replace('.', '')
37                 replaced = re.sub(entry, name, text)
38             with file(loop, 'w') as content:
39                 content.write(replaced)

```

```

40
41 if __name__ == "__main__":
42     replacer()

```

A.4 Analisador de frequências

TODO

A.5 Gerador de concordâncias

TODO

A.6 Protótipo de algoritmo de stemmer

```

1  #!/usr/bin/env python
2  # -*- coding: utf-8 -*-
3  # caio begotti <caio1982@gmail.com>
4  # this is under public domain
5
6  # reference: the schinke latin stemming algorithm in python
7  # http://snowball.tartarus.org/otherapps/schinke/intro.html
8
9  import sys
10
11  que = ['atque', 'quoque', 'neque', 'itaque', 'absque', 'apsque', 'abusque',
12        'adaeque', 'adusque', 'deniquep', 'deque', 'susque', 'oblique', 'peraeque',
13        'plenisque', 'quandoque', 'quisque', 'quaequep', 'cuiusque', 'cuique',
14        'quemque', 'quamque', 'quaque', 'quique', 'quorumque', 'quarumque',
15        'quibusque', 'quosque', 'quasque', 'quotusquisque', 'quousque', 'ubique',
16        'undique', 'usque', 'uterque', 'utique', 'utroque', 'utribique', 'torque',
17        'coque', 'concoque', 'contorque', 'detorque', 'decoque', 'excoque',
18        'extorque', 'obtorque', 'optorque', 'retorque', 'recoque', 'attorque',
19        'incoque', 'intorque', 'praetorque']
20
21  noun_suffix = ['ibus', 'ius', 'ae', 'am', 'as', 'em', 'es', 'ia', 'is',
22               'nt', 'os', 'ud', 'um', 'us', 'a', 'e', 'i', 'o', 'u']
23
24  verb_suffix = ['iuntur', 'beris', 'erunt', 'untur', 'iunt', 'mini', 'ntur',
25               'stis', 'bor', 'ero', 'mur', 'mus', 'ris', 'sti', 'tis', 'tur', 'unt',
26               'bo', 'ns', 'nt', 'ri', 'm', 'r', 's', 't']
27
28  nouns = []
29  verbs = []
30
31  # http://stackoverflow.com/questions/3411006/fastest-implementation-to-do-multiple-string-substitutions-in-python
32  # this is the multiple replacing algorithm proposed by matt anderson at stackoverflow in 2010

```

```

33 # it should perform faster than python's native replace method on huge corpora
34 def multi_replace(pairs, text):
35     stack = list(pairs)
36     stack.reverse()
37     def replace(stack, parts):
38         if not stack:
39             return parts
40         stack = list(stack)
41         from_, to = stack.pop()
42         # debug
43         # print 'split (%r=>%r)' % (from_, to), parts
44         split_parts = [replace(stack, part.split(from_)) for part in parts]
45         parts = [to.join(split_subparts) for split_subparts in split_parts]
46         # debug
47         # print 'join (%r=>%r)' % (from_, to), parts
48         return parts
49     return replace(stack, [text])[0]
50
51 def stemmer():
52     for entry in sys.stdin.readlines():
53         # step 2
54         entry = multi_replace([( 'j', 'i'), ('v', 'u')], entry.replace('\n', ''))
55
56         # step 3
57         if entry not in que:
58             if entry.endswith('que'):
59                 entry = entry[:-3]
60             else:
61                 nouns.append(entry)
62                 verbs.append(entry)
63
64         # step 4
65         for s in noun_suffix:
66             if entry.endswith(s):
67                 entry = entry[:-len(s)]
68
69         # step 5
70         if len(entry) >= 2:
71             nouns.append(entry)
72
73         # step 6
74         i = ['iuntur', 'erunt', 'untur', 'iunt', 'unt']
75         bi = ['beris', 'bor', 'bo']
76         eri = ['ero']
77         for s in verb_suffix:
78             if entry.endswith(s):
79                 if s in i:
80                     entry.replace(s, 'i')
81                 elif s in bi:
82                     entry.replace(s, 'bi')
83                 elif s in eri:
84                     entry.replace(s, 'eri')
85                 else:
86                     entry = entry[:-len(s)]

```

```
87
88     # step 7
89     if len(entry) >= 2:
90         verbs.append(entry)
91
92     print zip(verbs, nouns)
93
94 if __name__ == "__main__":
95     # step 1
96     stemmer()
```

APÊNDICE B – Cronograma

As tarefas necessárias para a conclusão do trabalho devem ser divididas em etapas sequenciais, visto que uma depende da outra. Muito provavelmente as etapas serão repassadas após uma primeira “rodada” para afinar os resultados e aparar problemas encontrados durante o caminho, antes da conclusão total da monografia.

As etapas e seus tempos de conclusão serão divididas como a seguir:

B.1 Definição

Definir o escopo do projeto e o que se pretende fazer. Especificar limites para não perder o foco em objetivos nem se perder em superficialidade. Este projeto de monografia propriamente dito. Tempo estimado: Set-Dez 2011. Complexidade: média. Concluída.

B.2 Coleta

Agregação do material para montagem dos *corpora*, pensando já na relevância dos textos, utilidade deles e abrangência linguística. Será preciso orientação próxima para que o material coletado seja relevante e representativo. Tempo estimado: Jan-Mar 2012. Complexidade: média. Concluída.

B.3 Montagem

A construção efetiva dos *corpora* de acordo com os formatos esperados pelo projeto NLTK e uma categorização básica dos textos para facilitar análise futura. Nessa etapa serão disponibilizados os *corpora* do trabalho para que outros do setor e departamento também possam utilizá-lo. Tempo estimado: Abr-Mai 2011. Complexidade: média. Em curso.

B.4 Filtragem

Especificação de certos filtros a serem utilizados em análises dos *corpora* (como a necessidade de se catalogar *stopwords* da língua, por exemplo). Essa etapa também precisará de orientação próxima, pois um conhecimento profundo do latim será necessário para que os dados não fiquem poluídos. Tempo estimado: Mai 2012. Complexidade: baixa. Em curso.

B.5 Análise

Etapa final da análise de todo o material organizado, bem como obtenção de dados estatísticos da língua para uso futuro em sala de aula ou pesquisa acadêmica. Gerar um modelo de uso do material resultante desse trabalho. Tempo estimado: Mai-Jun 2012. Complexidade: alta. Em curso.

APÊNDICE C – Anotações

- Devo arrumar ortografias de 'v' para 'u' e de 'j' para 'i'? Importante pras análises estatísticas!
- Melhor limpar ou considerar macron? Acho que limpar já que não tá nos originais?
- Seria bom botar o verbo ser todo declinado nas stopwords?
- A lista de stopwords tá exagerada? (tem que testar o modelo, deve estar sim, chodchod é interessante por exemplo)
- Segundo o Leumann quicum e quocum são comuns no Cícero, mas eles não são advérbios na realidade?