Introdução aos textos clássicos na era digital do terceiro milênio

letraria

Introdução aos textos clássicos na era digital do terceiro milênio

ANISE D'ORANGE FERREIRA (ORG.)

INTRODUÇÃO AOS TEXTOS CLÁSSICOS NA ERA DIGITAL DO TERCEIRO MILÊNIO

1ª edição

Araraquara
LETRARIA
2015

PROJETO EDITORIAL
Letraria

CAPA
Anise D'Orange Ferreira

REVISÃO
Anise D'Orange Ferreira

ORGANIZAÇÃO
Anise D'Orange Ferreira

TEXTOS ORIGINAIS
Corpus Linguistics, Treebanks and the Reinvention of Philology / Building a Dynamic Lexicon from a Digital Library / Fragmentary Texts and Digital Libraries / Digital Humanities in the Classroom – Technical Approach to Platform Integration / A Nearest-Neighbor Approach to the Automatic Analysis of Ancient Greek Morphology / Social Networks and the Language of Greek Tragedy

AUTORES
David Bamman / Gregory Crane / Monica Berti / Bridget Almas / Marie Claire Beaulieu / John Lee / Jeff Rydberg-Cox

TRADUTORES
Ana Luiza Iaria / Caio Vieira dos Reis Camargo

Ferreira, Anise D'Orange (Org.)
ISBN: 978-85-69395-01-0
Introdução aos textos clássicos na era digital do terceiro milênio / Anise D'Orange Ferreira (Org.) – Araraquara: Letraria, 2015.
144p. 21 x 29,7 cm.

1. Textos clássicos. 2. Era digital. 3. Terceiro milênio.

SUMÁRIO

INTRODUÇÃO

Os tradicionais estudos clássicos têm uma história de apropriação da tecnologia digital, no que se refere à elaboração de *corpora* eletrônicos. No Brasil, desde os anos 90 (FERREIRA; PIQUÉ, 1997[1]), nos acostumamos a trabalhar com o famoso *corpus* eletrônico do TLG (*Thesaurus Linguae Graecae*), de acesso restrito, com sede na Universidade da Califórnia em Irvine, cujo projeto teve início nos anos 70. Da mesma forma, nos familiarizamos com o não menos conhecido Projeto Perseu. Iniciado em mídias digitais, como vídeo disco e CD-ROM, nos anos 80, organizou e mantém incrementando a biblioteca digital *Perseus Digital Library* (PDL), com um acervo de textos e ferramentas de acesso irrestrito e distribuído de forma aberta, *open source*, em *web*, com sede na Universidade Tufts, em Boston, nos EUA.

Tais empreendimentos não se fizeram sem grandes investimentos e esforços conjuntos. De início, o TLG foi financiado principalmente por três grandes fundações, a *National Endowment for the Humanities*, a *Andrew W. Mellon Foundation*, *the David and Lucile Packard Foundation*, sem contar as doações particulares e outras fundações. Recebe também pelas licenças de uso. O Projeto Perseu já recebeu apoio de várias fontes e fundações, entre elas: o *Annenberg/CPB Project*, a *Apple Computer*, o *Getty Grant program*, a *Modern Language Association*, a *National Endowment for the Arts*, o *Packard Humanities Institute*, a *Xerox Corporation*, a *Boston University* e a *Harvard University*. Atualmente, continua recebendo apoio de fontes diversas e fundações: *Alpheios Project*, a *Andrew W. Mellon Foundation*, o *Institute of Museum and Library Services*, a *National Endowment for the Humanities*, a *National Science Foundation*, doações particulares, e a própria *Tufts University*. Vincula-se, por meio do seu editor-chefe, Gregory Crane, aos projetos *Open Philology*, beneficiados pelo prêmio *Alexander von Humboldt Professorship*, sendo desenvolvidos no Departamento de *Digital Humanities* do Instituto de Computação da Universidade de Leipzig.

[1] FERREIRA, A. A. G. D.; PIQUÉ, J. F. Tecnologia de Informação e Letras Clássicas. Minicurso na X Reunião de Estudos Clássicos. 1997.

O avanço tecnológico voltado para os textos clássicos não parou desde então, acompanhando a tecnologia corrente. Do sistema original do TLG, Ibycus (cf. Hegelrso[2]), e armazenamento em mídias como disquetes e CDs, chegamos às bases de dados em rede pela Internet, com conectividade pelos aparelhos móveis, como telefones e *tablets*. Na primeira década do terceiro milênio iniciamos uma fase de produção de textos e serviços digitais em *web* (e em outros aplicativos) que devem mudar a visão e a forma de se estudar, pesquisar e de produzir publicações na área de estudos clássicos.

Por essa razão, tenho a satisfação de trazer aos leitores uma seleção de textos básicos em língua portuguesa sobre esse desenvolvimento das tecnologias produzidas, principalmente pela equipe e colaboradores de Gregory Crane, que é o responsável não só pela existência da Biblioteca Digital do Projeto Perseu, mas também por exercer uma liderança na inovação tecnológica dentro das humanidades digitais ligadas às línguas históricas.

Em 1988, Crane completou a primeira versão do *software* Morpheus 1.0, um *parser* ou analisador automático da morfologia do grego antigo, iniciado em Harvard em 1984. Esse analisador é capaz de reconhecer diferenças sutis da morfologia grega, como diferenças dialetais. A análise das palavras consiste no pareamento dos radicais e flexões, comparando-os com conhecimento padrão da morfologia grega. O Morpheus está integrado à Biblioteca Digital Perseu e a muitas outras ferramentas desenvolvidas por terceiros, em programas *stand-alone* (ex. Diogenes, Kalos, Sibylla), ou *on-line* (*Perseus at Chicago*, Logeion, Attika, etc). Até o atual serviço *web* do TLG, restrito a assinantes, se beneficia dessa mesma ferramenta para seus mecanismos de busca de palavras no *corpus*.

Em um movimento na direção ao classicismo digital, os processos têm evoluído para incorporar tecnologias mais recentes utilizadas nos estudos de linguística de corpus, processamento de linguagem natural e/ou linguística computacional. Nesse movimento, a meta de Crane e de outros adeptos é criar uma ciberinfraestrutura apropriada para disseminar o acesso aos textos clássicos em formato digital, não só para fins de leitura, mas também para atender a inúmeras e diversas necessidades de pesquisa na área, documentando as línguas clássicas e provendo serviços que facilitam o acesso a essas

[2] CD ROM and Scholarly Research in the Humanities, *Computer and the Humanities*, 22 (1988) 111-116.

línguas. Tal ciberestrutura é promissora em suprir várias necessidades acadêmicas, entre elas, ampliar o desenvolvimento das pesquisas na área de letras clássicas e demais línguas históricas. Nesse sentido, seus recursos visam a editar, estudar e rever uma grande quantidade de documentos, fontes primárias e *corpora* existentes, associando-os a fontes secundárias e a serviços multilíngues; a tornar públicos e disponíveis documentos para estudantes, pesquisadores e público em geral; a adotar padrões editoriais e práticas acadêmicas de alto nível, respeitadas na área. Dentre as necessidades pedagógicas, promete propiciar uma experiência de aprendizagem da língua que contribua para a produção acadêmica começando na graduação; uma prática produtiva em um ambiente colaborativo, em interação genuína com pesquisadores; e, por consequência, diversificar as oportunidades profissionais.

A posição de quem segue essa linha de trabalho talvez não esteja ainda bem clara para muitos dos nossos filólogos que acham que edições digitais são livros em formato PDF[3]. Sim, PDF é um formato de arquivo e é digital, todavia, a não ser pelo fato de ser um texto digital, não significa que se insira na concepção de textos digitais típicas do terceiro milênio. Além de estar em formato digital, a proposta atual é incrementar uma base universal de dados abertos, sem restrições de uso. Sim, novamente, há muitos livros em PDF que podem ser baixados gratuitamente pela Internet, sem restrições. Ainda, não é disso que se trata. O livro em PDF, ou em outros formatos fechados, é inerentemente um produto individual e isolado de outros dados. Ainda é a reprodução de um livro na prateleira da biblioteca à espera de consulta, mesmo que possa ser baixado, através do computador, até de forma gratuita. Ele contém informações que não fazem interlocução com dados de outros textos, a não ser pelas observações encontradas pelo autor, documentadas em notas de rodapé, que ficam estacionadas ali. Para que a comunicação entre dados possa ocorrer, as novas edições de textos e estudos devem ser elaboradas, com anotações de dados, em formatos universais e abertos, atualmente, em padrão TEI-XML, de modo a poderem ser distribuídas em diferentes serviços de "mineração de dados", tradução literal de *data mining*.

A grande diferença que se instaura, portanto, no estudo e produção de conhecimento, é que os estudiosos ao fazerem sua pesquisa estão produzindo dados não como unidades isoladas e estáticas, mas como uma contribuição para um sistema mais

[3] *Portable Document Format.*

amplo e dinâmico. O pesquisador precisará decidir se vai gerar um estudo monolítico que será lido por poucos, ou se vai produzir um estudo dinâmico, i.e., uma edição integrada em sistemas de conhecimento abertos, sem, contudo, desprezar o rigor e os procedimentos filológicos.

A questão que surge em seguida diz respeito às competências exigidas do filólogo ou classicista para aderir a essa prática. Pois não adianta esperar que outros façam esse serviço em seu lugar, na função de uma espécie de bibliotecários digitais, que organizam os livros numa biblioteca digital. Aliás, esses já têm um papel específico na catalogação e na padronização de identificadores dos elementos no mundo digitalizado. Os classicistas vão precisar aprender a fazer sua parte dentro desse mundo. A boa notícia é que as interfaces de usuário têm ficado cada vez mais acessíveis para facilitar as edições e exigem cada vez menos conhecimentos operacionais no computador por parte do pesquisador da área de Humanas. De fato, os conhecimentos exigidos da parte dos cientistas da computação para dar suporte aos trabalhos dos classicistas não é nada fácil para quem não recebeu uma formação técnica específica; requer conhecimentos avançados de programação. Nessa situação, visando ao avanço dos recursos, os classicistas trabalham conjuntamente com os cientistas da computação, lhes indicando os parâmetros e aprendendo a inserir os procedimentos filológicos nos aparatos tecnológicos. Decorre dessa necessidade de se trabalhar em equipe que a tradição de estudos autorais pode dar lugar a estudos coletivos, com créditos distribuídos entre membros da equipe.

Os passos do desdobramento da ciberinfraestrutura, que oferece recursos *open source* pela *web*, foram documentados no artigo-manifesto: "Cyberinfrastructure for Classical Philology"[4] e também na obra de Babeu (2011), *Roma não foi digitalizada em um dia*[5]. Todas as fases subsequentes desse desenvolvimento com discussões sobre o impacto da área vêm sendo continuadamente divulgadas em artigos do *blog Perseus Digital Library Updates*, cuja lista de tópicos aparece na *home page* da *Perseus Digital Library*. Os processos nesta área são tão rápidos que as publicações tradicionais não conseguem acompanhar. O desenho atual do modelo de onde se pretende chegar com a

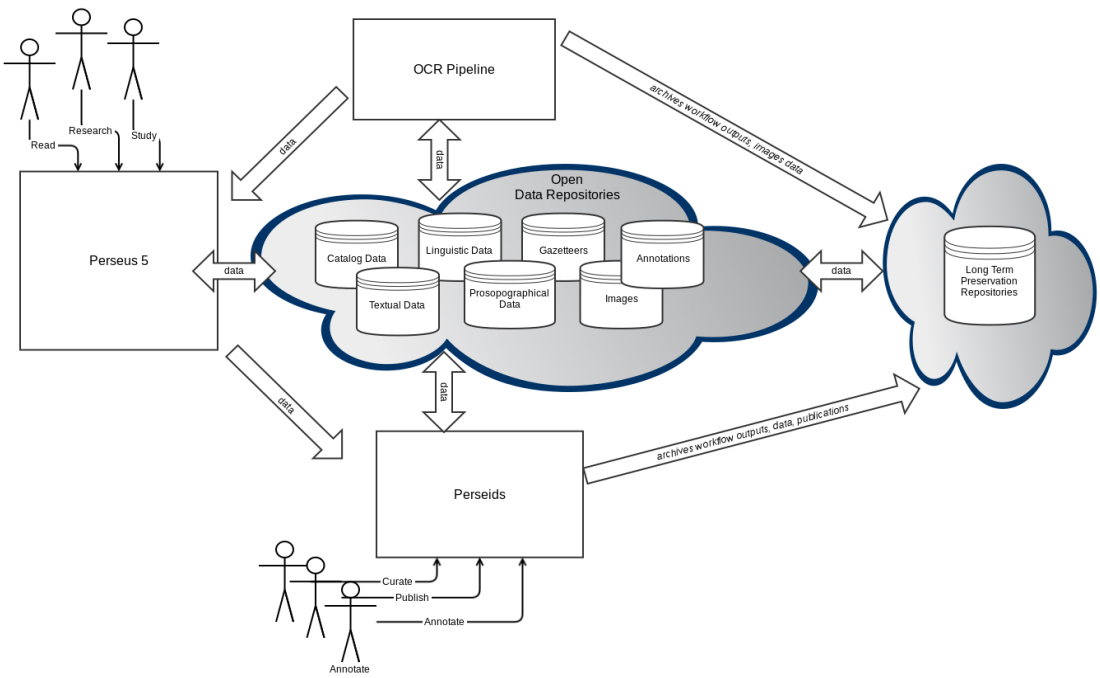
[4] "Cyberinfrastructure for Classical Philology". In Crane, G. and Terras, M. (eds) (2009). "Changing the Center of Gravity: Transforming Classical Studies Through Cyberinfrastructure". Digital Humanities Quarterly, Volume 3 Number 1, Winter 2009. Disponível on-line: http://www.digitalhumanities.org/dhq/vol/003/1/000023/000023.html

[5] Em inglês, *Rome wasn't digitized in a day: building a cyberinfrastructure for digital classicists.* CLIR n. 150. Washington: Council on Library and Information Resources. Disponível on-line: http://www.clir.org/pubs/abstract/pub150abst.html

infraestrutura está no texto "The road to Perseus 5 – why we need infrastructure for the digital humanities" escrito por Bridget Almas[6].

A figura do modelo apresentada nessa publicação mostra como se configura a proposta atual da referida infraestrutura. Na Figura 1, constatamos dois pontos principais em que os usuários interagem com a tecnologia: um como estudante, leitor ou pesquisador da *Perseus 5*, e o outro, como aquele que contribui para editar os textos e alimentar a base de dados, sendo curador, anotador, ou editor. No meio de tudo, temos o processamento do material e os vários tipos de repositórios, com diferentes tipos de dados, e a integração entre eles.

Figura 1 – Visualização da solução da ciberinfraestrutura da Perseus D.L. 5



Fonte: B. Almas, Perseus Updates, 18/05/2015

O desenvolvimento de projetos no chamado âmbito do classicismo digital é novo em nosso hemisfério lusófono, onde os recursos chegam a ser notícia para o usuário local quando relativamente prontos, sem que haja a participação ou contribuição de nossos

[6] http://goo.gl/IOHgje (URL reduzido) Perseus Updates 2015/05/18.

estudantes e pesquisadores em seu aperfeiçoamento. Estamos acostumados a receber comodamente o artefato tecnológico praticamente pronto para usar. E se há falhas, o artefato é desprezado como recurso suficientemente rigoroso para sustentar alguma tese. Contudo, a mim rompeu-se esse distanciamento quando, em 2012, tive a oportunidade de participar do NEH *Institute for Advanced Technology in the Digital Humanities*, intitulado "*Working with Text in a Digital Age*", que teve lugar na Tufts University, dentro do Projeto Perseu. Durante esse *workshop*, em que foram selecionados em torno de 24 participantes, colaboradores de Gregory Crane apresentaram-se com alguns de seus respectivos projetos:

- Anke Lüdeling, da Universidade Humboldt, Alemanha, e uma das organizadoras do evento, trabalhou com os convidados seu método de anotação manual de *corpus* e o emprego do serviço *web* ANNIS desenvolvido por sua equipe e usado em pesquisa de Linguística de Corpus.[7] O Projeto Perseu adaptou a ferramenta para pesquisa no *corpus* de grego e latim, anotado morfológica e sintaticamente, criando o serviço "Perseus Latin and Ancient Greek Treebank - Annis Query Tool"[8]. O serviço contém alguns *corpora* anotados apenas, como exemplos de uso da ferramenta no âmbito do grego e latim.

Para aqueles que contam com alguém que consiga instalar ANNIS em seu servidor *web*, e que seja capaz de acrescentar no sistema seus próprios *corpora* anotados, a ferramenta é gratuita, aberta e pode atender a projetos de pesquisa específicos. Um dos obstáculos enfrentados por docentes e pesquisadores na área de Humanas em instituições de ensino superior públicas no Brasil, no entanto, é que se parte do pressuposto de que humanistas não precisam de especialistas em computação para lhes dar suporte. Profissionais com esse perfil não são contratados, já que não são formados nas áreas de concentração de Humanas. A ajuda de tal recurso humano depende de bolsas que financiem a contratação temporária de serviços de terceiros ou de estagiários, o que dificulta o avanço da pesquisa na área.

[7] http://korpling.german.hu-berlin.de/ridges/index_de.html
https://annis2.sfb632.uni-potsdam.de/Annis/
[8] http://annis.perseus.tufts.edu/

- Monica Berti apresentou o projeto de identificação digital dos vários tipos de reuso de textos clássicos, "Detecting text re-use", com participação a distância de Marco Büchler do projeto E-traces[9]. Atualmente, em conjunto com Bridget Almas[10], vários de seus projetos são conduzidos pela plataforma *Perseids*, que reúne uma série de ferramentas para anotação manual de referências intertextuais. Desde 2013, ela é membro da cadeira de Humanidades Digitais em Leipzig, chefiada por Gregory Crane. Atualmente, ela também coordena o Consórcio Sunoikisis Digital Classics (DC), do qual fazemos parte, vinculado ao programa Sunoikisis tradicional do Centro de Estudos Helênicos de Harvard, coordenado por Kenny Morrell, e sediado em Leipzig, com apoio da Tufts. Reúne participantes de países europeus, Estados Unidos e Brasil, representado, por enquanto, apenas pelo nosso grupo, de docentes e discentes, na FCL-Ar/UNESP.

- Amir Zeldes, da Universidade de Humboldt, na época, colega de Anke Ludeling, trabalhou com os convidados a estatística para linguística de corpus, usando o *software* R. Ele é um dos desenvolvedores do ANNIS2, a ferramenta baseada em *web* para pesquisa e visualização de *corpora* contendo anotação de múltiplas camadas, i.e., morfológicas, sintáticas, etc. O *software* R está listado dentre os poderosos instrumentos de código aberto para análise de dados trabalhados na área de Humanas. Realiza análises estatísticas, gera gráficos com diferentes e riquíssimas visualizações dos dados. O lado desse *software* menos atraente ao humanista é que sua interface não é gráfica como uma planilha de Excel. Os dados podem ser importados de uma planilha, mas os comandos operacionais são efetuados por uma janela de terminal dentro da interface, para que o *software* execute a ação desejada. Um dos exemplos do uso dessa ferramenta na análise do texto clássico está em trabalhos de Jeff-Rydberg Cox, autor de um dos artigos aqui traduzidos.

- Bridget Almas, da equipe de Harry Diakoff e G. Crane, no Projeto Alpheios e Perseus Project, Tufts, apresentou e trabalhou conosco o procedimento de edição digital de traduções paralelas e *treebank* em XML e XHTML utilizando a ferramenta chamada oXygen. Ela é a chefe de programação dos desenvolvedores das interfaces e da integração de dados na plataforma criada em seguida ao evento, *Perseids*, que integrou os editores

[9] E-traces - http://etraces.e-humanities.net/publications-etraces.html
Fragmentary texts - http://www.fragmentarytexts.org/
http://demo.fragmentarytexts.org/

[10] Almas & Berti, "Perseids Collaborative Platform for Annotating Text Re-Uses of Fragmentary Authors", DH-case '13, September 10 2013, Florence, Italy, http://dx.doi.org/10.1145/2517978.2517986

de alinhamento de tradução e de edição de *treebank* da ferramenta Alpheios. O editor de *treebank* foi aprimorado na interface de outra ferramenta, Arethusa, dentro da mesma plataforma.

- J. Matthew Harrington, da equipe do Projeto Perseu, nos apresentou o projeto do *treebank* como processo pedagógico: "Treebank as pedagogy: the role of syntatic control in language aquisition".[11] O *Ancient Greek Dependency Treebank* 1.0 foi elaborado por David Bamman e Gregory Crane, para anotação sintática dos textos gregos e latinos. O projeto AGDT atualmente é desenvolvido por Giuseppe Celano, que elaborou uma complementação de anotação, adicionando elementos semânticos, de acordo com a gramática grega tradicional em língua inglesa, de Herbert W. Smyth. Como todo sistema de anotação, ele é acompanhado de um manual ou guia. Tanto AGDT 1.0 quanto sua versão seguinte foram traduzidas para o português para uso de nossos discentes e docentes. Elas serão alvo de publicações independentes, já que passam por modificações e ajustes.

- Neel Smith da *Holy Cross Classics* nos apresentou o projeto *Homer Multitext*, com apoio do *Center of Hellenic Studies*, voltado para a formação de edições digitais indexadas e interligadas de todos os manuscritos de Homero disponíveis[12] no acervo de imagens digitalizadas do CHS, para que os estudantes e pesquisadores possam rever, futuramente, os estudos relacionados à obra.

- Marie Claire Beaulieu e Bridget Almas, membros da equipe do Projeto Perseu, apresentaram o projeto que reúne, numa plataforma, várias ferramentas digitais do Projeto Perseu e Alpheios para edição e tradução de novos manuscritos: *Digital Humanities in the Classroom Introducing a New Editing Platform for Source Documents in Classics*[13]. Tal plataforma foi denominada *Perseids*[14]. Esta plataforma está sendo empregada pelo projeto de uma das participantes do evento, Michèle Brunet, da Universidade de Lyon, para a edição digital e tradução das epigrafias gregas do Museu do Louvre, dentro do seu projeto

[11] Greek and Latin Dependency Treebank Editor: http://nlp.perseus.tufts.edu/syntax/treebank/
[12] http://katoptron.holycross.edu/~nsmith/greek/index.html, http://www.homermultitext.org/
[13] http://perseids.org/, http://shell.perseus.tufts.edu/sosol/
[14] O desenvolvimento dessa plataforma de edição representa um importante papel na obtenção de grandes *corpora* anotados dentro do projeto *Open Philology*. Ver adiante.

chamado E-PIGRAMME.[15]

- Helma Dik, professora da Universidade de Chicago, responsável pelo Perseus under Philologic, que tem outra interface de pesquisa do acervo Biblioteca Digital do Perseu, e também pelo serviço *web* Logeion, apresentou os fundamentos e finalidades da mineração de dados e os novos paradigmas da pesquisa em Estudos Clássicos. Os classicistas brasileiros tiveram a chance de assistir a uma conferência de Helma Dik, no evento da Sociedade Brasileira de Estudos Clássicos em Brasília, em 2013.

- John Lee, que trabalha com Linguística Aplicada e tecnologias para língua grega e chinês na Universidade da Cidade de Hong Kong, apresentou os recursos em linguística computacional desenvolvidos por ele para o ensino de língua grega e chinês. Ele fez pesquisa sobre a ampliação dos métodos do analisador morfológico automático da língua grega, a partir do Morpheus. Ele também veio a supervisionar o estágio PDSE de um dos nossos orientandos, Caio Camargo em 2013-14.

- Jeff Rydberg-Cox, professor da Universidade de Missouri, Kansas City, apresentou um projeto em pedagogia Grega, utilizando "blended learning" ou sistema híbrido *on-line* com interfaces para *tablets* e *smartphones*. Ele tem outros projetos interessantes como Léxico Grego-Inglês e *Visual Explorer For The Language of Greek Tragedy*[16].

- Bruce Robertson, do Departamento de Clássicas da universidade canadense Mount Allison, apresentou, via Skype, o desenvolvimento do seu OCR para grego politônico[17]. Recentemente, publicou, via lista de correspondência eletrônica, o atual estágio de sua ferramenta[18], que exige um servidor muito potente para processamento das rotinas necessárias para um reconhecimento excelente de edições muito antigas, com anotações e mistura de grego e latim. Esse projeto tem o apoio do Google[19]. Um exemplo do desenvolvimento atual do seu OCR pode ser conferido *on-line* na página do

[15] https://univ-lyon2.academia.edu/MicheleBRUNET/ANR-E-PIGRAMME
[16] http://daedalus.umkc.edu/?page_id=16
[17] http://www.perseus.tufts.edu/publications/dve/RobertsonGreekOCR/
[18] http://heml.mta.ca/lace/catalog
[19] http://www.google.com/googlebooks/ancient-greek-and-latin.html

Comicorum atticorum fragmenta 3, da edição de Kock do século XIX[20]. Seu projeto tem código aberto e está disponível sob licença aberta – *open source* – no servidor github[21] para os desenvolvedores.

- David Mimno, professor do Departamento de Ciência da Informação na Cornell University, apresentou-nos o Projeto Mallet: MAchine Learning for LanguagE Toolkit, mostrando procedimentos de aprendizagem de máquina para mineração de dados (*data-mining*) de textos na área de estudos clássicos. Ele também fez parte da equipe do Projeto Perseu. [22]

A partir de 2013, Gregory Crane assume a cadeira Humboldt de Humanidades Digitais na Universidade de Leipzig, um prêmio recebido da Fundação Humboldt para financiar a revisão e a aceleração dos projetos na área, durante cinco anos, sob o título guarda-chuva *Open Philology Project.* Nesse projeto, Crane apresenta três linhas de ação interligadas: produzir dados filológicos, ampliar ao público a educação das línguas históricas e obter uma integração das diversas fontes de dados filológicos. Essas três linhas se resumem em três projetos. Um **Projeto de Grego e Latim Aberto**, que organiza o conteúdo das línguas históricas e suas traduções em línguas modernas; um **Projeto de Aprendizagem on-line de Língua Histórica**, que investiga formas de dar apoio ao ensino; e o **Projeto da Biblioteca Digital Scaife** que faz a integração dos materiais de herança cultural sob licenças abertas.

Durante o início desse professorado, as novas implementações, como as da plataforma *Perseids*, foram aprimoradas, assim como o *The Perseus Catalog* além da integração desses com outros serviços complementares desenvolvidos por terceiros. O catálogo é um serviço *web* de identificação de passagens de obras, de obras, de edições e de traduções. É bom lembrar que uma mesma obra pode ter diversas edições, daí não se misturar o identificador da obra com o texto em si. Dessa forma, todo o acervo digitalizado pode ser adequadamente identificado e relacionado em outros serviços. Muitos devem ter reparado que, atualmente, a Biblioteca Digital Perseu (PDL) exibe um bloco chamado "Stable Identifiers". Ali, temos: URI da citação, URI do texto, URI da

[20] http://heml.mta.ca/lace/render_page?hocrtype_id=4804
[21] https://github.com/brobertson/rigaudon
[22] http://people.cs.umass.edu/~mimno/publications.html

obra e URI do catálogo, onde URI significa, em português, **Nome Uniforme de Recursos.** Por exemplo, o URI da "A República" de Platão é na PDL http://data.perseus.org/catalog/urn:cts:greekLit:tlg0059.tlg030.perseus-grc1. O catálogo Perseu fornece, além de todos os nomes que Platão recebe em várias línguas, o seu URN: urn:cts:greekLit:tlg0059.tlg030.perseus-grc1, onde URN significa, em português, **Nome Uniforme de Recursos.**

Durante os anos 2013 e 2014, com apoio recebido do CNPq, conduzimos, na UNESP, com colegas da área de grego da FCLAr[23], colaboradora da USP[24] e com suporte dos recursos desenvolvidos pela equipe e colaboradores de Crane, uma pesquisa coletiva de produção de alinhamento de tradução para o português de textos gregos utilizando o editor de alinhamento da plataforma Alpheios em sala de aula e em projetos individuais, sobre *corpora* previstos na grade curricular e no próprio *site* do Alpheios. Traduzimos as interfaces dos editores de alinhamento e de *treebank* na página do Alpheios. Essa foi a forma inicial que encontramos para nos beneficiar dos recursos digitais e, ao mesmo tempo, contribuir com dados nossos e de estudantes. Do ponto de vista da pesquisa, verificamos e analisamos as dificuldades e objetivos do alinhamento, e começamos a estabelecer critérios de alinhamento de tradução em português para se compartilhar um padrão. Para o professor, tornou-se um instrumento de materialização e visualização da compreensão, ou a falta dela, sobre o funcionamento da língua grega ou de aspectos gramaticais por parte dos alunos. No projeto maior do professorado em Leipzig, estão se criando formas automáticas de comparação entre diferentes anotadores/tradutores.

A partir de 2015, com a criação do Consórcio Sunoikisis DC, passamos a participar com mais proximidade aos recursos em sua versão mais avançada, dentro da Plataforma Perseids, tanto para anotação de nomes, lugares e eventos e de registros materiais, bem como anotação de *treebank* e tradução alinhada. O Sunoikisis DC tem por objetivo promover cursos com metas de ensino que incluem habilidades operacionais das ferramentas desenvolvidas. Esses cursos, com sessões *on-line* compartilhadas, têm sua contrapartida com sessões tutoriais ou aulas locais, organizadas oficialmente pelas instituições e professores de cada país. O primeiro curso deste consórcio foi realizado no primeiro semestre de 2015, tendo como estudo de caso as publicações e fontes sobre a

[23] Edvanda B. Rosa, Fernando B. dos Santos e M. Celeste C. Dezotti.
[24] Filomena Y. Hirata.

Guerra do Peloponeso, focalizando especialmente a seção Pentekontaetia, da *História da Guerra do Peloponeso*, de Tucídides. Todas as apresentações e materiais estão disponíveis *on-line* a partir das comunidades Google + Sunoikisis DC[25].

Existem muitos outros grupos atuando na área do classicismo digital, seja como desenvolvedores, seja como contribuintes de dados. Todos, porém, se esforçam para conversar na mesma linguagem dos dados abertos. Estão atualmente disponíveis serviços surpreendentes realizados com dados de edições digitais, visualizações de lugares, eventos, nomes, associados ao texto e a anotações sintáticas e morfológicas. Um exemplo é o *Hellespont Project*[26], elaborado e mantido pelo DAI, *Deutsches Archäologisches Institut*. O próprio Centro de Estudos Helênicos de Harvard em Washington oferece um conjunto de projetos sob a rubrica *Digital Humanities Projects at the Center for Hellenic Studies*[27], além do famoso curso *on-line* de Gregory Nagy sobre o herói grego. Mais que abrigar o tradicional Sunoikisis[28] e dar apoio ao projeto *Homer Multitext*[29], há ainda o projeto do *Derveni Papyrus*, dentro do *CHS-iMouseion Project*, *Plato's similes*, etc. O Centro tem-se esforçado na divulgação de popularização dos seus recursos digitais, apresentando-os, inclusive, a classicistas brasileiros, durante o *workshop* organizado por Yannis Petropoulos, no mesmo Centro, em 2014, *Classics in Brazil in the digital age.* Quem quiser se familiarizar com os vários projetos em andamento na área, pode acompanhar a lista de correspondência, *Digital Classicist*, provida pelo servidor inglês que apoia a comunidade DigitalClassicist.org[30], onde o leitor pode encontrar também uma boa base bibliográfica sobre o assunto.

Os textos aqui apresentados não representam a última novidade em termos de resultados da pesquisa da inovação tecnológica, posto que as mudanças são muito rápidas no que concerne aos avanços da tecnologia. Mas descrevem alguns de seus fundamentos, princípios e procedimentos básicos. A maioria são textos que podem ser encontrados na internet em inglês e todos têm acesso irrestrito. A tradução desses tem por objetivo concentrar o tópico em uma mínima coleção, em língua portuguesa, para atingir o maior número possível de leitores da nossa língua, principalmente de estudantes da área de

[25] https://goo.gl/pfrfhn (URL reduzido)
[26] http://gapvis.hellespont.dainst.org/#index
[27] http://chs.harvard.edu/CHS/article/display/5417?menuId=134
[28] http://wp.chs.harvard.edu/sunoikisis/courses/greek-seminar-and-course-archives/
[29] http://chs.harvard.edu/CHS/article/display/1169
[30] http://www.digitalclassicist.org

Humanas, facilitando o acesso e agilizando a leitura dentro de uma abordagem que não lhes é nem habitual, nem familiar.

O artigo de Bamman e Crane, publicado originalmente em 2010, trata da implementação de um sistema de anotação manual de elementos sintáticos para a língua grega. Outro texto dos mesmos autores diz respeito à implementação de um léxico dinâmico, criado a partir da identificação automática de lemas extraídos de *corpora* paralelos de traduções alinhadas e também de *corpora* etiquetados com análise morfológica (POS) e sintática (*treebank*). O texto de Berti apresenta o modelo para estabelecer as relações entre textos fragmentários e suas fontes em uma edição digital. O artigo de Almas e Beaulieu é uma continuidade dos descritores técnicos da plataforma Perseids. O artigo que dá origem a esse, publicado pela *Language & Literary Computing*, não pode ser reproduzido em português, uma vez que a editora Oxford não permite a publicação de traduções. Convido o leitor que tiver dificuldade em acompanhar esse texto a fazer a leitura de um artigo anterior[31] apresentado em congresso. Outro texto que não pôde ser reproduzido nesta coletânea é o de múltiplos autores, colaboradores de Gregory Crane, que discute sobre estudantes pesquisadores e acadêmicos cidadãos, cujo artigo, traduzido, está disponível em outro lugar[32]. John Lee nos contempla, em seu artigo, com uma técnica, chamada "a abordagem do vizinho mais próximo", que utilizou para incrementar o analisador automático Morpheus, empregado originalmente na PDL, apresentando os desafios da língua grega no que concerne à ambiguidade morfológica. Por fim, o trabalho de Jeff-Rydberg Cox nos apresenta como redes sociais, como a de uma tragédia de Ésquilo, podem ser extraídas automaticamente e configuradas visualmente.

As traduções foram realizadas com permissão dos autores e auxílio do Programa de Pós-Graduação em Linguística e Língua Portuguesa da Faculdade de Ciências e Letras da Universidade Estadual Paulista em Araraquara, São Paulo, Brasil, e minha organização, como parte de apoio ao projeto do MCTI /CNPq /MEC/CAPES n. 18/2012 – Ciências Humanas, Sociais e Sociais Aplicadas, proc. nº 406845/2012-3. Sobre a tradutora de todos os textos, com exceção do artigo de John Lee: Ana Luiza Iaria, bacharel em Letras (português e inglês) e Direito/USP; MSc em Tradução Científica, Médica e

[31] http://goo.gl/63Gq8p (URL reduzido)
[32] https://goo.gl/lxpj92 (URL reduzido).

Técnica pela Imperial College London; Fellow do Chartered Institute of Linguists, Membro do Institute of Translation and Interpreting (UK) e da American Translation Association; Chartered Linguist. A tradução do artigo de John Lee foi realizada por Caio Vieira dos Reis Camargo, doutorando do referido Programa.

Anise D'Orange Ferreira.

Araraquara, 21 de julho de 2015.

LINGUÍSTICA DE CORPUS, *TREEBANKS* E A REINVENÇÃO DA FILOLOGIA[33]

David Bamman[34]
Gregory Crane[35]

Introdução

Há muitos anos, humanistas em geral, e estudantes do mundo greco-romano em particular, trabalham com materiais digitais, mas o mundo digital emergente, nesta primeira geração, até então teve relativamente pouco efeito sobre os objetivos, as práticas e a cultura intelectual geral das ciências humanas. Alunos do passado usaram novas ferramentas para fazer as mesmas perguntas e aprimorar atividades bem estabelecidas – usaram suas enormes coleções como gigantes concordâncias e o e-mail acelerou, ao invés de alterar, o fluxo de publicações eletrônicas. Os *treebanks* gregos e latinos em desenvolvimento pelo Projeto Perseu na Universidade de Tufts começaram a refletir as mudanças mais fundamentais. *Treebanks* são coleções de texto com amplas categorias morfológicas, sintáticas e similares de anotação e são instrumentos familiares para pesquisa de linguística de corpus e computacional. Ao elaborar os *treebanks* para línguas históricas, como o grego e o latim, encontramos um novo espaço intelectual que combina elementos de linguística computacional e de *corpus* e da antiga disciplina de filologia. O artigo abaixo descreve os trabalhos sobre os *treebanks* e, em seguida, descreve as implicações deste trabalho para o latim, o grego e outras línguas históricas.

Análise sintática

O ressurgimento de métodos estatísticos na linguística computacional nos últimos 25 anos deu origem a um grande investimento na criação de *treebanks* – corpora grandes

[33] Publicado originalmente em 2008, com o título *Corpus Linguistics, Treebanks and the Reinvention of Philology* pelo *Perseus Project*, Tufts University. Em português, distribuído sem fins comerciais, com permissão dos autores.

[34] Perseus Project – Department of Classics Tufts University – Medford MA – 02140 – USA – david.bamman@tufts.edu

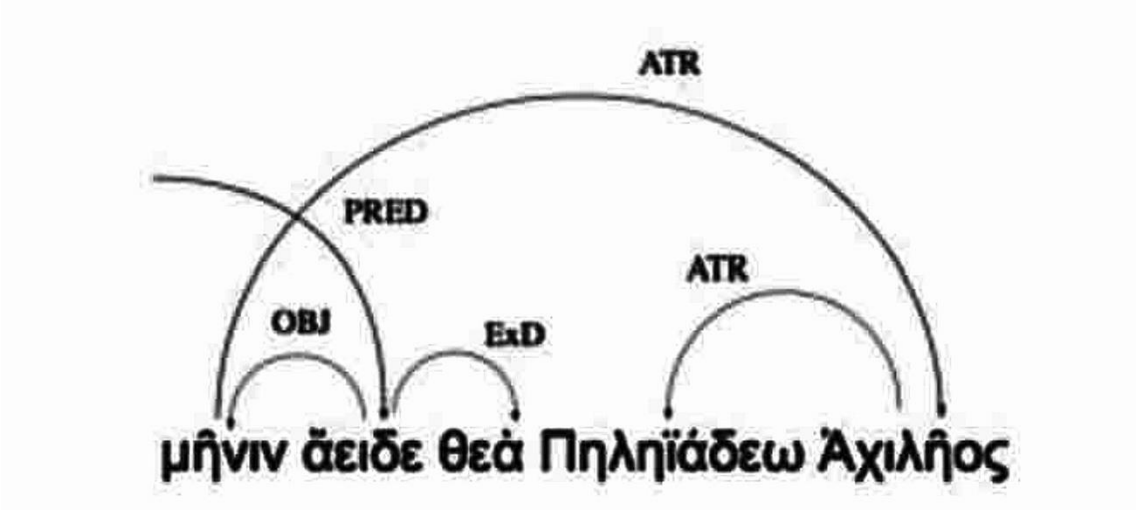
[35] Perseus Project – Department of Classics Tufts University – Medford MA – 02140 – USA – crane@tufts.edu.

e sintaticamente anotados. A maior parte do trabalho teve por foco o inglês (MARCUS *et al.*, 1993) e outras línguas modernas como tcheco (HAJIC, 1998), alemão (BRANTS *et al.*, 2002), espanhol (MORENO *et al.*, 2000), francês (ABEILLE *et al.*, 2000), italiano (MONTEMAGNI *et al.*, 2000) e japonês (KUROHASHI; NAGAO, 1998); recentemente, apareceram vários incluindo outras línguas também outras línguas históricas, como inglês médio (KROCH; TAYLOR, 2000), inglês moderno anterior (KROCH *et al.*, 2004), inglês antigo (TAYLOR *et al.*, 2003b), português medieval (ROCIO *et al.*, 2000), ugarítico (ZEMANEK, 2007) e várias traduções indo-europeias do Novo Testamento (HAUG; JOHNDAL, 2008).

Com o financiamento da National Science Foundation, começamos a desenvolver um *treebank* para o latim clássico em 2006. Os resultados deste trabalho levaram diretamente ao financiamento privado para o desenvolvimento de um *treebank* com 400.000 palavras para poesia grega antiga. Em julho de 2010, lançamos publicamente mais de 280.000 palavras sintaticamente analisadas a partir destas duas línguas (230.953 palavras do grego antigo e 53.143 palavras do latim)[36]. Uma vez que o latim e o grego antigo são línguas com muitas declinações e com um alto grau de variação na ordem das palavras, baseamos nosso estilo de anotação na gramática de dependência do Prague Dependency Treebank (HAJIC 1998) para o tcheco (outra língua não projetiva) que, desde então, tem sido amplamente adotado por vários projetos de anotação para outras línguas, inclusive o árabe (HAJIC *et al.*, 2004), esloveno (DZEROSKI *et al.* 2006) e grego moderno (PROKOPIDIS *et al.*, 2005). A Figura 1 ilustra uma árvore de dependência no antigo *Treebank* de Dependência do Grego Antigo, usando o primeiro verso da *Ilíada* de Homero.

[36] Todos os dados analisados sintaticamente estão disponíveis publicamente em: http://nlp.perseus.tufts.edu/syntax/*treebank*/.

Figura 1 – Árvore de dependência (texto grego)
(μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος - "Canta, deusa, a ira de Aquiles, filho de Peleu"),
Homero, *Ilíada* 1.1. Foram desenhados arcos que vão dos nós aos dependentes.



Infraestrutura da anotação

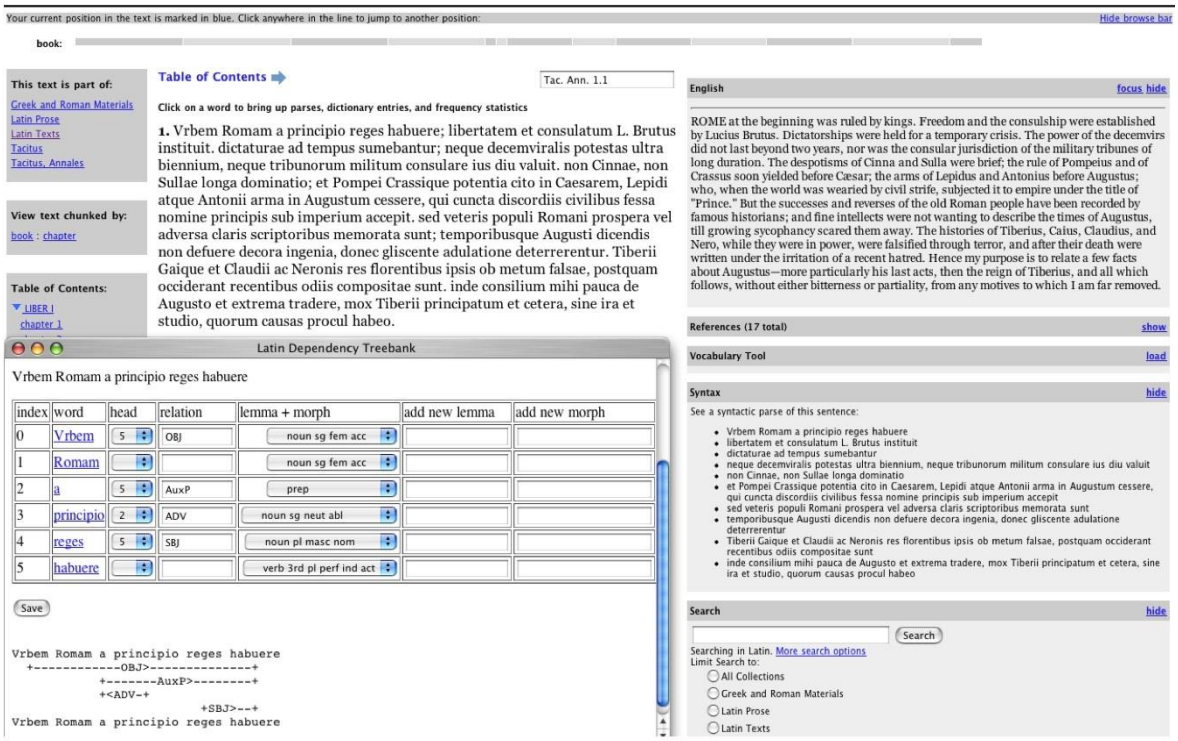
A anotação eficiente do latim e do grego antigo é dificultada pelo fato de que não existem falantes nativos e os textos à nossa disposição são tipicamente altamente estilizados em termos de natureza. Para ajudar com este problema, incorporamos nosso ambiente de anotação à Biblioteca Digital Perseu (PDL, *Perseus Digital Library*)[37]. Criado em 1987 com o intuito de construir uma coleção grande e heterogênea de materiais textuais e visuais sobre o mundo grego arcaico e clássico, hoje, o Perseu serve como um laboratório para tecnologias de bibliotecas digitais e também é amplamente usada por estudantes, acadêmicos e outros para acessar informações sobre o mundo greco-romano (CRANE 1987a; CRANE 1987b; CRANE 1998; CRANE *et al.* 2006; CRANE *et al.*, no prelo).

O conhecimento resultante dos textos históricos desde que foram escritos produziu uma riqueza de materiais contextuais para ajudar os falantes não nativos a compreendê-los, incluindo comentários, traduções e léxicos especializados. O ambiente de leitura do Perseu apresenta o texto de origem grego ou latino e o contextualiza com

[37] N. E. P. Perseu é o nome tanto do Projeto, quanto da Biblioteca Digital; o primeiro é o local institucional responsável pelo conjunto de ambientes e processos em que a segunda, como produto, reside. Entenda-se, assim, Biblioteca Digital do Projeto Perseu.

estas publicações secundárias em conjunto com uma análise morfológica de cada palavra do texto bem como as leituras do manuscrito variante. A Figura 2 apresenta uma imagem da biblioteca digital com uma ferramenta de anotação sintática incorporada à interface. No *widget* à direita, o texto de origem, visualizado (a primeira parte de *Anais* de Tácito) foi automaticamente segmentado em frases; um anotador pode clicar em qualquer frase para lhe atribuir uma anotação sintática. Aqui, o usuário clicou na primeira frase *(Vrbem Romam a principio reges habuere);* esta ação abre uma tela de anotação em que uma análise automática parcial é fornecida, juntamente com a análise morfológica mais provável para cada palavra. O anotador pode então corrigir este resultado automático e passar para a próxima oração segmentada, com todos os recursos contextuais ainda à vista.

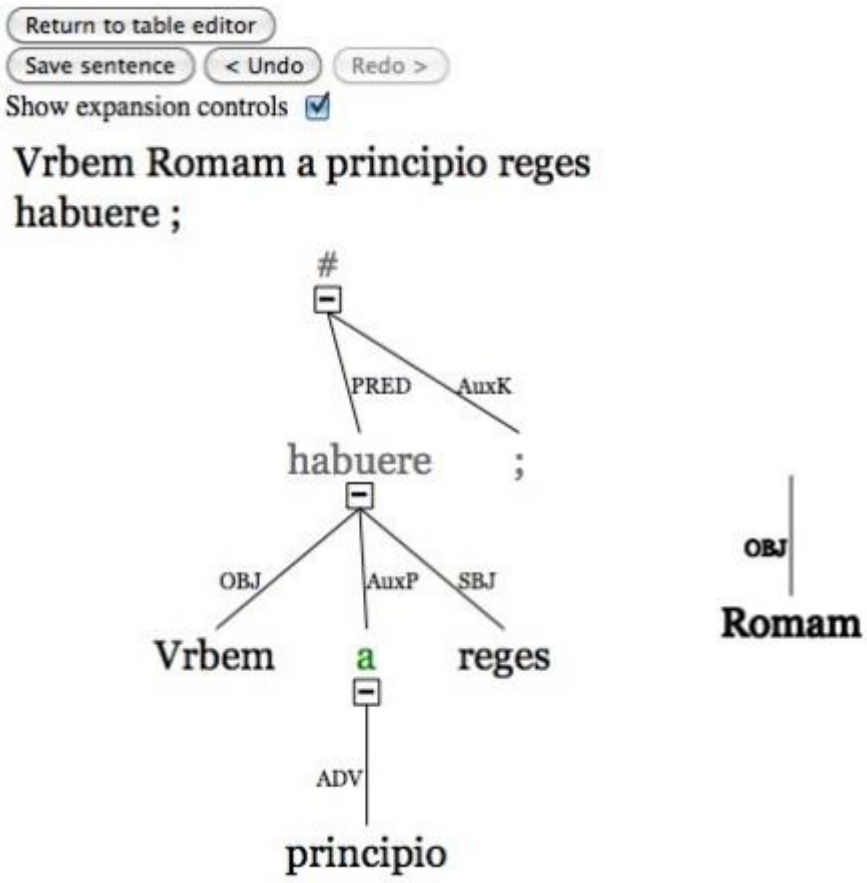
Figura 2 – Uma captura de tela de *Anais* de Tácito da PDL



Nossa colaboração com o Projeto Alpheios também nos permitiu integrar um editor de *treebank* gráfico ao nosso processo de anotação para tornar a construção de árvores mais intuitiva e dar aos anotadores maior flexibilidade quanto ao seu método preferencial de inserção. A Figura 3 mostra uma árvore em processo de construção, com

uma única palavra (Romam) que está sendo arrastada para a cabeça sintática.

Figura 3 – Captura de tela do editor de *treebank* gráfico Alpheios



Além de proporcionar análise morfológica e dicionários digitalizados, como *Lewis e Short's Latin Dictionary* ou o LSJ, as traduções e comentários do Perseu também são especialmente úteis para a compreensão de um texto. Ao situar nosso ambiente de anotação em meio a estes recursos de contextualização, estamos oferecendo suporte para falantes não nativos da língua para maximizar suas contribuições para o *treebank* e reduzindo obstáculos para contribuírem para nosso trabalho. Estas informações contextuais têm maior impacto em alunos iniciantes do que em especialistas, mas é útil para qualquer anotador que queira consultar as interpretações de autoridades publicadas no campo.

Ao incorporar nosso ambiente de anotação a esta infraestrutura on-line, pudemos construir uma rede de anotadores distribuídos não apenas nos Estados Unidos como também em todo o mundo (nossos anotadores não estão apenas na Universidade

de Tufts, Universidade da Califórnia-Berkeley, Universidade da Pensilvânia mas também na Hungria, Reino Unido e Austrália). A colaboração em andamento com vários professores de Clássicos nos permitiu introduzir o conceito de *treebanks* em salas de aula na Universidade de Tufts, Universidade de Missouri-Kansas City, Universidade Furman, The Colllege of Holy Cross e a Universidade de Nebraska-Lincoln.

Métodos de anotação

Com o desenvolvimento de nossos *Treebanks* de Dependência do Grego Antigo e Latim, otimizamos três diferentes métodos de anotação. Com o método de produção "sala de aula", solicitam-se anotações dos alunos na sala (por exemplo, o curso de grego sobre a *Ilíada* de Homero) que, por sua vez, são reconciliados pelo professor; o método de produção "padrão" envolve solicitar anotações de dois anotadores independentes e altamente treinados, cujas diferenças são então reconciliadas por um terceiro; e o método "erudito" segue a tradição de criar uma edição crítica, na qual um único estudioso, com amplo treinamento no tema, cria uma anotação sintática de uma obra e é o único responsável por ele como um ato de interpretação.

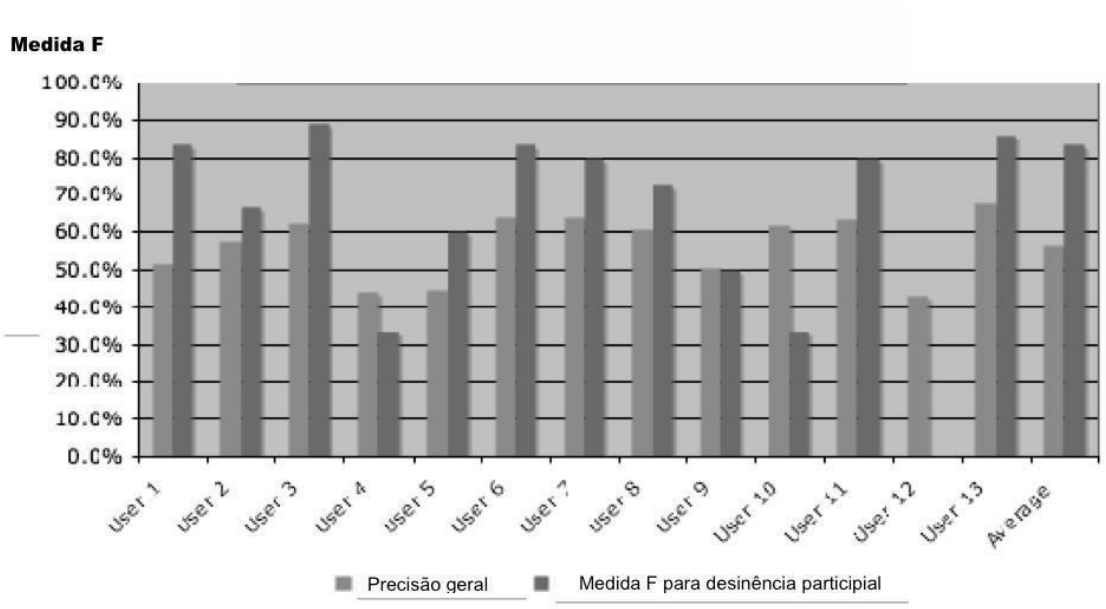
Método de produção "sala de aula"

Apoiamos o uso de *treebanks* em sala de aula em seis universidades nos Estados Unidos – Tufts, Bradeis, The College of the Holly Cross, Furman, Missouri em Kansas City e Nebraska em Lincoln. A principal motivação para este trabalho tem sido pedagógica, uma vez que professores e alunos afirmam que o ato de usar *treebank* é útil para aprender fenômenos gramaticais complexos. Além desta utilidade fundamental, também otimizamos as anotações resultantes como matéria-prima para os nossos *treebanks* publicados. Segundo este método, os alunos fornecem vários fluxos primários de anotação que o professor, como especialista, é responsável por corrigir e enviar.

Em um estudo para avaliar o potencial para esse tipo de contribuição, avaliamos as anotações de um grupo de treze alunos de graduação no *College of the Holy Cross*. Ao contrário dos anotadores no modelo de produção padrão, que passam por meses de treinamento com *feedback* constante sobre seu desempenho, este grupo recebeu apenas

treinamento limitado pelo professor e acesso a um manual on-line de orientações para anotação. A média de precisão geral entre anotadores é de apenas 54,5% devido os diferentes níveis de habilidades dos alunos em sala de aula, porém, o mais importante, descobrimos que alunos diferentes têm (naturalmente) diferentes conjuntos de habilidades – enquanto todos eles apresentam bom desempenho em algumas tarefas (como a modificação atributiva, com uma média da medida F de 91,9% em toda a classe), em outras tarefas a precisão apresenta ampla variação. A Figura 4, por exemplo, mapeia a capacidade destes usuários para identificar corretamente uma ligação participial (isto é, distinguir um uso adverbial de um particípio, como "Reclinado na cama, eu li o livro," a partir de um atributivo de número um, como "o rei errante"). Aqui vemos uma gama muito maior de precisão (relatado novamente como medida F), de 0% (usuário 12) até 89,0% (usuário 3).

Figura 4 – Precisão da anotação de usuário para a desinência participial



Uma recompensa pedagógica de incorporar *treebanks* à sala de aula é a capacidade de identificar automaticamente os pontos fortes e fracos de cada aluno – a Figura 4, por exemplo, identifica que o aluno 12 precisa claramente de mais ajuda na compreensão da desinência participial em grego. Podemos aproveitar este trabalho simultaneamente para a produção de dados de alta qualidade sintaticamente analisados de duas maneiras: primeiro, um professor pode corrigir os fluxos de anotação produzidos pelos alunos na

sala e enviá-los como anotação acabada e final (em que toda a classe e o professor são reconhecidos como proprietários); em segundo lugar, usamos a anotação em sala de aula para ajudar o professor a identificar os alunos com melhor desempenho que podem, então, passar a receber mais treinamento e fornecer as anotações primárias no modelo "padrão").

Método de produção "padrão"

De acordo com o modelo de produção padrão de *treebank*, os anotadores que contribuem para nossos *treebanks* gregos e latinos existentes recebem treinamento extenso com *feedback* constante sobre seu desempenho. A formação destes anotadores varia de estudantes de graduação avançados a doutores recentes e professores; a maioria sendo alunos em programas de pós-graduação em Clássicos. Além de um período de treinamento inicial, os anotadores estão ativamente engajados em um novo aprendizado[38] através um fórum online onde podem fazer consultas mútuas e aos editores do projeto, o que permite que conheçam as atualizações mais recentes das codificações de anotação ao mesmo tempo em que ajudam a treinar novos anotadores. Dois anotadores independentes anotam cada frase e as diferenças são então reconciliadas em um terceiro. Esta reconciliação (ou anotação "secundária", codificada na versão XML) é realizada por um editor/anotador mais experiente, normalmente um doutor com especialização no tema específico (como Homero).

Contudo, análises especializadas são de criação lenta e onerosa, especialmente dada a dificuldade e distância histórica dos textos clássicos. O *Penn Treebank* reporta uma taxa de produtividade entre 750 a 1000 palavras por hora para os anotadores, após quatro meses de treinamento (TAYLOR *et al.*, 2003a) e o *Penn Treebank* para chinês reporta uma de 240 a 480 palavras por hora (CHIOU *et al.*, 2001), mas não há falantes nativos de línguas históricas como o grego. Nossas velocidades de anotação, portanto, são significativamente mais lentas, variando de 92 palavras por hora para 224, com uma média de 130.

[38] Os fóruns de latim e grego antigo estão disponíveis aqui: http://treebank.alpheios.net/forum/.

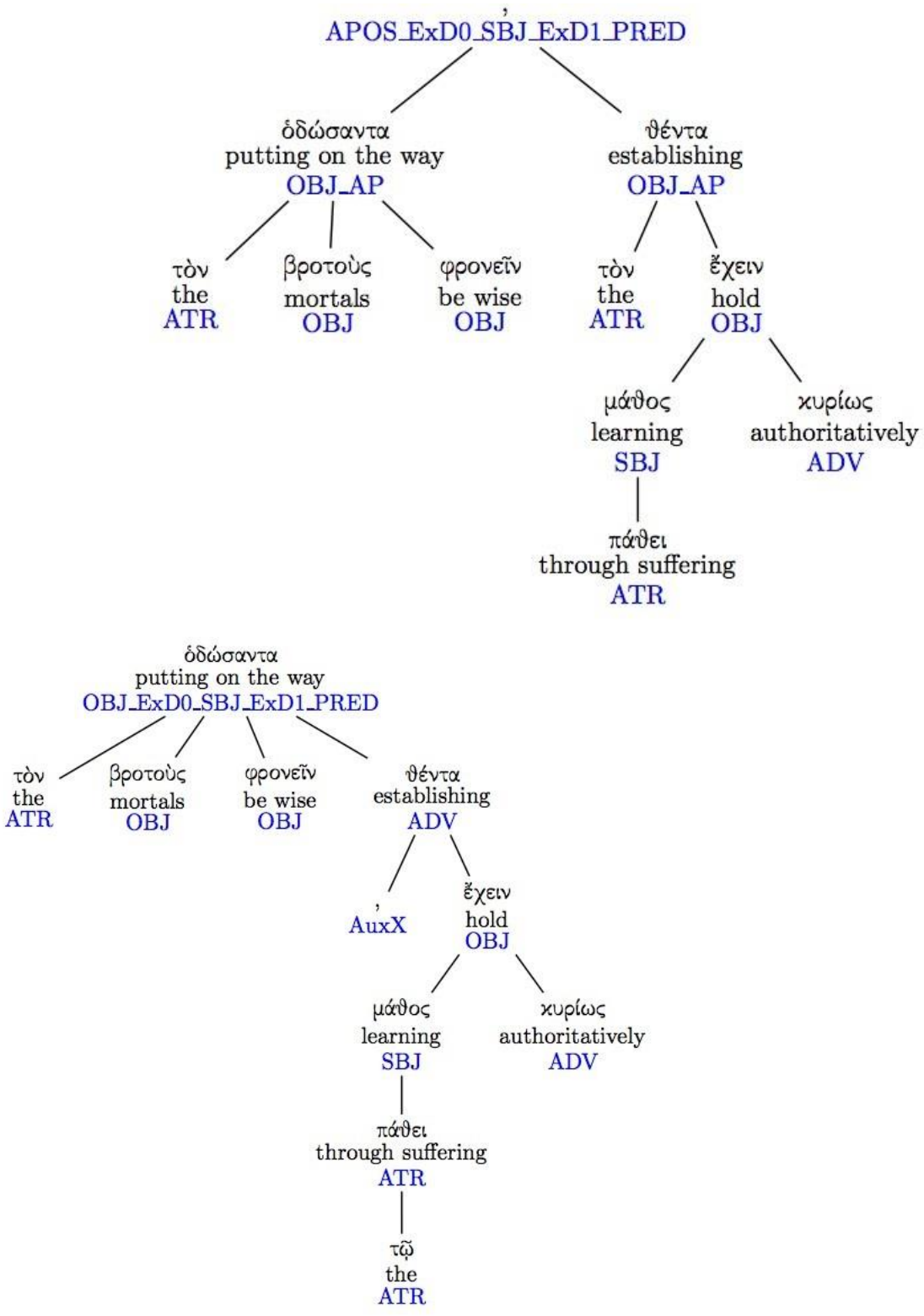
Método de produção "erudito"

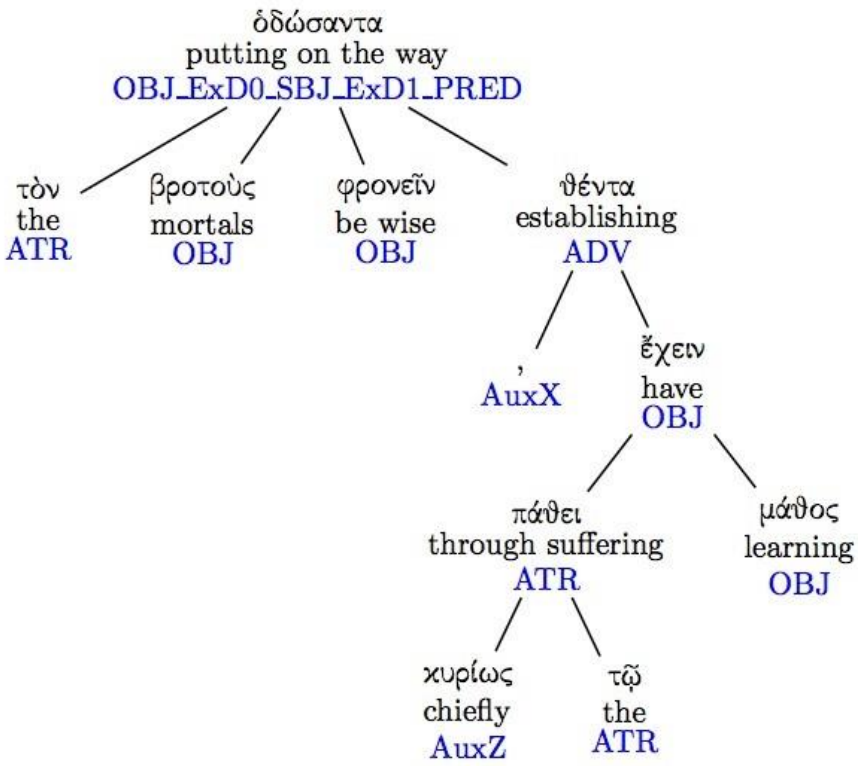
Com nosso *treebank* das obras completas de Ésquilo, investigamos um novo modo de produção: um único estudioso preenche uma anotação sintática de uma obra completa e a trata como uma interpretação autônoma do texto.

A motivação para este trabalho é fundamentalmente a diferente natureza de *treebanks* históricos comparados aos modernos. Enquanto um artigo do *Wall Street Journal* é certamente mais representativo de como os falantes nativos de inglês realmente falam do que a *Ilíada* de Homero é para os gregos antigos, a *Ilíada* tem sido o foco de estudo por quase 3.000 anos, com alunos e professores examinando cada palavra, anotando sua sintaxe, semântica e outros níveis linguísticos de forma particular à margem de livros ou em comentários publicados. Obviamente, a ambiguidade está presente em todas as línguas, porém as decisões individuais e específicas que os anotadores tomam ao resolver ambiguidade sintática ao criar *treebanks* modernos há séculos são discutidas para outros textos flagrantemente clássicos ou históricos; teses e carreiras inteiras foram feitas com o estudo de única obra de um único autor.

A Figura 5, por exemplo, ilustra a complexidade que envolve a interpretação textual de uma única sentença de *Agamêmnon* de Ésquilo (τὸν φρονεῖν βροτοὺς ὁδώσαντα, τὸν πάθει μάθος θέντα κυρίως ἔχειν) "[Zeus]... que colocou os homens no caminho da sabedoria, que estabeleceu que a lei 'aprender através do sofrimento' estará em vigor", linhas 176-8).

Figura 5 – Três interpretações diferentes de uma sentença de *Agamêmnon* de Ésquilo como análises sintáticas processáveis por máquina. Árvore sintática de *Ag.* 176-178 (DENNISTON-PAGE (a), FRAENKEL (b) e BOLLACK(c)).





A fórmula πάθει μάθος ("aprender através do sofrimento") é citada e comentada em muitas introduções gerais para o teatro de Ésquilo (foi até citada por Robert F. Kennedy em seu discurso sobre o assassinato de Martin Luther King Jr. (KENNEDY, 1968), porém, o texto e a interpretação sintática da frase são altamente controversos (BAMMAN *et al.*, 2009). Os três comentários mais recentes sobre a peça – Fraenkel (1950), Denniston-Page (1957) e Bollack (1981) – adotaram três soluções muito diferentes com base em seu próprio peso de evidência filológica, cada um resultando em uma árvore sintática marcadamente diferente.

A variedade de interpretações textuais e sintáticas para apenas estas três linhas de Ésquilo começa a apontar as deficiências de um modelo de produção padrão de *treebank* para textos em debate acadêmico continuado. Ao criar um *corpus* anotado de um idioma para o qual não existem falantes nativos (e para o qual, consequentemente, não se pode confiar em intuições nativas), estamos construindo sobre uma montanha conhecimento anterior que formou nossa compreensão fundamental do texto.

Conclusão

Os *treebanks* gregos e latinos não são simplesmente bancos de dados para a pesquisa, mas catalisadores para a nova vida intelectual. Duas implicações em particular se destacam. Primeiro, eles abrem para alunos de graduação oportunidades no grego e latim semelhantes aos seus colegas em outras áreas de ciências para fazer contribuições concretas. Em segundo lugar, os *treebanks* não são apenas bancos de dados produzidos industrialmente, são repositórios de interpretações processáveis por máquina. A frase sintaticamente analisada é uma nova forma para a publicação das conclusões acadêmicas sobre a língua – uma forma que é, em si, em grande parte independente de língua. Se o idioma de publicação preferido pelo pesquisador for inglês ou alemão, árabe ou chinês, a árvore de análise parece a mesma. Os *treebanks* gregos e latinos, assim, abriram novas possibilidades para estudantes e para os pesquisadores avançados para participar de forma mais completa do estudo da cultura greco-romana do que era possível em textos impressos.

Referências

ABEILLE, A.; CLEMENT, L.; KINYON, A.; TOUSSENEL; F. Building a Treebank for French. In: **Proceedings of the Second Conference on Language Resources and Evaluation,** Athens, p. 87-94, 2000.

BAMMAN, D.; MAMBRINI; F.; CRANE; G. An Ownership Model of Annotation: The Ancient Greek Dependency Treebank. In: **Proceedings of the Eighth International Workshop on Treebanks and Linguistics Theories,** 2009.

BOLLACK, J.; COMBE, P. J. de L. **L'Agamemnon d'Eschyle: le texte et ses interpretations**. Lille : Presses universitaires de Lille, 1981.

BRANTS, S.; DIPPER, S.; HANSEN, S.; LEZIUS, W.; SMITH; G. The TIGER Treebank. In: **Proceedings of the Workshop on Treebanks and Linguistic Theories.** Bulgaria: Sozopol, 2002.

CHIOU, F.; CHIANG, D.; PALMER; M. Failitating Treebank Annotation Using a Statistical Parser. In: **Proceedings of the First International Conference on Human Language Technology Research Hlt '01,** p. 1-4, 2001.

CRANE, G. Clay Balls and Compact Disks: Some Political and Economic Problems of New Storage Media. **Favonius Supplement**, v. 1, p. 1-6, 1987.

CRANE, G. From the Old to the New: Integrating Hypertext into Traditional Scholarship. In: **Hypertext '87:** Proceedings of the 1st ACM conference on Hypertext, p. 51-56, 1987.

CRANE, G. New Technologies for Reading: The Lexicon and the Digital Library. **Classical World**, p. 471-501, 1998.

CRANE, G.; BAMMAN, D.; CERRATO, L.; JONES, A.; MIMNO, D. M.; PACKEL, A.; SCULLEY, D.; WEAVER, G. Beyond Digital Incunabula: Modeling the Next Generation of Digital Libraries. In: **Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries**, ECDL, p. 353-366, 2006.

CRANE, G.; BAMMAN, D.; JONES; A. Philology in an Electronic Age. In: **Greek Lexicography After Liddell and Scott. Pre-publication.** Disponível em: http://geryon.perseus.tufts.edu/data/Papers and Props/Philology.pdf.

PAGE, D. **Aeschylus Agamemnon**. Edited by the late John Dewar Denniston and Denys Page. Oxford: Clarendon Press, 1957.

DZEROSKI, S.; ERJAVEC, T.; LEDINEK, N.; PAJAS, P.; ZABOKRTSKY; Z.; ZELE, A. Towards a Slovene Dependency Treebanks. In: **Proceedings of the Fifth International Conference on Language Resources and Evaluation**, ELRA, Genoa, 2006.

FRAENKEL, E. **Aeschylus. Agamemnon**. Oxford: Clarendon Press, 1950.

HAJIC, J. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In: HAJICOVA, E. (Ed.). **Issues of Valency and Meaning**. Studies in Honor of Jarmila Panevova. Prague: Charles University Press, 1998.

HAJIC, J.; SMRZ, O.; ZEMANEK, P.; SNAIDAUF, J.; BESKA, E. Prague Arabic Dependency Treebank: Development in Data and Tools. In: **Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools**, 2004.

HAUG, D. T. T.; JOHNDAL, MX. Creating a Parallel Treebank of the Old Indo-European Bible Translations. In: **Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data**, LaTeCH, 2008.

KENNEDY, R. F. **Statement on the assassination of Martin Luther King.** Indianapolis, Indiana, April 4, 1968.

KROCH, A.; TAYLOR, A. **Penn-Helsinki Parsed Corpus of Middle English**. 2. ed. 2000. Disponível em: http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-2/.

KROCH, A.; SANTORINI, B.; DELFS, L. **Penn-Helsinki Parsed Corpus of Early Modern English.** 2004. Disponível em: http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-1.

KUROHASHI, S.; NAGAO, M. Building a Japanese Parsed Corpus while Improving the Parsing System. In: **Proceedings of the First International Conference on Language Resources and Evaluation** (Granada).

MARCUS, M. P.; MARCINKIEWICZ, M. A.; SANTORINI, B. Building a Large Annotated Corpus of English: The Perm Treebank. **Computational Linguistics**, v. 19, p. 313-330, 1998.

MONTEMAGNI, S.; BARSOTTI, F.; BATTISTA, M.; CALZOLARI, N.; CORAZZARI, O.; LENCI, A.; ZAMPOLLI, A.; FANCIULLI, F.; MASSETANI, M.; RAFFAELLI, R.; BASILI, R.; PAZIENZA, M. T.; SARACINO, D.; ZANZOTTO, F.; MANA, N.; PIANESI, F.; DELMONTE, R. The Italian Syntactic-Semantic Treebank: Architecture, Annotation, Tools and Evaluation. In: **Proceedings of the COLING Workshop on Linguistically Interpreted Corpora** (LINC-2000), 2000.

MORENO, A.; GRISHMAN, R.; LOPEZ, S.; SANCHEZ, F.; SEKINE, S. A Treebank of Spanish and its Application to Parsing. In: **Proceedings of the Second Conference on Language Resources and Evaluation**, 2000.

PROKOPIDIS, P.; DESIPRI, E.; KOUTSOMBOGERA, M.; PAPAGEORGIOU, H.; PIPERIDIS, S. Theoretical and Practical Issues in the Construction of a Greek Dependency Treebank. In: **Proceedings of the 4th Workshop on Treebanks and Linguistic Theories** (TLT), p. 149-160, 2005.

ROCIO, V.; ALVES, M. A.; GABRIEL LOPES, J.; XAVIER, M. F.; VICENTE, G. Automated Creation of a Medieval Portuguese Partial Treebank. In: ABEILLE, A. (Ed.). **Treebanks: Building and Using Parsed Corpora.** Dordrecht: Kluwer Academic Publishers, 2000. p. 211-227.

TAYLOR, A.; MARCUS, M.; SANTORINI, B. The Penn Treebank: An Overview. In: ABEILLE, A. (Ed.). **Treebanks: Building and Using Parsed Corpora.** Dordrecht: Kluwer Academic Publishers, 2003. p. 5-22.

TAYLOR, A.; WARNER, A.; PINTZUK; S.; BETHS; F. **York-Toronto-Helsinki Parsed Corpus of Old English Prose**. University of York, 2003.

ZEMANEK, P. A Treebank of Ugaritic: Annotating Fragmentary Attested Languages. In: **Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories** (TLT2007), Bergen, p. 213-218, 2007.

CONSTRUINDO UM LÉXICO DINÂMICO A PARTIR DE UMA BIBLIOTECA DIGITAL[39]

David Bamman[40]
Gregory Crane[41]

Introdução

Desde o projeto COBUILD (SINCLAIR, 1987) na década de 1980, lexicógrafos têm explorado grandes corpora de conhecimento estruturado, muitas vezes extraindo a contagem de frequência e colocações – a informação de frequência de uma palavra é especialmente importante para aprendizes da segunda língua, e locuções (o "companheiro" de uma palavra) são fundamentais para delimitar seu significado. Esta abordagem baseada em *corpus* para a construção de léxico desde então foi ampliada em duas dimensões: Por um lado, os dicionários e recursos lexicográficos estão sendo construídos em coleções textuais cada vez maiores: o projeto alemão *elexiko* (KLOSA *et al.*, 2006), por exemplo, é construído sobre um *corpus* moderno alemão de 1,3 bilhão de palavras e podemos esperar projetos muito maiores no futuro à medida que a *web*[42] é explorada como um *corpus*.

Ao mesmo tempo, os pesquisadores estão também submetendo seus corpora a processos automáticos mais complexos para extrair mais conhecimento. Enquanto a frequência de palavras e análise da locução é fundamentalmente uma tarefa de contagem simples, projetos como o Sketch Motor de Kilgarriff (2004) também permitem aos lexicógrafos influir a informação sobre o comportamento gramatical de uma palavra.

Atualmente, estamos no processo de criação de um vocabulário dinâmico personalizado a partir dos textos clássicos na Biblioteca Digital Perseu (CRANE, 1987;

[39] Publicado originalmente em 2008, com o título: *Building a Dynamic Lexicon from a Digital Library*, pela JCDL'08. Em português, distribuído sem fins comerciais, sob permissão dos autores.
[40] Perseus Project – Department of Classics Tufts University – Medford MA – 02140 – USA – david.bamman@tufts.edu gregory.
[41] Perseus Project – Department of Classics Tufts University – Medford MA – 02140 – USA – crane@tufts.edu.
[42] Em 2006, por exemplo, o Google lançou a primeira versão do seu *corpus* de 1T 5-gram na web [6], uma coleção de n-gramas (n = 1-5) e suas frequências calculadas a partir de 1 trilhão de palavras do texto na web.

CRANE *et al.*, 2001). Este léxico vai apresentar um inventário de sentido (juntamente com as informações de frequência) para qualquer lexema grego ou latim da forma como é usado em qualquer autor, época ou gênero encontrado em nossa coleção, juntamente com informações estatísticas sobre seus quadros de subcategorização comuns e preferências selecionais.

Figura 1 – Gráfico de dependência da anotação *treebank* para *quem ad finem sese effrenata iactabit audacia* ("Até que ponto (tua) audácia se gabará?"), Cícero, *In Catilinam* 1.1

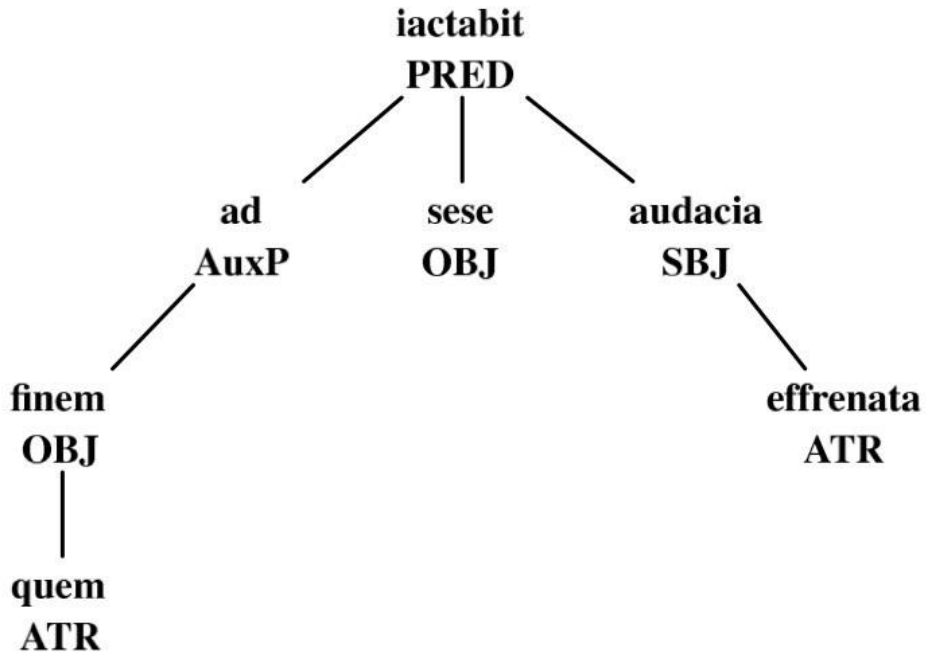


Figura 2 – Versão XML da anotação *treebank* para *quem ad finem sese effrenata iactabit audacia*, Cícero, In 1.1

```
<sentence id="74" document_id="Perseus:text:1999.02.0010" subdoc="text=Catil.:Speech=1:chapter=1" span="quem0:audacia0">
  <word id="1" form="quem" lemma="quis1" postag="p-s---ma-" head="3" relation="ATR"/>
  <word id="2" form="ad" lemma="ad1" postag="r--------" head="6" relation="AuxP"/>
  <word id="3" form="finem" lemma="finis1" postag="n-s---ma-" head="2" relation="OBJ"/>
  <word id="4" form="sese" lemma="sui1" postag="p-s---fa-" head="6" relation="OBJ"/>
  <word id="5" form="effrenata" lemma="effreno1" postag="t-srppfn-" head="7" relation="ATR"/>
  <word id="6" form="iactabit" lemma="jacto1" postag="v3sfia---" head="0" relation="PRED"/>
  <word id="7" form="audacia" lemma="audacia1" postag="n-s---fn-" head="6" relation="SBJ"/>
</sentence>
```

O inventário de sentido em si depende de tecnologias de indução de sentido da palavra e desambiguação, porém, a extração dos quadros de subcategorização e preferências selecionais de uma palavra baseia-se na marcação morfológica automática e

análise sintática. Etiquetadores morfológicos de última geração podem alcançar taxas de precisão de mais de 96% para inglês (RATNAPARKHI, 1996; SCHMID, 1994) e 92% para as línguas altamente flexionadas como checo (HAJLIC; HLADKÁ, 1998); os analisadores de dependência podem alcançar taxas de precisão rotulada para as mesmas línguas de 86% (NIVRE *et al.*, 2007) e 80% (COLLINS *et al.*, 1999) respectivamente[43]. Esses serviços, no entanto, atingem estas precisões altas devido a treino com grandes[44] volumes de dados anotados manualmente, geralmente mais de um milhão de palavras.

Em contraste, temos um *treebank* de latim com 30.457 palavras. O pequeno tamanho do treino deste conjunto de dados gera uma etiquetagem e análise geral previsivelmente inferiores. Como Church e Hovy (1993) comentaram para tradução automática, no entanto, a avaliação do desempenho de um sistema depende da aplicação. As 30.457 palavras podem não ser suficientes para a análise sintática precisa como um fim em si, mas as frases imperfeitamente analisadas resultantes são suficientes para causar informação léxica robusta dado o grande número suficiente delas. Usando as mesmas técnicas simples de teste de hipóteses usadas para encontrar **locuções** (entre frases cheias de ruído), podemos identificar as preferências selecionais comuns de uma palavra quando a análise automática tem ruído. Isto é promissor para outras línguas com recursos baixos e bibliotecas digitais que procuram otimizar pequenas fontes de conhecimento estruturadas em comparação a grandes coleções não estruturadas. Enquanto o trabalho aqui foi desenvolvido no contexto de uma única biblioteca digital, as técnicas de aprendizado supervisionadas que descrevemos podem ser usadas por um conjunto de qualquer tamanho, com um pequeno conjunto de dados anotados.

Recursos

Temos dois tipos diferentes de recursos em nossa biblioteca digital: um conjunto pequeno, mas com curadoria humana de dados sintaticamente anotados e um corpo muito maior com textos não anotados.

[43] A precisão de análise não rotulada (em que apenas o nó é avaliado não a relação sintática), apresenta taxas de precisão mais elevadas de 91% para inglês (COLLINS *et al.*, 1999) e 84% para checo (MCDONALD *et al.*, 2005).

[44] O Penn Treebank (MARCUS *et al.*, 1994), por exemplo, contém mais de um milhão de palavras (em estilo PTB-2), enquanto o Prague Dependency Treebank (HAJIC *et al.*, 1999) contém 1,5 milhão.

Dados anotados

A pequena fonte de conhecimento estruturado à nossa disposição é o *treebank* do latim clássico com 30.457 palavras. Agora, na versão 1.4, o *Treebank* de Dependência do Latim é composto de excertos de cinco textos: *Commentarii de Bello Gallico* de César, *Oratio in Catillinam* Cícero, *Bellum Catilinae* de Salústio, de Virgílio e *Vulgata* de Jerônimo, como mostrado na tabela 1.

Tabela 1 – Composição do *treebank* de dependência latina

Data	**Autor**	**Palavras**	**Sentenças**
séc I a.C	César	1.488	71
séc I a.C	Cícero	5.663	295
séc I a.C	Salústio	12.311	701
séc I a.C	Virgílio	2.613	178
séc IV-V d.C.	Jerônimo	8.382	405
	Total	30.457	1.650

O *treebank* é uma grande coleção de frases que foram sintaticamente anotadas. O conhecimento codificado nesta estrutura é um trabalho extremamente intenso, pois dois anotadores independentes anotam cada frase e suas anotações são reconciliadas por uma terceira. O próprio processo de anotação envolve a especificação da relação sintática exata de cada palavra em uma frase (por exemplo, qual é o sujeito, qual é o objeto, onde a locução prepositiva deve ser anexada, qual adjetivo modifica qual substantivo, etc.) Além disso, para o índice do seu nó sintático e o tipo de relação com ele, cada palavra no *treebank* também é anotada com o lema a partir da qual ela é flexionada (por exemplo, *est* é uma forma flexionada do lema *sum*) e seu código morfológico (por exemplo, *est* é 3ª pessoa singular do verbo indicativo ativo).

As Figuras 1 e 2 apresentam dois pontos de vista de uma anotação sintática para uma única frase *(quem ad finem sese effrenata iactabit audacia)*[45]. A Figura 1 mostra a estrutura conceitual para a árvore de dependência que resulta da anotação (sujeitos e objetos, por exemplo, são ambos filhos dos verbos que modificam) e a Figura 2

[45] "Até que ponto (tua) audácia se gabará?" (Cícero, *In Catilinam* 1.1).

apresenta uma serialização em XML da árvore (o formato em que liberamos nossos dados).

Uma vez que o latim tem uma ordem de palavras altamente flexível, baseamos nosso estilo de anotação na gramática de dependência usada pelo *Prague Dependency Treebank* (PDT) (HAJIC, 1998) para o checo adaptando-o para o latim através da gramática da Pinkster (1990). Gramáticas de dependência diferem das gramáticas com base em termos constituintes por ignorar categorias frasais não terminais (como NP ou VP), pelo contrário, ligando as próprias palavras a seu nó imediato. Esta é uma forma especialmente adequada de representação para as línguas com ordem de palavras ou moderadamente livre (como o latim e o checo), onde a ordem linear dos constituintes é quebrada com elementos de outros constituintes.

Para tornar nosso estilo de anotação tão útil quanto possível, também estamos colaborando com outros *treebanks* latinos (por exemplo, o *Thomisticus Index* (PASSAROTI, 2007; BUSA, 1974-1980) sobre as obras de Tomás de Aquino) para criar um conjunto de orientações de anotação a ser usado como um padrão para o latim de qualquer período (BAMMAN *et al.*, 2007a, 2007b). Este trabalho também nos permitiu compartilhar nossos dados à medida que anotamos os respectivos textos (BAMMAN *et al.*, 2007a, 2007b).

Dados não anotados

Tabela 2 – Precisão morfológica por característica

	Precisão
Caso	90,10%
Grau	99,92%
Gênero	92,90%
Modo	98,68%
Número	95,15%
Parte do discurso	95,11%
Pessoa	99,56%
Tempo	98,62%
Voz	98,89%
Todos	83,10%

O conjunto de dados sintaticamente anotados que temos em nossa coleção é ínfimo em comparação com o tamanho do *corpus* não anotado. A Biblioteca Digital Perseu contém cerca de 3,5 milhões de palavras de textos de origem latina, junto com 4,9 milhões de palavras de grego. Embora estes textos não sejam estruturados sintaticamente, cada um contém metadados extensos detalhando autor e as subcoleções à qual a obra pertence (como a poesia latina ou prosa latina).

Nossa abordagem para extrair informação lexical desta grande coleção envolve primeiro atribuir análises sintáticas a todas as frases que ele contém. Não podemos, é claro, analisar manualmente cada frase à mão, assim, a estrutura sintática deve ser atribuída automaticamente. A análise sintática de última geração é um processo de aprendizado supervisionado em que um analisador é treinado em um conjunto de dados humanos anotados. A análise de desempenho está fortemente ligada ao tamanho dos dados de treinamento; grandes *treebanks* (mais de um milhão de palavras) apresentam melhor desempenho. Antes de pesquisar este *corpus*, avaliamos o desempenho do próprio algoritmo de análise em nosso pequeno conjunto de dados e da etiquetagem morfológica automática em que se baseia.

Avaliação

Avaliamos a precisão de etiquetagem morfológica automática, usando o analisador TreeTagger (SCHMID, 1994) e da análise sintática automática usando o analisador de dependência de McDonald *et al.* (2005). Em todos os testes que se seguiram, as taxas de precisão reportadas são o resultado de um teste com 10 etapas nas 30.457 palavras do *treebank*, nas quais o etiquetador ou analisador é treinado em 90 por cento do *treebank* (ca. 27411 palavras) e testados nos dez por cento restantes; este teste é realizado um total de dez vezes, uma para cada décimo mantido fora; a precisão reportada corresponde à média de todos os testes.

Etiquetagem morfológica

Como parte de uma língua altamente flexionada, as palavras latinas têm uma análise morfológica composta de nove características: parte do discurso, pessoa, número, tempo, modo, voz, gênero, caso e grau. O analisador TreeTagger realizou com uma precisão de 83% na desambiguação correta da análise morfológica completa. Para resolver a parte simples do discurso, seu desempenho é próximo ao de línguas com maior número de recursos (95%), mas a flexão complexa do latim apresenta mais dificuldades para resolver gênero e caso. Estas duas características têm uma entropia superior na língua devido à sua ambiguidade sobreposta[46].

Análise sintática

A maioria das avaliações de precisão de análise pressupõe um padrão-ouro para as marcas morfológicas subjacentes com o intuito de isolar o ganho ou a perda específica no próprio analisador. Na determinação da precisão funcional, poderíamos esperar de um analisador atribuir uma análise sintática a todas as frases em nosso *corpus* (para o qual devemos atribuir automaticamente também uma análise morfológica), apresentamos duas avaliações: uma para analisar um *corpus* com etiquetas morfológicas conhecidas (“ouro”) e outra para a análise de um *corpus* para o qual as etiquetas morfológicas foram automaticamente atribuídas (“automático”). A precisão sem rótulo mede quantas vezes o nó sintático de uma palavra é correta, enquanto a precisão rotulada também mede se a etiqueta sintática correta (como sujeito *versus* objeto) foi aplicada também.

Tabela 3 – Precisão de análise

	Sem etiqueta	**Etiquetada**
Ouro	64,99%	54,34%
Automático	61,49%	50,00%

[46] Por exemplo, uma palavra como *magna* (grande) pode ser um adjetivo nominativo feminino ou neutro acusativo (além do ablativo feminino ou neutro nominativo).

Como esperado, a precisão para a avaliação de ouro é muito mais baixa do que a relatada para idiomas como inglês e tcheco. Com as etiquetas morfológicas automáticas, podemos esperar encontrar cerca de metade das relações sintáticas em uma frase. Contudo, podemos dividir este valor ainda mais. A precisão global relatada na Tabela 3 é um conjunto de todos os autores, gêneros e relações sintáticas. Se dividirmos esses resultados por autor (Tabela 4), encontramos uma forte correlação entre o rigor de análise e a não projetividade do autor – a relação com os quais constituintes frasais são quebrados por outros constituintes[47]. Jerônimo, um autor de prosa no século IV d. C., tem uma baixa taxa de não projetividade de 1,8%, enquanto Virgílio, um poeta da Idade de Ouro, tem a maior, 12,2%. A não projetividade alta é uma característica da poesia latina como uma forma de efeito retórico (hipérbato), assim, podemos esperar a queda de nossas menores taxas de precisão no futuro entre as obras dos poetas estilizados e a presença das maiores as em autores de prosa estrita. Felizmente (a esse respeito), o *corpus* da poesia latina é muito menor do que a prosa (a Biblioteca Digital Perseu, por exemplo, inclui 593.000 palavras de poesia latina e 2,9 milhões de palavras de prosa).

Tabela 4 – Precisão de análise etiquetada por autor

	Ouro	**Automático**
Jerônimo	61,44%	58,15%
Salústio	53,04%	46,99%
César	51,34%	46,24%
Cícero	49,97%	44,41%
Virgílio	48,99%	40,60%

Outra variável incluída nesta taxa de precisão geral é o desempenho do analisador por etiqueta individual. Conforme mostra a Tabela 5, precisão[48] e recall[49] são muito maiores para adjetivos atributivos (ATR), desinências de frase preposicional (AuxP), sujeitos (SBJ), objetos (OBJ) e advérbios (ADV) do que para a desinência de conjunção subordinativa (AuxC) e qualquer relação envolvida na coordenação (CO). Isso é um bom

[47] Veja Nivre [30] para uma definição formal de projetividade.
[48] Definimos precisão aqui como o número de vezes que uma etiqueta X é atribuída corretamente ao nó correto dividido pelo número de ocorrências desta etiqueta X no *corpus* analisado automaticamente.
[49] Definimos recall aqui como o número de vezes em que uma etiqueta X é atribuída corretamente ao nó correto, dividido pelo número de ocorrências desta etiqueta X no *corpus* de teste.

sinal para a extração de preferências selecionais a partir de um *corpus*, uma vez que as relações que vamos procurar serão exatamente estes – enquanto a precisão de sujeitos e objetos ainda paira em torno de 50%, a precisão de atributos, pelo menos, é mais elevado, 63%.

Tabela 5 – Precisão/recall rotulada por etiqueta sintática

Ouro			**Automático**		
	Precisão	**Recall**		**Precisão**	**Recall**
ATR	68,17%	71,20%		63,09%	62,41%
AuxP	67,38%	69,80%		63,66%	66,81%
SUJ	61,95%	62,24%		50,93%	51,10%
OBJ	59,33%	62,84%		50,90%	55,12%
ADV	53,72%	59,90%		49,24%	55,31%
AuxC	39,30%	39,00%		34,80%	36,04%
SUJ CO	38,20%	39,81%		26,58%	29,04%
OBJ CO	37,90%	38,48%		31,84%	30,85%
ATR CO	34,38%	27,76%		30,35%	25,17%
ADV CO	34,27%	27,84%		30,29%	22,22%

Extraindo preferências selecionais

A preferência selecional de um predicado especifica o tipo de argumento com que geralmente aparece. O verbo comer, por exemplo, normalmente requer que seu objeto seja algo que possa ser comido e que seu sujeito seja animado, exceto se usado metaforicamente. A preferência selecional, no entanto, também pode ser muito mais detalhada, o que reflete não apenas uma classe de palavra (como animada ou humana), mas também as próprias palavras. Por exemplo, o tipo de argumento usado com o verbo latino *libero* (libertar) são muito diferentes em Cícero e Jerônimo, com base em um estudo pequeno manual de 100 ocorrências do verbo (BAMMAN; CRANE, 2007): Cícero, como orador da República, geralmente usa-o utiliza para falar da libertação do *periculum* (perigo), *metus* (medo), *cura* (tratamento) e *aes alienum* (dívida); Jerônimo, por outro lado, usa-o para falar de libertação de um conjunto muito diferente de coisas, tais como

manus Aegyptorum (a entrega dos egípcios), *os leonis* (a boca do leão) e *mors* (morte). Estas são qualidades sintáticas, pois cada um desses argumentos tem uma relação sintática direta com seu nó assim como ocupa um lugar semântico dentro da estrutura do argumento subjacente.

As preferências selecionais são uma variedade de locuções e podem ser extraídas através de métodos semelhantes (CHURCH; HANKS, 1989) – onde locuções podem ser encontradas por comparação da contagem de duas palavras que ocorrem em conjunto (em geral, em um intervalo fixo de palavras), com a probabilidade independente de cada uma ocorrendo sozinha. As preferências selecionais podem ser encontradas ao se estabelecer a probabilidade de que uma palavra apresente uma relação sintática específica com outra – a mais informativa são objetos diretos (OBJ). Usando clusters (ROOTH *et al.*, 1999) ou semelhança métricas do WordNet (STEPHEN; WEIR, 2002), podemos também usar as frequências individuais para as palavras a generalizar para a classe de palavra que um predicado prefere.

Etiquetando os dados

Para extrair preferências selecionais de nosso *corpus* de latim com 3,5 milhões de palavras, primeiro treinamos nosso etiquetador e analisador no *treebank* completo; em seguida, usamos estes modelos treinados para etiquetar morfologicamente todo o *corpus* e, depois, atribuímos uma estrutura sintática dos textos etiquetados automaticamente.

Extraindo conhecimento

Com todo o corpo etiquetado e analisado, agora, podemos extrair dele preferências selecionais. Contudo, a robustez da associação é enviesada por frequência global de uma palavra em um *corpus*, de modo que a palavra de alta frequência, naturalmente, seria um argumento comum para muitos verbos transitivos. Podemos superar isso usando as mesmas técnicas de teste de hipóteses utilizadas para encontrar locuções comuns. O teste de probabilidade de log (λ) (DUNNING, 1993), por exemplo, mede quantas vezes duas palavras ocorrem juntas em uma frase em comparação com a frequência que seria de se

esperar para encontrá-las juntas, dadas as suas frequências no *corpus* geral[50]. Para adotar esta medida para encontrar preferências selecionais comuns, podemos definir as contagens relevantes deste modo:

c_1 = contagem de $lema_1$ no *corpus*
c_2 = contagem de $lema_2$ no *corpus*
c_{12} = contagem de $lema_2$ dependendo de um argumento do $lema_1$ Com o valor de log λ:
$\log \lambda = \log L (c_{12} , c_1 , p) + \log L (c_2 - c_{12}, N - c_1, p) - \log L (c_{12} , c_1 , p_1) - \log L(c_2 - c_{12} , N - c_1, p_2)$ onde

p = c2 , p_1 = c_{12} , p2 = c2-c12, N **contagem do *corpus***
N *c1* N-c1
e $L (a, b, c) = c^a (1 - c)^{b-a}$

Para alcançar um nível de confiança de $\alpha = 0,05$ de que um lema é um argumento comum de outro seu valor de $-2 \log \lambda$ log deve estar acima de 3,84.

Resultados

A força do teste de hipóteses é que ele também nos permite superar os dados com ruído. Dado o tamanho do nosso *corpus*, podemos tolerar erros de análise ocasionais uma vez que a contagem da maioria dos lemas é relativamente alta: se uma palavra é um argumento comum verdadeiro de outra palavra, ela aparecerá como este argumento várias vezes em 3,5 milhões de palavras.

Pode-se ver a força desta abordagem nas Tabelas 6 e 7. A Tabela 6 apresenta as relações mais fortes encontradas entre duas palavras em todo o *corpus* (e não as preferências ou argumentos selecionais simplesmente, mas todas as palavras que têm alguma relação sintática entre si). Nove dos dez pares de palavras são conectados com uma relação atributiva e contêm locuções fortes.

[50] Usamos probabilidade de log como distinta de informação mútua para evitar privilegiar locuções em palavras de baixa frequência em vez pares melhor comprovados. Para o nosso fim, a probabilidade de log e $\chi 2$ é, em grande parte, intercambiável – uma avaliação $\chi 2$ de fazer, por exemplo, oferece a mesma lista idêntica classificada como a encontrada usando probabilidade (Tabela 7) abaixo.

Tabela 6 – Dez colocados sintáticos mais fortes (formas de raiz não declinada exibidas)

Latim	**Português**[51]	**Relação**	**−2logλ**
res publicus	república	ATR	3.840,0
populus romanus	povo romano	ATR	2.450,8
pater conscribo	pai conscrito	ATR	612,6
filius Israhel	filho de Israel	ATR	524,1
deus dominus	Senhor Deus	ATR	346,2
terra Aegyptius	terra do Egito	ATR	324,3
do opera	dar-se ao trabalho	OBJ	254,2
rex Babylon	rei da Babilônia	ATR	249,0
deus immortalis	deus imortal	ATR	238,6
bellum Civilis	guerra civil	ATR	190,1

A Tabela 7, em contraste, apresenta as preferências selecionais comuns para um único lema, *do* **(dar).**

Tabela 7 – OBJ mais forte de *do* (dar). A Coluna OLD lista o verbete para o qual é dado como um uso exemplar no *Oxford Latin Dictionary*

Latim	**Português**	**OLD**	**−2logλ**
opera	serviço (= dar-se ao trabalho)	22c	254,2
obses	refém	11a	21,8
signum	sinal	-	12,6
velum	vela (= navegar)	18f	7,9
pecunia	dinheiro (= pagar)	6a	7,3
negotium	negócio	-	6,2
poena	penalidade (= sofrer)	7b	5,6
possessio	posse	1c	4,8
littera	carta (para entrega)	10a	4,3
osculum	beijo	8a	4,1
tergum	costas (= virar)	18d	3,4

[51] Em inglês no original.

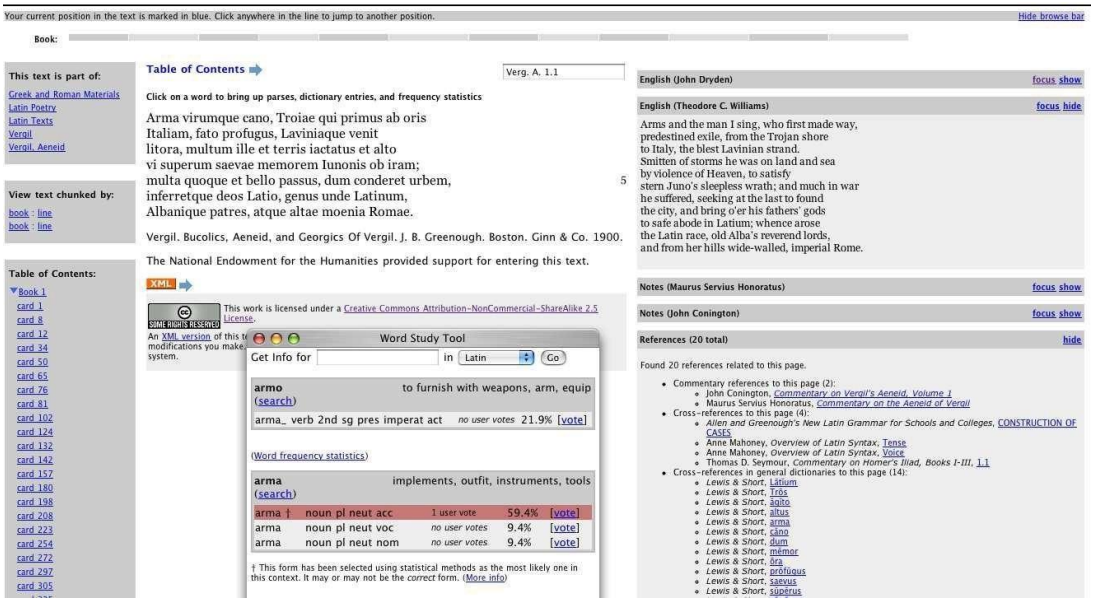
Aqui começamos a perceber a diferença entre uma expressão idiomática (*do opera*) (literalmente, "dar trabalho," idiomaticamente "empenhar-se"), com uma probabilidade de log de 254,2 e preferências selecionais com pontuação caindo abaixo do limite locucional (3,84), mas os argumentos ainda muito típicos. Podemos julgar a força destas associações comparando-as com verbetes em um léxico latino tradicional, o *Oxford Latin Dictionary* (OLD) (CLARE, 1968-1962): nisto, não estamos determinando uma precisão padrão ouro, mas um acordo entre anotadores entre o resultado automatizado e um editor humano. Para isso, 9 de seus 11 objetos mais fortes são citados como usos exemplares na coluna OLD – apenas *signum* e *negotium* são omitidos.

As preferências selecionais fortes também nos permitem distinguir entre lemas com significados semelhantes. As palavras latinas *ago* (dirigir) *gero* (gerar) e *duco* (conduzir) são frequentemente usadas em um sentido indefinido para especificar que uma ação ocorreu (*ago*, por exemplo, é a raiz latina da palavra agente em inglês). Esta abstração dá origem a discussões com significados especiais. Ao extrair as preferências selecionais desses verbos, podemos compará-los e isolar os argumentos que os distinguem uns dos outros. A Tabela 8 apresenta todos os argumentos de *ago*, *gero* e *duco* com uma probabilidade de pontuação acima de 1. Muitos desses objetos formam expressões idiomáticas com o nó (por exemplo, *ago* + *gratia*, "agradecer") e, à exceção de um, encontram-se como uso exemplar no OLD. O uso do verbo *gero* destaca, em particular, a possibilidade de agrupar ainda mais essas palavras individuais em classes maiores: três de seus objetos mais fortes são os escritórios oficiais (*praetura*, "pretoria"; *censura*, "ofício de censor" e *magistratus* "magistratura").

Tabela 8 – OBJ mais forte de *ago* (dirigir), *gero* (gerar) e *duco* (conduzir). A Coluna OLD lista o verbete para o qual é dado como um uso exemplar no *Oxford Latin Dictionary*

Latim	**Português**	**OLD**	**−2logλ**
ago			
gratia	agradecer (= dar graças)	28b	50,6
paenitentia	arrependimento (= arrepender-se)	28a	21,8
nugae	bagatela (= regatear)	22b	7,7
causa	processo (judicial)	42c	1,2
aetas	idade (= ter X anos)	31a	1,0
gero			
bellum	guerrear	8b	10,5
praetura	magistratura	10	3,2
mos	costume	8d	2,7
censura	censura	-	1,5
magistratus	magistratura	10	1,4
duco			
uxor	esposa (= casar)	5a	11,6
exercitus	exército (= marchar)	6a	1,3
pompa	desfile	7a	1,0

Figura 3 – Uma captura de tela de *Eneida* de Virgílio da Biblioteca Digital Perseu



Informações lexicais em uma biblioteca digital

A interação de uma arquitetura de biblioteca digital com este conhecimento ocorre de três maneiras: primeiro, permite contextualizar ainda mais nossos textos-fonte para os usuários da biblioteca digital existente; em segundo lugar, permite-nos apresentar relatórios personalizados para uso da palavra de acordo com os metadados associados com os textos a partir dos quais são retirados e, com isso, criamos um léxico dinâmico que não só observa como uma palavra é usada em latim em geral, mas também em qualquer autor, gênero ou era específico (ou a combinação desses). E, em terceiro lugar, permite-nos continuar a minerar mais textos quanto ao conhecimento que eles contêm à medida que são acrescentados à coleção da biblioteca, essencialmente tornando-se um serviço aberto.

Contextualização

A Figura 3 mostra uma imagem de nossa biblioteca digital existente. Nela, o leitor vê as sete primeiras linhas da *Eneida* de Virgílio. O texto fonte é fornecido no centro e as informações de contextualização preenchem a coluna da direita. Esta informação inclui:

- Traduções. Aqui são apresentadas duas traduções ao inglês; uma por John Dryden, poeta inglês do século XVII e outra mais moderna por Theodore Williams.
- Comentários. Dois comentários também são fornecidos, um em latim pelo gramático romano Sérvio, e um em inglês por John Conington, acadêmico do século XIX.
- Citações em obras de referência. As obras de referência clássicas como gramáticas e léxicos costumam citar passagens específicas em obras literárias como exemplos de uso. Aqui, todas as citações em tais obras de referência para qualquer palavra ou frase nestas sete linhas são apresentadas à direita.

Além disso, cada palavra no texto de origem está ligada à sua análise morfológica, que lista todos os lemas e características morfológicas associadas a essa forma específica da palavra.

Aqui o leitor clicou em **arma** no texto fonte. Esta ferramenta revela que a palavra pode ser derivada a partir de dois lemas (o verbo **armo** e o substantivo **arma**), e apresenta

uma análise morfológica completa para cada. Um sistema de recomendação seleciona automaticamente a análise mais provável, dado o contexto, e os usuários também podem voltar para a forma que acharem correta.

As informações de preferência selecional que temos extraído de nossa coleção é um outro método de dar mais informações contextuais a nossos usuários. Apesar de todas as palavras em um texto de origem estarem vinculadas às suas entradas lexicais por meio de sua[52] análise morfológica, pudemos oferecer uma fonte de conhecimento que complementa os léxicos de curadoria humana fornecendo também informações de frequência (e pontuações de probabilidade de log) como comprovação para a predominância de um objeto.

Criando subcorpora personalizados

Os resultados sobre o uso do verbo **fazer** apresentados acima são retirados das 3,5 milhões de palavras de nosso corpora latino. A vantagem de ter esse conhecimento em uma biblioteca digital é a estrutura imposta pela arquitetura da biblioteca. Os textos em nossa coleção têm metadados associados a eles que especificam seu autor, gênero e todas as várias coleções a que pertencem (por exemplo, a *Eneida* de Virgílio é parte das obras completas de Virgílio, que faz parte da poesia latina, que faz parte de textos em latim). Esta mesma arquitetura é preservada nos dados analisados automaticamente para que possamos consultar e apresentar informações sob medida para os autores ou gêneros específicos.

Conduzir esta mesma pesquisa em três subconjuntos de todo o nosso *corpus* – todas as obras de autoria de César, Jerônimo e Ovídio – fornece os resultados apresentados na Tabela 9. Aqui vemos claramente a importância de buscar essas seleções de todo o nosso *corpus*, pois o uso do verbo difere claramente de acordo com o gênero de cada autor. César caracteristicamente usa *do* que pode ser chamado de um sentido "militar", como com *obses* ("reféns"); Jerônimo, um padre apostólico cujas obras em latim são predominantemente compostas da Bíblia Vulgata, usa *do* para fornecer bebida, comida, descanso e glória, enquanto os objetos mais comuns em Ovídio, um poeta do amor,

[52] Todos os lemas latinos, por exemplo, estão ligados a seus verbetes no *Lewis Fundamental Latin Dictionary* e *Lewis and Short Latin Dictionary*.

incluem beijos *(osculum)* e presentes *(munus).* Observe que não precisamos simplesmente nos limitar à pesquisa por autor –podemos pesquisar por qualquer elemento dos metadados presentes nestes textos ou qualquer combinação de campos (por exemplo, toda a histórica romana escrito, exceto as obras de Tácito, além de toda a poesia elegíaca latina escrita antes da virada do milênio).

Tabela 9 – OBJ mais forte de *do* por autor individual

Latim	Português	log λ
César		
obses	refém	18,4
opera	serviço	11,9
suspicio	suspeita	2,2
facultas	faculdade	1,8
signum	sinal	1,5
Ovídio		
osculum	beijo	8,5
velum	velejar	5,9
munus	presente	3,5
signum	sinal	2,6
Jerônimo		
potus	beber	16,6
esca	comida	3,3
requies	descansar	3,0
gloria	glória	2,4
terra	terra	1,6

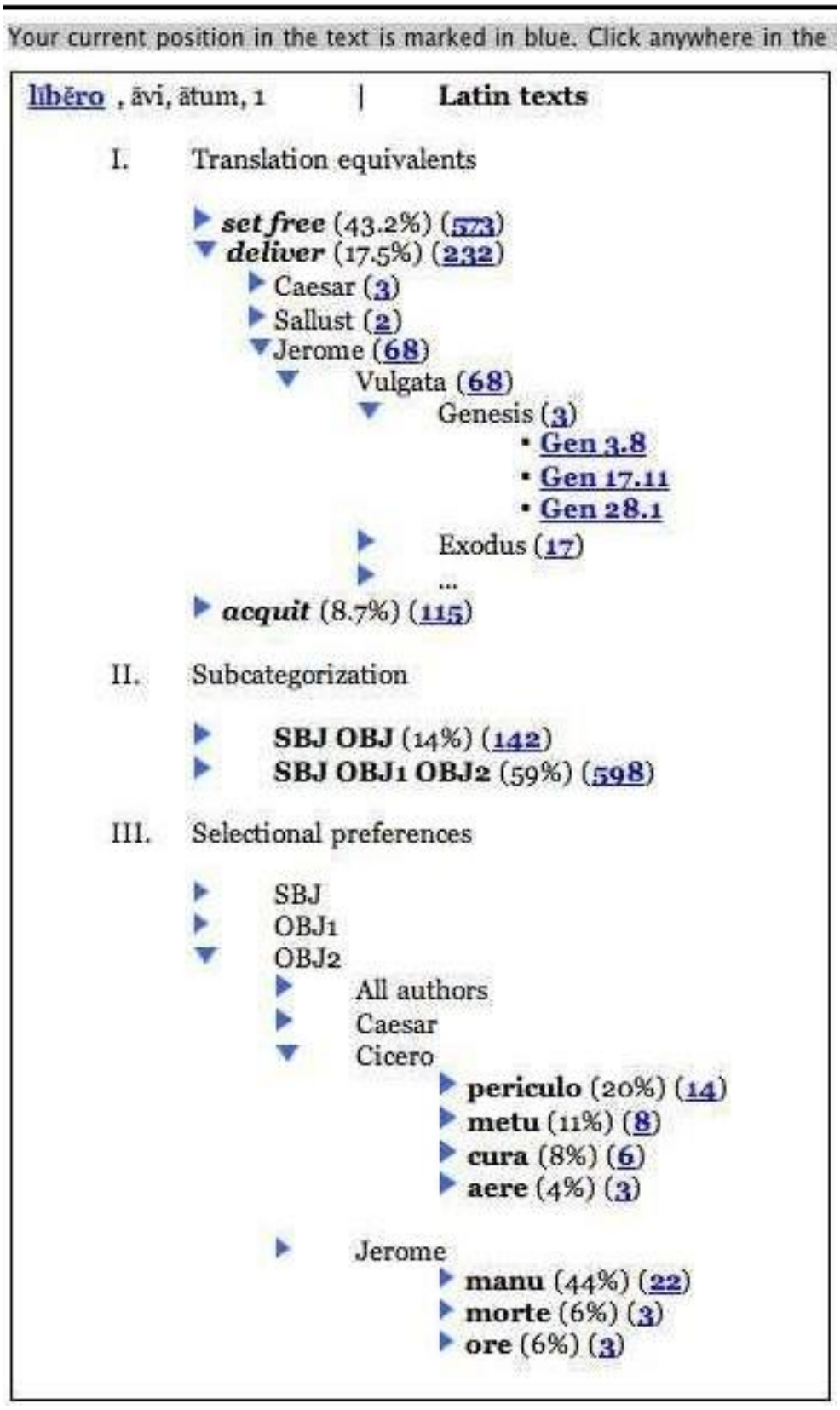
A Figura 4 apresenta um exemplo de como um verbete lexical completo ficaria no contexto de uma biblioteca digital. O verbete apresenta uma visão de alto nível da palavra e se uso em latim, porém, todas as suas categorias são divididas por subcorpora específicos, o como seu uso em autores individuais.

Coleção aberta

A Biblioteca Digital Perseu em si contém apenas um subconjunto muito pequeno de latim – suas coleções são compostas principalmente de textos da época clássica (ca. 200 a. C. a 200 d. C.) com alguns dados além deste período (*Vulgata* de Jerônimo, por exemplo, foi composta no século IV d. C.). Os textos que sobreviveram a partir deste período, geralmente, fazem parte de um cânone fixo; eles formam um conjunto fechado e são semelhantes a qualquer número de corpora linguísticos controlados que tenham entrado em vigor ao longo dos últimos 40 anos (como o *corpus* Brown (KUCERA; FRANCIS, 1967) ou o British National Corpus (LEECH *et al.*, 1994) – que apresentam um conjunto de casos de teste equilibrado e bem delineado no qual se realizam experiências repetidas, mas m seu alcance é extremamente pequeno em comparação com os volumes de textos que existem fora deles.

Enquanto a "Idade de Ouro" da literatura latina floresceu próximo à virada do milênio (abrangendo desde o século I a.C. até o século I d.C.), o latim continuou a ser uma língua produtiva nos dois milênios seguintes. Como língua franca, seu uso atravessou igualmente fronteiras nacionais e gêneros. Mesmo nos primórdios da era moderna, é a língua não apenas de importantes obras científicas como a *Astronomia Nova* de Johannes Kepler (1609) ou *Systema Naturae* de Carolus Linnaeus (1735) e tratados religiosos, como os escritos de Desiderio, Erasmo e Martinho Lutero, além de milhares de obras obscuras e anônimas. As Figuras 5, 6 e 7 apresentam três desses exemplos retirados do Google Books – um tratado matemático escrito por Robert Simson em 1735, a história religiosa escrita por Johann Friedrich Gruner em 1764 e uma dissertação de filosofia escrito em 1836 (FREYSTADT, 1832). Estas obras representam apenas três exemplos das milhares de obras latinas que estão fora do cânone controlado da era clássica, mas que ainda podem ser encontrados em bibliotecas digitais já existentes.

Figura 4 – Simulação de um verbete de vocabulário dinâmico para o verbo latino *libero* (libertar)



O volume de textos latinos que podem ser encontrados em bibliotecas digitais é de magnitude maior do que o encontrado em qualquer *corpus* controlado. Este volume se apresenta como uma oportunidade para um léxico dinâmico. Como mostrado acima, até mesmo a análise automática abaixo da média é mais do que compensada pelo volume de textos que são analisados – quanto mais dados, melhor o desempenho. Além disso, uma amostra mais ampla de latim de diferentes épocas e autores também nos permite isolar esses recursos no uso de palavras que diferencia qualquer autor – como César usa a palavra do ***fazer*** difere do uso de Ovídio e, talvez, podemos justamente imaginar que seu uso em conjunto é distinto do de um autor cuja língua mãe não seja o latim escrevendo no século XVIII. Ao analisar os textos, como tratados matemáticos do século XVIII e dissertações filosóficas do século XIX, podemos ampliar significativamente o escopo de nosso léxico.

A biblioteca digital também difere de um *corpus* controlado de textos, pois sua coleção é dinâmica enquanto um *corpus* é laboriosamente curado à mão para apresentar um equilíbrio de textos que refletem o uso atual, uma biblioteca digital está constantemente adicionando novos textos à sua coleção. Sem um *corpus* fixo de onde retirar seu conhecimento, um léxico que analisa automaticamente todo texto novo que é adicionado a uma coleção digital está sempre atualizado; simplesmente ao adicionar novos textos, mesmo que obscuros, podemos coletar informações sobre como os seus autores usam a linguagem de uma maneira que é semelhante a (ou radicalmente diferente de) os outros autores da coleção. Analisar um texto e incluindo a sua informação lexical em uma obra de referência maior é simplesmente outra maneira de contextualizá-lo.

Conclusão

A aplicação do conhecimento estruturado a coleções muito maiores, mas não estruturadas, aborda uma lacuna deixada pelos imensos esforços de digitalização de grupos como Google e da Open Content Alliance (OCA). Estes grandes projetos estão criando milhões de coleções de livros, contudo, os serviços que prestam são genéricos (por exemplo, extração de termo chave, análise de entidade nomeada, obras afins) e refletem a grande variedade de textos e línguas que eles contêm. Ao aplicar o conhecimento específico do idioma de especialistas (como codificado em nosso

treebank), somos capazes de criar serviços mais específicos para complementar os genéricos já existentes. Ao criar um vocabulário dinâmico, construído a partir da interseção de um *corpus* de 3,5 milhões de palavra com um *treebank* de 30.457 palavras, estamos enfatizando o imenso papel que até mesmo pequenas fontes de conhecimento estruturado pode ter.

No futuro pretendemos continuar a investigar o conhecimento e os serviços que podem surgir a partir dessa interação entre pequenos dados estruturados e grandes coleções não estruturadas (também usamos *treebanks*, por exemplo, para tipificar a descoberta automática de alusões nos textos) (BAMMAN; CRANE, 2008). Será importante também avaliar este léxico em seu papel final como um recurso em nossa biblioteca digital, inclusive as oportunidades existentes para melhorias para a comunidade. Os serviços morfológicos e de dicionário atualmente existentes no Perseu já permitem aos usuários a leitura de um texto para "votar" na análise morfológica ou sentido da palavra que é apropriado, dado o contexto envolvente, com melhor precisão decorrente do maior número de votos (CRANE *et al.*, 2006). Com este tipo de interação humana, é possível melhorar o recurso global anotando onde nosso sistema errou para mudarmos o foco para sua correção automática no futuro.

Figura 5 – Trecho de 1735 tratado matemático *Sectionum Conicarum Libri V* de Robert Simson, do Google Books

SECTIONUM
CONICARUM
LIBER PRIMUS.

De Parabola.

DEFINITIONES.

LIT recta linea AB, & punctum extra ipſam C, & plano in quo ſunt recta & punctum impoſitatur norma DEF, ita ut latus ipſius DE applicetur rectæ AB, alterum vero EF ſit ad eas partes ipſius AB ad quas eſt C; & extremitati F lateris EF annectatur extremitas una fili FGC ejuſdem longitudinis cum eo latere, altera vero fili extremitas in puncto C figatur; & adducatur pars fili FG ope paxilli G ad latus normæ EF, & juxta ipſum tendatur; dein moveatur normæ latus DE ſecundum rectam AB, & interea filum paxillo diſtentum ſemper lateri EF appli-

FIG. 1.

A

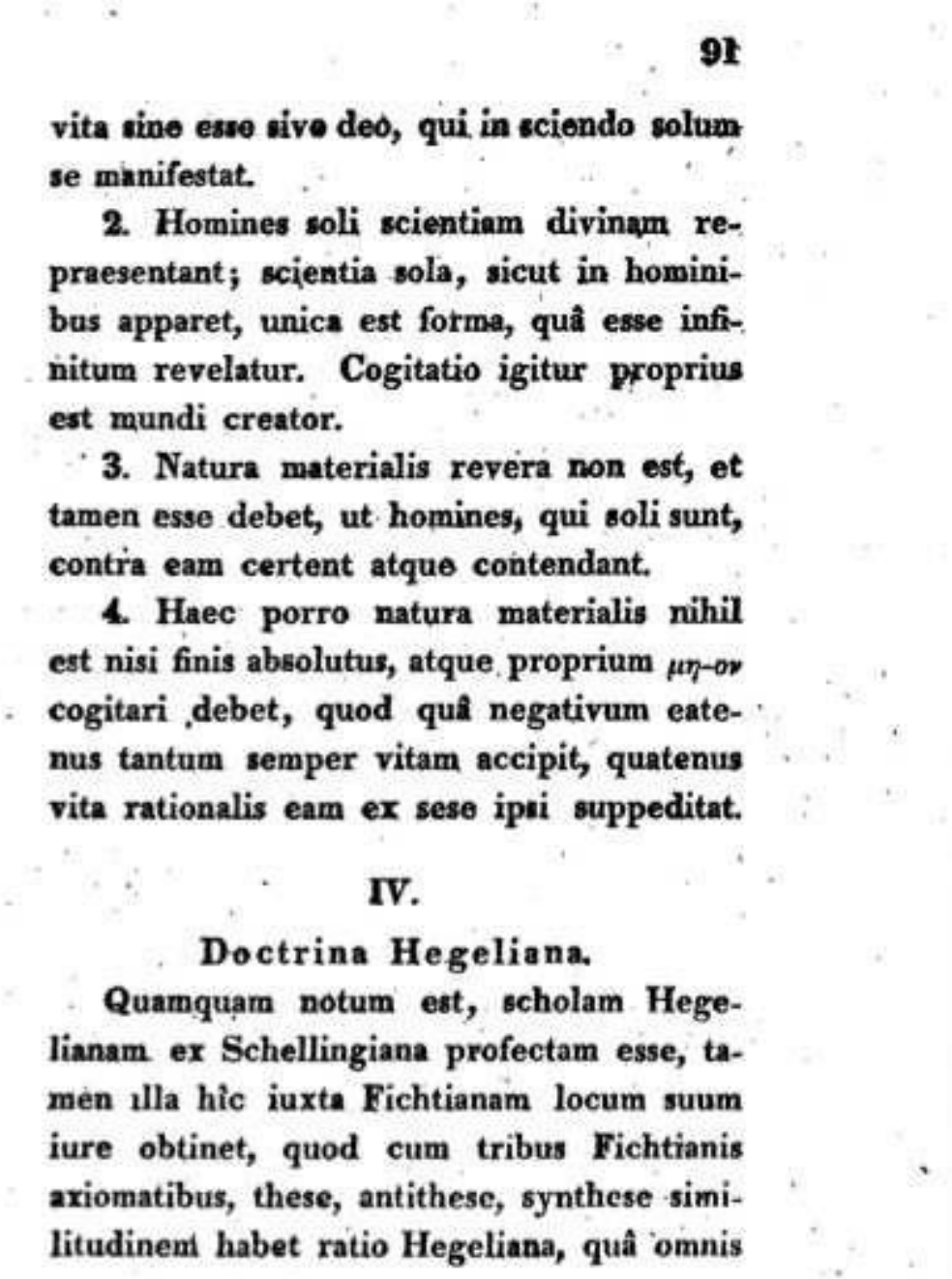
Figura 6 – Trecho da história religiosa *De origine episcoporum eorumque in ecclesia primitiva iure exercitatio,* de Johann Friedrich Gruner, 1764 do Google Books

XXIV DE ORIG. EPISCOP. EORVMQ. caet.

Deo ita ſtudeam, ut fidem certe, diligentiam atque induſtriam deſiderari in me nunquam patiar. Eum animum meum, academiae huic, eiusque Ciuium Ornatiſſimorum commodis plane deuotum, ne uerbis tantum oſtendere uidear; conabor re et opere, quantum quidem praeſentis temporis ratio permittit, declarare. Igitur, quod Deus ter Optimus Maximus felix fauſtumque eſſe iubeat, nouum munus *praelectionibus exegeticis in epiſtolam* PAVLI *Apoſtoli ad Romanos* auspicabor. Quo quidem in labore ita uerſari ſtudebo, ut ſenſu uerborum legitime indagato, uia Auditoribus Lectiſſimis ad rerum, quae diuino illo libro continentur, intelligentiam muniatur; memor magni olim uiri, PHIL. MELANCHTHONIS effati, „non poſſe „ ſcripturam intelligi theologice, niſi intellecta ante ſit grammatice.„ Neque uero iis deero, qui *diſputandi exercitio* acuere ingenium uelint. Faxit Deus, ut his meis, tenuibus licet, conatibus nominis ipſius gloria illuſtretur! Scribebam in Regia Fridericiana,

a. d. XXI. m. Nouembris, cIↃIↃCCLXIV.

Figura 7 – Trecho da dissertação filosófica *Philosophia cabbalistica et pantheismus,* de M. Freystadt, 1832, do Google Books. Observe os neologismos inventados a partir dos nomes dos filósofos alemães do século 19 (hegeliana, schellingiana e fichtiana)



91

vita sine esse sive deo, qui in sciendo solum se manifestat.

2. Homines soli scientiam divinam repraesentant; scientia sola, sicut in hominibus apparet, unica est forma, quâ esse infinitum revelatur. Cogitatio igitur proprius est mundi creator.

3. Natura materialis revera non est, et tamen esse debet, ut homines, qui soli sunt, contra eam certent atque contendant.

4. Haec porro natura materialis nihil est nisi finis absolutus, atque proprium μη-ον cogitari debet, quod quâ negativum eatenus tantum semper vitam accipit, quatenus vita rationalis eam ex sese ipsi suppeditat.

IV.

Doctrina Hegeliana.

Quamquam notum est, scholam Hegelianam ex Schellingiana profectam esse, tamen illa hîc iuxta Fichtianam locum suum iure obtinet, quod cum tribus Fichtianis axiomatibus, these, antithese, synthese similitudinem habet ratio Hegeliana, quâ omnis

Além disso, uma vez que o léxico é construído a partir de tecnologias modulares, ele se beneficia de quaisquer melhorias nesses serviços individuais (como a etiquetagem morfológica ou a análise sintática) e, desde que a etiquetagem e a precisão de análise geralmente dependem do tamanho de seu *corpus* de treinamento, esperamos melhorias

adicionais com o crescimento de nosso *treebank*. Atualmente, estamos no processo de adicionar Petrônio (um autor de prosa latina tardia) e também vários textos de Ovídio e Propércio (ambos poetas da Idade de Ouro).

A obra descrita até o presente também se concentra exclusivamente no latim, mas os textos na Biblioteca Digital do Perseu contêm uma coleção muito maior de grego (4,9 milhões de palavras). Nosso objetivo no desenvolvimento deste trabalho é criar uma arquitetura que pode ser facilmente aplicada a ambas as línguas – tudo o que precisamos para extrair preferências selecionais para grego é um grande *treebank* o suficiente para treinar um analisador estatístico e estamos na fase inicial estágios de desenvolvimento. De fato, as tecnologias descritas acima não são específicas para uma língua ou mesmo uma biblioteca: elas simplesmente dependem de uma pequena fonte de conhecimento estruturado e uma grande coleção textual. Como provam as bibliotecas de um milhão de livros, começam a surgir grandes coleções textuais em diversos idiomas; contudo, ainda perduram as fontes de conhecimento que podem ser criadas apenas por profissionais da área.

Agradecimentos

Fundos da Digital Library Initiative Phase 2 (IIS-9817484), National Science Foundation (BCS-0616521) Andrew W. Mellon Foundation (#40700635) financiaram este trabalho. Agradecemos também a Meg Luthin e Skylar Neil por sua inestimável assistência nas pesquisas.

Referências

BAMMAN, D.; CRANE, G. The design and use of a Latin dependency treebank. In: **Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories** (TLT2006), Prague, p. 67-78, 2006. ÚFAL MFF UK.

_____. The Latin Dependency Treebank in a cultural heritage digital library. In: **Proceedings of the Workshop on Language Technology for Cultural Heritage Data** (LaTeCH 2007). Prague: Association for Computational Linguistics, 2007. p. 33-40.

_____. The logic and discovery of textual allusion. In: **Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data** (LaTeCH 2008).

BAMMAN, D.; PASSAROTTI, M.; CRANE, G.; RAYNAUD, S. A collaborative model of treebank development. In: **Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories** (TLT2007), Bergen, p. 1-6, 2007a.

_____. **Guidelines for the syntactic annotation of Latin treebanks**. version 1.3. Technical report. Medford: Tufts Digital Library, 2007b.

BRANTS, T.; FRANZ, A. **Web 1T 5-gram Version 1**. Philadelphia: Linguistic Data Consortium, 2006.

BUSA, R. **Index Thomisticus: sancti Thomae Aquinatis operum omnium indices et concordantiae, in quibus verborum omnium et singulorum formae et lemmata cum suis frequentiis et contextibus variis modis referuntur quaeque / consociata plurium opera atque electronico IBM automato usus digessit Robertus Busa SI**. Frommann-Holzboog, Stuttgart-Bad Cannstatt, 1974-1980.

CHURCH, K. W.; HANKS, P. Word association norms, mutual information, and lexicography. In: **Proceedings of the 27th annual meeting on Association for Computational Linguistics**. Morristown: Association for Computational Linguistics, 1989. p. 76-83.

CHURCH, K. W.; HOVY, E. H. Good applications for crummy machine translation. **Machine Translation**, v. 8, n. 4, p. 239-258, 1993.

CLARK, S.; WEIR, D. Class-based probability estimation using a semantic hierarchy. **Computational Linguistics**, v. 28, n. 2, p. 187-206, 2002.

COLLINS, M.; RAMSHAW, L.; HAJIC, J.; TILLMANN, C. A statistical parser for Czech. In: **Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL).** Morristown: Association for Computational Linguistics, 1999. p. 505-512.

CRANE, G. From the old to the new: Integrating hypertext into traditional scholarship. In: **Hypertext'87: Proceedings of the 1st ACM conference on Hypertext.** ACM Press, 1987. p. 51-56.

CRANE, G.; BAMMAN, D.; CERRATO, L.; JONES, A.; MIMNO, D. M.; PACKEL, A.; SCULLEY, D.; WEAVER, G. Beyond digital incunabula: Modeling the next generation of digital libraries. In: GONZALO, J.; THANOS, C.; VERDEJO, M. F.; CARRASCO, R. C. (Ed.). **ECDL**, Springer, v. 4172 of Lecture Notes in Computer Science, p. 353-366, 2006.

CRANE, G.; CHAVEZ, R. F.; MAHONEY, A.; MILBANK, T. L.; RYDBERG-COX, J. A.; SMITH, D. A.; WULFMAN; C. E. Drudgery and deep thought. **Communications of the ACM,** v. 44, n. 5, p. 34-40, 2001.

DUNNING, T. Accurate methods for the statistics of surprise and coincidence. **Computational Linguistics**, v. 19, p. 61-74, 1993.

FREYSTADT, M. **Philosophia cabbalistica et pantheismus**. Regimontii Prussorum, Borntraeger, 1832.

GLARE, P. G. W. (Ed.). **Oxford Latin Dictionary**. Oxford: Oxford University Press, 1968-1982.

GRUNER, J. F. **De origine episcoporum eorumque in Ecclesia primitiva iure exercitatio. Litteris Grunertianis**, Halae Magdeburgicae, 1764.

HAJIC, J. Building a syntactically annotated corpus: The Prague Dependency Treebank. In: HAJICOVÁ, E. (Ed.). **Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová**. Prague: Karolinum, Charles University Press, 1998. p. 12-19.

HAJIC, J.; HLADKÁ, B. Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset. In: **COLING-ACL**, p. 483-490, 1998.

HAJIC, J.; PANEVOVÁ, J.; BURÁANOVÁA, E.; URESOVÁA, Z.; BÉMOVÁ, A. Annotations at analytical level: Instructions for annotators (English translation by Z. Kirschner). **Technical report**, ÚFAL MFF UK, Prague, Czech Republic, 1999.

KILGARRIFF, A.; RYCHLY, P.; SMRZ, P.; TUGWELL; D. The sketch engine. In: **Proceedings of the Eleventh EURALEX International Congress**, p. 105-116, 2004.

KLOSA, A.; SCHNÖRCH, U.; STORJOHANN, P. ELEXIKO – a lexical and lexicological, corpus-based hypertext information system at the Institut für deutsche Sprache, Mannheim. In: **Proceedings of the 12th Euralex International Congress**, 2006.

KUCERA, H.; FRANCIS, W. N. *C***omputational Analysis of Present-Day American English**. Providence: Brown University Press, 1967.

LEECH, G.; GARSIDE, R.; BRYANT, M. CLAWS4: the tagging of the British National Corpus. In: **Proceedings of the 15th conference on Computational Linguistics.** Morristown: Association for Computational Linguistics, 1994. p. 622-628.
LEWIS, C. T. (Ed.). **An Elementary Latin Dictionary**. Oxford: Clarendon Press, 1891.

LEWIS, C. T.; SHORT, C. (Ed.). **A Latin Dictionary**. Oxford: Clarendon Press, 1879.

MARCUS, M. P.; SANTORINI, B.; MARCINKIEWICZ; M. A. Building a large annotated corpus of english: The penn treebank. **Computational Linguistics**, v. 19, n. 2, p. 313-330, 1994.

MCDONALD, R.; PEREIRA, F.; RIBAROV, K.; HAJIC, J. Non-projective dependency parsing using spanning tree algorithms. In: **Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing**, p. 523-530, 2005.

NIVRE, J. Constraints on non projective dependency parsing. In: **EACL**. The Association for Computer Linguistics, 2006.

NIVRE, J.; HALL, J.; NILSSON, J.; CHANEV, A.; ERYIGIT, G.; KU¨BLER, S.; MARINOV, S.; MARSI, E. Maltparser: A language-independent system for data-driven dependency parsing. **Natural Language Engineering,** v. 13, n. 2, p. 95-135, 2007.

PASSAROTTI, M. Verso il Lessico Tomistico Biculturale. La treebank dell'Index Thomisticus. In: RAFFAELLA, P.; DIEGO, F. (Ed.). **Il filo del discorso. Intrecci testuali, articolazioni linguistiche, composizioni logiche. Atti del XIII Congresso Nazionale della Società di Filosofia del Linguaggio, Viterbo,** set. 2006, p. 187-205. Roma: Aracne Editrice, Pubblicazioni della Società di Filosofia del Linguaggio, 2007.

PINKSTER, H. **Latin Syntax and Semantics**. London: Routledge, 1990.

RATNAPARKHI, A. A maximum entropy model for part-of-speech tagging. In: BRILL, E.; CHURCH, K. (Ed.). **Proceedings of the Conference on Empirical Methods in Natural Language Processing.** Somerset, New Jersey: Association for Computational Linguistics, 1996. p. 133-142.

ROOTH, M.; RIEZLER, S.; PRESCHER, D.; CARROLL, G.; FRANZ, B. Inducing a semantically annotated lexicon via EM-based clustering. In: **Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics.** Morristown: Association for Computational Linguistics, 1999. p. 104-111.

SCHMID, H. Probabilistic part-of-speech tagging using decision trees. In: **International Conference on New Methods in Language Processing**, Manchester, 1994.

SIMSON, R. **Sectionum Conicarum Libri.** Edinburgh: V. T. and W. Ruddimannos, 1735.

SINCLAIR, J. M. (Ed.). **Looking Up: an account of the COBUILD project in lexical computing.** Collins, 1987.

TEXTOS FRAGMENTÁRIOS E BIBLIOTECAS DIGITAIS[53]

Monica Berti[54]

Introdução

O objetivo deste artigo é descrever um novo modelo para a representação de textos fragmentários em uma biblioteca digital de fontes clássicas. Um fragmento é a peça sobrevivente de algo irremediavelmente perdido ou nunca terminado. Neste sentido, a palavra é aplicada a uma grande variedade de material de restos de provas antigas, como ruínas monumentais, cacos cerâmicos, pedaços de papiros ou inscrições quebradas[55]. Os limites destes fragmentos são marcados por margens, cuja materialidade chama a atenção para a exterioridade da prova, influenciando nossa reconstrução da totalidade a que o fragmento pertencia e nossa percepção das razões da sua fragmentação, geralmente devido a um evento violento externo como destruição ou consumo. Se um fragmento desse tipo tem evidência textual, a materialidade do fragmento se estende também para o texto, que passa a ser uma peça quebrada de uma escrita antiga[56].

Quanto à evidência textual, há também uma outra categoria de fragmentos, que se refere a um fenômeno completamente diferente, porque estes trechos não são parte de um todo original maior, mas o resultado de um trabalho de interpretação realizado por estudiosos, que extraem e coletam informações relativas a obras perdidas embutidas em outros textos sobreviventes. Estes fragmentos incluem uma grande variedade de formatos que vão desde citações literais a alusões vagas, mas eles são apenas uma imagem mais ou menos sombria do original de acordo com a distância maior ou menor de uma citação

[53] Texto originalmente publicado no *Workshop on Historical Texts* ocorrido na Tufts University em 13-14 de janeiro de 2010 (http://www.linguistik.hu-berlin.de/en/institut-en/professuren-en/korpuslinguistik/events-en/nehdfg). Em português, distribuído sem fins comerciais, sob permissão da autora.

[54] Alexander von Humboldt-Lehrstuhl für Digital Humanitites Institut für Informatik – Universität Leipzig – DE – 04109 – Leipzig – Deutschland – monica.berti@uni-leipzig.de.

[55] Para uma definição do termo, ver OED2, VI, fragmento sv. Os principais conceitos que expressam o significado do termo fragmento estão também representados por synsets (conjuntos de sinônimos cognitivos) em WordNet, que é uma base de dados lexical para o idioma inglês (Disponível em: http://wordnet.princeton.edu/).

[56] Gumbrecht 1997, p. 320. Entre os muitos exemplos deste tipo de evidência "fragmentada", veja o *Marmor Parium* e o *Hellenica Oxyrhynchia.*

literal. Este uso do termo fragmento pode ser enganoso, porque o texto original do trecho é geralmente coberto pelo contexto de transmissão e distorcido pelo estilo e propósito do autor que o extraiu e citou (geralmente chamado de "testemunha" do fragmento[57]). Além disso, as citações literais podem estar incorretas e, especialmente, no caso da prosa, pode ser muito difícil distinguir as citações literais de paráfrases ou resumos, uma vez que o sentido original do texto pode ser alterado por omissões, deformações ou razões polêmicas[58].

A coleção de fragmentos impressos compreende trechos textuais tirados de muitas fontes diferentes e organizadas de acordo com vários critérios, como a ordem cronológica ou disposição temática. O comprimento destes trechos pode ser significativamente diferente de uma edição para outra e depende da opção do editor: ele pode decidir publicar um extrato maior ou menor se ele atribui uma parcela maior ou menor do texto incorporado ao fragmento ou se ele quer para proporcionar ao leitor o contexto mais longo possível para lhe dar uma melhor compreensão da citação preservada na mesma[59]. Em qualquer caso, quando uma parte extraída do texto é impressa, ela imediatamente adquire uma espécie de materialidade, devido à sua representação tipográfica: tem margens bem definidas, como um fragmento de verdade, mas na verdade é o resultado de uma extração e interpretação modernas; ela pode criar falsas ilusões, pois o fragmento em si não existe e é apenas uma sombra, cuja forma é turva e pode levar a uma percepção distorcida da realidade[60]. No entanto, a obtenção de fragmentos é uma tradição bem estabelecida e os grandes feitos de estudiosos desde o Renascimento têm permitido redescobrir e preservar um patrimônio cultural inestimável que, de outra forma, estaria perdido e esquecido[61]. Ao mesmo tempo, a procura de restos de obras perdidas é um exercício metodológico muito útil para a prática de reconstrução de testemunhos antigos e também é um estímulo para a interdisciplinaridade, uma vez que um editor tem que enfrentar uma série de problemas decorrentes da grande variedade de assuntos e diferentes tipos de textos que geralmente

[57] Schepens (1997a, p. 166; 2000, p. 4-13).
[58] Brunt (1980, p. 478, 482); Bowersock (1997, p. 174); Lenfant (2007a, p. 47, 53-63); Bamman-Crane (2008b, p. 2).
[59] Compare, por exemplo, FHG I 54 fr. [73 com FGrH 323a F 14. Para uma explicação sobre essas abreviações, ver nota 10.
[60] Brunt (1980, p. 477).
[61] Dionisotti (1997). Sobre a importância de textos fragmentários para o nosso conhecimento da literatura antiga, consulte Strasburger (1977, p. 922); Schepens (1997a, p. 14445). Cf. Também Berti *et al.* (2009, p. 259). Dionisotti de (1997, p. 27).

formam uma coleção de fragmentos[62].

Uma das principais preocupações ao se levantar evidências de trabalhos perdidos é reconstruir a complexa relação entre o fragmento e sua fonte de transmissão, o que significa pesar o nível de interferência desempenhado pelo autor que reutilizou e transformou o contexto original do fragmento, medindo, portanto, a distância entre o texto original e o texto derivado, e tentar perceber o grau de reutilização de texto e seus efeitos sobre o texto de destino resultante[63]. Este processo interpretativo é geralmente explicado no comentário de uma edição de fragmento ou em obras e monografias relacionadas a vários aspectos de autores e obras fragmentárias, mas é completamente perdido na representação da impressão dos fragmentos, que são reproduções simplesmente tipográficas de extratos de derivados textos.

Nosso objetivo é repensar a questão fundamental da relação entre o fragmento e sua testemunha, proporcionando um novo modelo para a representação de fontes antigas, baseadas em tecnologias de informação que permitem a construção de acervos digitais concebidos não só para preservar, mas também estender as ontologias que o estudo tradicional desenvolveu ao longo de gerações, e ao mesmo tempo representar cada elemento das convenções impressas de uma forma mais dinâmica e interligada Mesmo que muitos gêneros diferentes de textos fragmentários tenham sido preservados, nossas observações centram-se em historiadores fragmentários gregos porque, em muitos aspectos, eles podem ser considerados representativos para a construção de um acervo digital de autores fragmentários. Além disso, as coleções monumentais de fragmentos históricos gregos publicados nos dois últimos séculos deram origem a questões fundamentais na coleta e edição de fragmentos: as bibliotecas digitais emergentes de fontes clássicas nos desafiam a repensar tais questões e as características dos fragmentos textuais[64].

Antes de abordar estas questões, gostaríamos de discutir alguns pontos referentes

[62] Dionisotti de (1997, p. 27).
[63] Lee (2007, p. 472).
[64] As duas obras de referência fundamentais para todas as pesquisas sobre a historiografia grega fragmentária são FHG *(Fragmenta Historicorum Graecorum* IV, coll K. e Th. Müller. Parisiis 1841-1884) e FGrH *(Die Fragmente der Griechischen Historiker* I-III, v F. Jacoby. Berlin - Leiden 1923-1958). Quanto ao projeto internacional com o objetivo de publicar as seções que Jacoby planejou, mas nuncapublicou, consulte Schepens 1997a e 1998. Para outros projetos recentes que foram realizados para atualizar a obra de Jacoby e torná-la mais útil e acessível, consulte Marincola (2000 e 2005); Worthington (2005); Lanzillotta (2006).

às novas possibilidades que estão sendo oferecidas pelas tecnologias digitais para criar uma nova infraestrutura para estudos clássicos, onde o objetivo é fornecer uma ampla gama de serviços para representar e estudar antigas fontes de uma forma que não era viável na cultura impressa[65]. Esses serviços fazem parte de uma Infraestrutura Cibernética em Ciências Humanas e Sociais Humanidades e Ciências Sociais que foi proposta pelo *American Council of Learned Societies*, não apenas para difundir inovações tecnológicas no mundo acadêmico, mas para desenvolver novos modelos, ferramentas e padrões para a representação de novas edições digitais de textos e para construir uma cultura digital "cumulativa, colaborativa e sinérgica[66]".

As tecnologias da informação estão possibilitando publicar grandes quantidades de recursos referentes a estudos clássicos na *web*. Eles são apresentados em uma grande variedade de formatos e incluem vários tipos de fontes, como textos literários, epígrafes, papiros e imagens de evidências arqueológicas, além de fornecer traduções e transcrições de textos, comentários, artigos, livros, enciclopédias, atlas, bases de dados bibliográficas e registros de áudio de conferências e workshops sobre todos os aspectos do mundo antigo. Os fenômenos de informação mais impressionantes, no entanto, que estão afetando o trabalho dos clássicos são os projetos de digitalização em massa de Google Books e do Internet Archive, o projeto Wikipédia e blogs de grupos de discussão eletrônicos. A cada dia um número crescente de textos digitalizados aparece no Google Books: mesmo que a maioria deles seja vista apenas em um trecho ou pré-visualização limitada, existem muitos textos não protegidos por direitos autorais que são totalmente visíveis. Grandes coleções destes documentos pertencem ao campo dos estudos clássicos e, assim, permitem que os estudiosos consultem e baixem muitas edições críticas publicadas nos séculos XIX e XX. Desnecessário dizer que este é o primeiro passo em direção a uma contribuição extraordinária para a preservação de um patrimônio inestimável de erudição mais antiga que, muitas vezes, é ignorado, não só por ser considerado antigo e fora de moda, mas também porque, em muitos casos, são difíceis de

[65] Para o recente debate sobre as novas perspectivas oferecidas nas áreas dos chamados eClassics e ePhilology, consulte Crane-Bamman (2007); Crane *et al.* (1991, 2006; 2007; 2009a, 2009b); Crane-Seales-Terras de (2009); Blackwell-Crane (2009); Bodard-Mahony (2010); Crane (2010); Numerico-Fiormonte-Tomasi (2010). Para um panorama da história da computação em clássicos e uma pesquisa profunda das bibliotecas digitais clássicas multidisciplinares e os requisitos para a infraestrutura cibernética para clássicos digitais, consulte Babeu (2010).

[66] Welshons (2006). Para um exemplo de aplicação destes princípios aos estudos clássicos, ver Pritchard (2008).

localizar e consultar em bibliotecas tradicionais[67].

Por outro lado, o crescente sucesso da Wikipédia e dos *blogs* incentiva todos os especialistas a questionar o futuro da produção e difusão do conhecimento. Classicistas estão enfrentando um novo mundo onde os temas da pesquisa são compartilhados por muitas pessoas; eles têm a responsabilidade de prestar atenção a esses fenômenos de modo a preservar o patrimônio cultural e proporcionar as melhores fontes de informação e discussão possíveis. Ao mesmo tempo, o livre acesso aos recursos da *web* apresenta uma oportunidade extraordinária aos classicistas para participar do debate cultural e revitalizar o papel de estudos clássicos na civilização moderna. Por outro lado, a *web* está ajudando a criar novos modelos de colaboração acadêmica, rompendo o forte individualismo que caracteriza a investigação na área das humanidades e contribuindo para a troca de sinergia entre as diferentes especializações, como regularmente acontece nas ciências[68].

No entanto, mesmo se os dados publicados na *web* estão aumentando constantemente, nossa capacidade de absorver e processá-los é relativamente constante e enfrentamos o risco de sobrecarga de informações, além de perder a capacidade de localizar e selecionar material útil e de alta qualidade para nosso trabalho. Os motores de busca têm suas limitações, pois permitem apenas procurar palavras específicas em documentos e não as relações entre eles e seus conteúdos. Este problema surgiu desde o nascimento da *web* e originou a evolução da chamada Web Semântica: seu objetivo é desenvolver métodos e linguagens para descrever a semântica de documentos e recursos na *web* para organizá-los e expressar suas relações, com o intuito de ampliar as capacidades sintáticas e semânticas da *web*, integração e combinação de dados provenientes de diferentes 16 fontes, o que significa compartilhar conceitos e não apenas palavras-chave[69].

Estes esforços estão se movendo para o desenvolvimento de um novo campo interdisciplinar chamado Ciência da Web: seu objetivo é reunir especialistas de cada ramo

[67] Em novembro de 2008, a Comunidade Europeia lançou um protótipo do projeto Europeana cujo objetivo é coletar conteúdos digitais disponíveis em museus, bibliotecas, arquivos e coleções audiovisuais da Europa: http://www.europeana.eu/.

[68] Hardwick (2000); McManus-Rubino (2003); Rosenzweig (2006). Para uma discussão anterior sobre o papel do computador na pesquisa clássica, consulte Ireland (1976); Bolter (1984 e 1991) (cf. também 2012); Wright (1994). Há um grande debate sobre se a erudição deve ser acessível a todos e gratuita: Rosenzweig (2005); Willinsky (2005).

[69] Sobre a evolução da *Web* Semântica, consulte Berners-Lee, Hendler, Lassila (2001); Shadbolt *et al.* (2006).

do conhecimento para estudar a *web* e lidar com seus desafios técnicos e sociais e, posteriormente, oferecer soluções para a modelagem da *World Wide Web* e compreensão do seu impacto social[70]. Os estudiosos têm o dever de participar deste debate porque precisam garantir a construção de um ambiente de informação acadêmica que atenda adequadamente às suas necessidades de produção e divulgação dos resultados das pesquisas. Os classicistas podem derivar benefícios duradouros e fazer uma contribuição eficaz ao participar desta iniciativa. Dado o alto nível de especialização exigido por muitos campos referentes a estudos clássicos, os estudiosos não podem simplesmente ignorar a interdisciplinaridade e novas tecnologias, se quiserem preservar o passado[71]. Ao mesmo tempo, podem também contribuir para o desenvolvimento de linguagens e métodos para armazenar e organizar recursos na rede, tal como mostrado, por exemplo, pelo fato de que muitos conceitos lógicos utilizados na Web Semântica derivam do pensamento e da tradição filosófica antigos[72].

Representando Fragmentos

Na segunda metade do século XX, as novas tecnologias vêm produzindo cada vez mais ferramentas informatizadas que foram personalizadas para coleta e digitalização de textos antigos, levando à formação de coleções digitais de todas as 20 grandes fontes clássicas[73]. Com o desenvolvimento destas ferramentas e a contínua expansão de repositórios completos, os classicistas lidam com casos textuais complexos como representar fragmentos de uma biblioteca digital, lançando as bases não apenas para a construção de uma nova geração de coleções fragmentárias que expressam toda a complexidade da erudição clássica, mas também oferecendo uma representação visual mais precisa e dinâmica de fragmentos de texto, elaborando uma estrutura e uma interface completamente diferente de coleções produzidas na cultura impressa. As edições

[70] Hendler *et al.* (2008).
[71] Epigrafia e papirologia está obtendo muitos benefícios de novas tecnologias: ver Cayless *et al.* (2009); Bodard (2010).
[72] Parodi-Ferrara (2002); Benjamins *et al.* (2004).
[73] Nas últimas décadas, o foco da pesquisa conduzida por classicistas está em proporcionar corpora grandes de fontes antigas e em desenvolver a marcação semântica de determinados textos, como fontes gregas e romanas em TEI XML Biblioteca Digital Perseus e documentos epigráficos codificados por EpiDoc, que é uma extensão das Diretrizes TEI (BURNARD-BAUMAN, 2009) para representar textos documentais preservados em pedras, aplicáveis também a outros campos, como papirologia e numismática (ver Cayless *et al.*, 2009). Cf. também Ciotti Albonico (2005).

impressas de fragmentos contêm extratos de muitas fontes diferentes e são, portanto, representações em papel de hipertextos[74]. Com a crescente disponibilização em formato digital das edições fonte a partir que vêm sendo disponibilizadas, é possível construir as edições que sejam verdadeiramente hipertextuais, incluindo não apenas trechos, mas links para as fontes acadêmicas de onde os excertos são extraídos[75].

A construção de um *corpus* digital de autores fragmentários significa lidar com o problema de codificar e representar o texto e a estrutura de um fragmento. É de aceitação geral que a representação digital das características internas e externas de um texto não se limita apenas a um processo simples de reprodução e mecânica, mas de um ato interpretativo[76]. Assim, codificar fragmentos é, antes de tudo, o resultado de interpretá-los, desenvolvendo uma linguagem adequada para a representação de todos os elementos de suas características textuais, criando, assim, meta-informação através de uma marcação semântica precisa e sofisticada. Portanto, editar fragmentos significa produzir meta-edições que são diferentes das impressas porque não são apenas citações isoladas, mas também ponteiros para os contextos originais de onde os fragmentos foram extraídos. Enquanto os editores devem ser capazes de definir as seções de texto precisas que acreditam ser relevantes e poder anotá-las de várias formas (por exemplo, distinguir o que consideram ser paráfrase de citação direta), estes fragmentos também devem ser dinamicamente vinculados aos seus contextos originais e informações atualizadas para contextualização. Em um nível mais amplo, o objetivo de uma edição digital de fragmentos é representar várias relações transtextuais, como definidas na crítica literária, que incluem a intertextualidade (a presença de um texto dentro de outro texto, como citações, alusões e plágio), paratextualidade (ou seja, todos os elementos que não fazem parte do texto, como títulos, subtítulos, prefácios, notas, etc.), metatextualidade (relações críticas entre os textos, ou seja, comentários e textos críticos), arquitextualidade (o que significa a qualidade genérica e status de um texto) e hipertextualidade (ou seja, a derivação de um texto de um hipotexto preexistente através de um processo de transformação ou imitação[77]).

Conceber uma edição digital de fragmentos também significa encontrar

[74] Sobre a definição de hipertexto na computação, consulte Landow (2006).
[75] Sobre o impacto do hipertexto no estudo clássico, ver Crane (1987).
[76] Cf. Fiormonte (2003, p. 163-172); Ciotti (2005).
[77] Para estes conceitos, consulte Genette (1997, p. 1-7). Sobre intertextualidade, consulte também Polacco (1998).

paradigmas digitais e soluções para expressar informações sobre edições críticas impressas e suas características editoriais e convencionais. Trabalhar em uma edição digital significa converter as ferramentas tradicionais e recursos utilizados por estudiosos como referências canônicas, tabelas de concordâncias e índices em conteúdos para processo automatizado.

Para mostrar como um fragmento deve ser representado em uma biblioteca digital, consideremos um exemplo complexo constituído por uma série de referências fragmentárias incorporadas em uma longa seção da *Vida de Teseu* de Plutarco (24-28), que pertence à unificação da Ática e início da democracia, a anexação do território de Mégara à Ática, a instituição dos jogos ístmicos e a guerra contra as amazonas (ver Anexo 1). Nestes capítulos, Plutarco menciona diversos tipos de fontes: 1) três oráculos[78]; 2) o texto de uma inscrição[79]; 3) autores preservados, como Aristóteles, Homero, o próprio Plutarco e Píndaro[80]; 4) uma série de historiadores fragmentários, como Helânico, Ândron de Halicarnasso, Filócoro, Ferecides, Herodoro, Bíon, Menécrates, Clidemo e o autor da *Teseida*[81].

Estas citações foram recolhidas em muitas coleções diferentes de textos fragmentários. Em particular, o texto de Plutarco foi dividido pelos irmãos Müller e por Jacoby em extratos espalhados e repetidos nas seções de suas coleções de 29 fragmentos históricos gregos correspondentes aos autores citados por Plutarco[82]. Desse modo, o resultado da representação impressa destes fragmentos é que o mesmo texto da *Vida de Teseu* não só é interrompido em diversos fragmentos, como também se repete tantas vezes quantos são os autores citados por Plutarco. Além disso, dado que não é possível identificar os limites das citações de Plutarco, os editores adotaram critérios diferentes e os mesmos fragmentos podem ter comprimentos e divisões diferentes de uma edição para

[78] Dois oráculos de Delfos (24 = Parke-Wormell 2.154; 26.4 = Parke-Wormell 2.411), um oráculo da Sibila (24.5 = Hendess 23).
[79] O pilar no Istmo (25,3).
[80] Aristóteles (25.2 = *Constituição dos Atenienses* 41.2; F 384 Rose3); Homero (25.2 = *Ilias* 2.547); próprio Plutarco (27.6 = *Vida de Demóstenes* 19.2); Píndaro (28.2 = F 176 Sn.-Mae).
[81] Helânico (25.5 = FHG I 55 fr. 76 = FGrH 4 F 165 = FGrH 323a F 15; 26,1 = FHG I 55 fr. 76 = FGrH 4 F 166 = FGrH 323a F 16a; 27,2 = FGrH 4 F167a = 323a F 17); Ândron (25,5 = FGrH 10 F 6); Filócoro (26,1 = FHG I 392 fr. 49 = FGrH 328 F 110;); Ferecides (26.1 = FGrH 3 F 151); Herodoro (26.1 = FGrH 31 F 25a); Bíon (26.2 = FHG II 19 fr. 1 = FGrH 14 F 2 = FGrH 332 F 2); Menécrates (26.2 = FHG II 345 fr. 8 = FGrH 701 F 1); Clidemo (27.3 = FHG I 360 fr. 6 = FGrH 323 F 18); o autor da *Teseida* (28.1 = EGF 217 Kinkel).
[82] Para as referências a essas coleções, ver notas anteriores.

o outro[83].

As tecnologias digitais são as ferramentas para os estudiosos irem além desses limites, porque as normas, protocolos e ferramentas agora disponíveis nos permitem expressar a natureza hipertextual e hermenêutica de textos fragmentários, proporcionando aos estudiosos um *corpus* interligado de fontes primárias e secundárias de fragmentos que também inclui aparatos críticos, comentários, traduções e bibliografia moderna em textos antigos. O primeiro requisito para a construção de um acervo digital de textos fragmentários, portanto, é tornar os conteúdos semânticos de edições críticas impressas legíveis por máquina, definindo uma arquitetura geral para representar, pelo menos, os principais elementos seguintes relativos ao domínio de textos fragmentários em uma biblioteca digital[84]:

1) Citação como *Link* Processável por Máquina. Os fragmentos dos autores citados por Plutarco no exemplo mencionado acima devem ser vinculados ao texto completo de *A Vida de Teseu* (ver Anexo 3). Esta é a primeira função de uma representação adequada dos textos fragmentários: desta forma, é possível ver o trecho diretamente dentro de seu contexto de transmissão, evitando a ideia enganosa de uma existência material independente de textos fragmentários, que deriva da representação tipográfica de trechos que são, na verdade, o resultado de reconstruções modernas de obras perdidas. Esta função tem outra vantagem importante em uma biblioteca digital porque elimina o problema da repetição do mesmo texto dentro de um conjunto, tal como acontece, por exemplo, na biblioteca digital TLG[85].

2) Início e Fim de um Fragmento. Vincular o fragmento à sua origem significa recolocá-lo em seu contexto original. O próximo passo é oferecer um mecanismo para marcar o início e o fim de um fragmento neste contexto de acordo com as opções de

[83] Um comprimento diferente e apresentação do mesmo fragmento são perceptíveis nos casos de Filócoro (FHG I 392 fr. 49 = FGrH 328 F 110), Bíon (FHG II 19 fr. 1 = FGrH 14 F 2 = F2 FGrH 332) e Clidemo (FHG I 360 fr. 6 = FGrH 323 F 18). Há também um caso em que o mesmo fragmento de Helânico tem dois comprimentos diferentes na coleção de Jacoby: FGrH 4 F 167a e 323a FGrH F 17a. Finalmente, Hellan. FHG I 55 fr. 76 corresponde a dois fragmentos de diferentes em Jacoby (FGrH 4 F 165 = FGrH 323a F 15 e FGrH 4 F 166 = FGrH 323a F 16a).

[84] Sobre os aspectos técnicos, consulte Berti *et al.* (2009); Romanello *et al.* (2009a, 2009b).

[85] Este aspecto é particularmente importante quando os textos de uma biblioteca digital são usados para análise estatística computacional, porque o número de ocorrências de uma palavra resultante da busca textual pode ser completamente errada se a coleção tiver duplicatas do mesmo texto. Na TLG, coleções de autores fragmentários são apresentadas separadamente das edições de autores sobreviventes e isso significa que o mesmo texto pode ser repetido muitas vezes afetando o resultado de pesquisas textuais e palavra-chave.

editores diferentes. O resultado é que o leitor, ao visualizar no interior do trecho sua fonte de transmissão, pode ver a representação simultânea de diferentes comprimentos do mesmo fragmento com base em edições que adotaram critérios textuais diferentes (ver apêndice 3).

3) Numeração e Ordenação de Fragmentos. Numerar e ordenar fragmentos pode variar de forma significativa de uma edição para outra. Estas diferenças dependem das escolhas do editor, que pode decidir a ordem dos fragmentos - e, consequentemente, seu número de acordo com diferentes características internas ou externas dos próprios fragmentos ou de suas fontes[86]. As diferenças também podem ser o resultado de diferentes fragmentações do mesmo texto ou a necessidade de adicionar novos textos a um conjunto de fragmentos. Nosso modelo oferece a possibilidade de codificar este tipo de informação, que normalmente é registrada na tabela de concordâncias de uma edição impressa: alinhar várias referências para o mesmo objeto textual pode ajudar o leitor a visualizar diferentes numerações e ordenações de fragmentos em diferentes edições; o modelo também permite a inclusão de novos dados se foram acrescentadas novas.

4) Representação de Informações sobre os Autores e Obras Fragmentárias. Na fonte de transmissão do fragmento, é necessário especificar que um dado segmento de texto é o nome do autor ao qual é atribuído o fragmento e, em alguns casos também o título da obra e do número do livro aos quais o fragmento originalmente pertencia. Atribuir um fragmento a um autor e a uma obra pode ser uma tarefa difícil, porque podemos ter autores homônimos e também porque gerenciar títulos de obras antigas pode ser bastante complexo: na maioria dos casos, as testemunhas não citam o título da obra de onde retiraram o fragmento; além disso, em fontes antigas, o título de uma obra pode ser atestado com variantes mais ou menos significativas e o resultado é que os editores diferentes podem atribuir o mesmo fragmento a autores e obras diferentes[87]. O objetivo é desenvolver um catálogo completo de identificadores únicos para cada autor e obra fragmentária que incluirá diversas expressões do mesmo autor e obra e onde cada entrada terá metadados associados, proporcionando ao acadêmico uma espécie de cânone que

[86] Em FHG, historiadores fragmentários gregos são organizados em ordem cronológica, enquanto em FGrH eles têm um número e são divididos por gêneros. Os fragmentos são agrupados por obras dentro de ambas as coleções.

[87] Ver, por exemplo, Harding (2008, p. 1), sobre as diferentes formas como os autores antigos se referem as obras dos historiadores atenienses. Para autores homônimos, consulte Crates de Atenas e Crates de Mallus, que são ambos considerados possíveis autores de obra em glosas de Ático, atribuído por fontes antigas a um Crates não especificado: Broggiato (2000).

inclui simultaneamente toda a informação disponível sobre autores e obras fragmentárias, com ponteiros para fontes primárias e secundárias[88]. Esta função, além de proporcionar ao estudioso uma ferramenta inovadora, pode ser muito útil para reforçar uma das "questões teóricas" sugeridas por Glenn Most ao coletar fragmentos, ou seja, a relação entre autores fragmentários e os "limites de mudança de formação de cânone ao longo do tempo[89]".

5) Classificação de Fragmentos. Os fragmentos podem ser classificados segundo vários critérios, desde fatores internos a externos. A primeira classificação é baseada no gênero literário, praticamente coberto pela poesia épica e oratória e historiografia. Dentro da mesma coleção, os fragmentos normalmente identificados como *testimonia* (ou seja, fragmentos com dados biográficos e bibliográficos sobre os autores fragmentários) e fragmentos (ou seja, fragmentos de obras perdidas)[90].

Também podem ser aplicados outros critérios para classificar fragmentos pertencentes ao mesmo gênero literário, como é mostrado na obra monumental de Jacoby ao editar fragmentos históricos gregos, que são um dos resultados mais importantes alcançados no campo da historiografia antiga[91]. No entanto, a representação impressa destas categorias apresenta muitas limitações, porque é impossível traçar uma linha demarcatória entre muitos gêneros diferentes de autores e obras fragmentárias que podem ser inseridos em diferentes categorias que se sobrepõem: o resultado é que o mesmo fragmento é, muitas vezes, repetido em várias seções diferentes correspondentes a categorias diferentes[92]. Uma coleção digital em que cada fragmento é preservado em seu contexto original e representado com várias peças de metadados pode expressar a complexidade das classificações modernas, apesar de não espalhar e repetir o mesmo trecho várias vezes. Desta forma, é possível evitar o rigor da categoria impressa, permitindo que os estudiosos comparem um fragmento com outros trechos e visualizem

[88] O trabalho inicial sobre a criação de um catálogo e registros de autoridade para os autores fragmentárias foi realizado pela Biblioteca Digital da Perseu: ver Babeu (2008). Sobre a identificação entidade nomeada ver Blackwell-Crane (2009, p. 44).
[89] Most (1997, p. vi).
[90] Vale ressaltar o critério adotado por Diels e Kranz em sua coleção de filósofos pré-socráticos *(Die Fragmente der Vorsokratiker* I-III. Berlim 1951-1952[6]): A = Leben, Schriften, Lehre (ou seja, *testimonia* sobre a vida, obras e doutrinas de autores), B = Fragmente (ou seja, citações das obras de autores); C = *Imitationen* (ou seja, as obras que levam o autor como um modelo). Nem sempre é fácil distinguir entre *testimonia* e fragmentos: cf. Laks 1997. Para um exemplo pertencente a fragmentos históricos, ver Schepens (1997b).
[91] Schepens (1997a; 1998).
[92] Schepens (1997a, p. 148-154); (1998, p. ix-x).

como pertencem a categorias diferentes de uma forma mais dinâmica e simultânea.

Representando Variantes Textuais e Conjecturas

Frequentemente, as coleções impressas de fragmentos incluem um aparato crítico que, normalmente, não tem por base um novo exame dos manuscritos originais que testemunham o texto, mas uma seleção de variantes e conjecturas extraídas das melhores edições críticas de fontes de fragmentos. Esta escolha se deve, principalmente, ao fato de que levaria muito tempo para examinar cada manuscrito e também porque uma obra deste tipo iria além das competências e finalidades dos editores de fragmentos, cujo interesse principal é reconstruir o conteúdo e características de obras perdidas[93].

Coleções digitais de referência de fontes gregas e latinas, como o TLG e do CD-ROM do Packard Humanities Institute (PHI) baseiam-se no texto de uma única edição para cada fonte, sem incluir o aparato crítico: assim, são reproduções parciais de textos impressos e quando ao procurar a transmissão textual de uma passagem, os estudiosos precisam voltar a consultar o original impresso, comparando-o com outras edições e trabalhos filológicos quanto às variantes e conjecturas específicas[94].

Tanto a infraestrutura cibernética emergente para ciências humanas quanto as pesquisas realizadas no campo da *ePhilology* criaram um novo conceito de corpora textuais de grego e latim, onde o objetivo é oferecer serviços e métodos acadêmicos para acompanhar e comparar várias versões do mesmo texto, ao longo do tempo, afetando de uma maneira fundamental o trabalho futuro em textos fragmentários[95]:

1) Várias Edições e Esquemas de Alinhamento de Citação. O primeiro passo é recolher cada edição das fontes preservando fragmentos e as coleções de obras fragmentárias para que uma determinada passagem possa ser visualizada em diferentes

[93] Ricos aparatos críticos são fornecidos nas mais recentes coleções de fragmentos trágicos e cômicos (TrGF e PCG). Quando aos historiadores gregos fragmentários, Jacoby optou por um breve aparato crítico; o mesmo critério é seguido pelos editores da continuação de seu trabalho: ver Schepens (1998, p. xiii; 2000, p. 13-16).

[94] Sobre chamados "incunábulos digitais", que significa os primeiros projetos digitais que mantêm os pressupostos e os limites da cultura impressa, consulte Crane *et al.* (2006); Crane-Seales-Terras (2009, p. 35-37). A nova coleção com aparatos críticos é *Musisque Deoque* (http://www.mqdq.it), que é um arquivo digital de poesia latina com variantes, conjecturas e outras ferramentas exegéticas.

[95] Ver Blackwell-Crane (2009, p. 60-64) (com bibliografia), onde os três elementos fundamentais que devem caracterizar uma edição digital são apresentados: 1) inclusão de imagens de manuscritos, inscrições, papiros e outros materiais que transmitem o texto; 2) representação de várias edições produzidas por autores diferentes; 3) inclusão de *apparatus critici* que podem ser processados por máquina que permitem aos estudiosos comparar observações textuais com leituras a partir de manuscritos e outros materiais fonte. Cf. também Kraus (2009).

versões do mesmo texto reconstruído por editores diferentes. No exemplo citado acima, o objetivo é coletar todas as edições digitais da *Vida de Teseu* e os autores fragmentários citados por Plutarco. Quando não houver edições on-line disponíveis em um formato cuidadosamente transcrito, serão usados textos gerados através de reconhecimento óptico de caracteres (OCR) para criar elos entre uma passagem e uma imagem de página de várias edições da mesma passagem[96]. Além disso, dado que os regimes de citação podem ser diferentes, o sistema pode agrupar várias edições com o intuito de alinhar vários esquemas de citação[97].

2) Agrupamento Dinâmico de Várias Edições e Crítica Digital. A coleta de várias edições críticas do mesmo texto significa construir um "multitexto", que é uma "rede de versões com uma única raiz reconstruída", de modo que os estudiosos possam comparar diferentes opções textuais e conjecturas produzidas por filólogos[98]. Este processo envolve uma nova forma de conceber a crítica literária porque produz uma representação e visualização de transmissão textual totalmente diferente das convenções impressas nas quais o texto que é reconstruído pelo editor é separado do aparato crítico que é impresso na parte inferior da página. Além disso, com a inclusão de imagens de manuscritos, papiros, e outros materiais fontes, o leitor tem uma visualização dinâmica da tradição textual e percebe os diferentes canais de transmissão e a produção filológica do texto que, normalmente, estão ocultos nos aparatos críticos estáticos, concisos e necessariamente seletivos das edições impressas padrão[99]. Produzir um multitexto significa, portanto, produzir várias versões do mesmo texto, que são a representação das diferentes etapas de sua transmissão e reconstrução a partir de variantes do manuscrito para conjecturas filológicas. Este processo tem consequências fundamentais para o estudo de fontes antigas em geral e para os fragmentários em particular, uma vez que, ao estudar os fragmentos e

[96] Em textos gerados por OCR para fontes gregas clássicas, ver Stewart *et al.* (2007); Boschetti *et al.* (2009).
[97] Um caso complexo de vários esquemas de citação é fornecido pelo *Deipnosophistae de* Ateneu, cujo texto pode ser citado conforme as enumerações de Casaubon ou de Kaibel: ver Lenfant (2007b, p. 384-385). Um sistema para o alinhamento dos esquemas de citação para *Deipnosophistae* de Ateneu foi concebido pela Biblioteca Digital Perseus, que produziu uma versão experimental em XML desta obra com as duas numerações por Casaubon e Kaibel: ver Berti *et al.* (2009). Sobre representar citações em um ambiente digital cf. Smith (2009).
[98] Blackwell-Crane (2009, p. 60). Cf. acima nota 41. O conceito de multitexto é o resultado do trabalho realizado pelo Homer Multitext Project do Centro de Estudos Helênicos que visa a produzir uma nova representação digital da tradição textual dos poemas homéricos: ver Dué-Ebbott (2009) e Smith (2010). Sobre os aspectos técnicos do alinhamento das variantes e conjecturas ao texto, ver Boschetti (2007a, 2007b). Dué-Abbott de (2009, p. 1 e 13-18). Cf. Também Mordenti (2001, p. 42).
[99] Dué-Abbott de (2009, p. 1 e 13-18). Cf. Também Mordenti (2001, p. 42).

avaliar sua distância em relação à versão original, é "obrigatório" examinar as variantes dos manuscritos do texto-fonte para ver o que pode ser atribuído à testemunha ou para 47 a transmissão do texto através de séculos[100].

O Fragmento e seu Contexto Incorporado

Ao examinar os fragmentos, é possível distinguir dois esquemas principais de citação: 1) fragmentos de textos sobreviventes; 2) fragmentos de textos não sobreviventes. Estes esquemas incluem vários tipos de reprodução textual resultantes da atitude das diferentes testemunhas de citações e também do fato de que fontes gregas datam de uma época em que não existia uma norma para citação. Além disso, também podemos considerar duas outras grandes categorias que pertencem ao domínio da reutilização textual em fontes antigas; 3) passagens onde a testemunha não cita a fonte e a fonte sobrevive ou não[101]; 4) passagens onde citações e paráfrases não estão marcadas ou são difíceis de encontrar[102].

Com o intuito de verificar a confiabilidade de citações antigas e tirar pelo menos um espectro sombrio dos hábitos de citação de autores clássicos, o ponto de partida é analisar citações de textos sobreviventes. Quanto à literatura grega, uma das obras mais representativas é *Deipnosophistae* de Ateneu, que inclui uma enorme coleção de fragmentos perdidos e autores preservados. Vários estudos têm sido dedicados à coleta e comparação de citações de Ateneu sobre historiadores sobreviventes, como Heródoto, Tucídides e Xenofonte (ver Anexo 2-3)[103]. Mesmo se não seja possível dar um julgamento definitivo e completo de seu comportamento em relação a citações, estas análises permitiram aos estudiosos enumerar uma série de padrões recorrentes nas citações de Ateneu, incluindo uma ampla tipologia de reproduções textuais e características linguísticas que podem ser úteis na identificação e classificação de citações de

[100] Cf. Lenfant (2007a, p. 45).
[101] Ver, por exemplo, o problema da identificação de obras perdidas, literárias e documentais, utilizadas na *Constituição dos Atenienses* de Aristóteles: ver Rhodes (1981, p. 15-30). Cf. também Strasburger (1977, p. 27-30), em "das *anonyme* historische Gut in der Sekundärtradition".
[102] Por exemplo, em citações e paráfrases de Platão na literatura posterior: para um projeto com o objetivo de investigar este tipo de informação, veja abaixo nota 56.
[103] Ambaglio (1990); Pelling (2000); Lenfant (2007a); Maisonneuve (2007). Sobre a importância destes "estudos de controle", veja Strasburger (1977, p. 22-24); Brunt (1980, p. 480-481); Schepens (1997a, p. 167, n. 66).

historiadores perdidos[104]. A coleta de fragmentos de Xenofonte mostrou que as citações de Ateneu são mais ou menos confiáveis de acordo com diferentes temas, enquanto Pelling concentrou sua pesquisa sobre os "hábitos de transição" de Ateneu de passar de um assunto para o outro ao citar fontes diferentes, e na possibilidade de individuação nos "clusters fragmentários" de Deipnosophistae que significam grupos de citações de determinados autores reunidos e organizados na mesma estrutura tópica de uma seção do banquete aprendido[105].

Novas tecnologias, como a identificação de citação, também estão ajudando a encontrar citações em grandes bibliotecas digitais, como o Google Books e o Internet Archive, onde muitos documentos não seguem as convenções de publicações acadêmicas e contêm citações mais ou menos precisas sem citar a fonte ou onde as citações não podem ser reconhecidas automaticamente, como acontece frequentemente em fontes antigas. O objetivo destes métodos é fornecer *links* para materiais primários que são fundamentais para a compreensão de obras secundárias, uma vez que cada citação é apenas uma sombra distante do texto original[106]. Estas técnicas também começaram a ser aplicadas a obras de referência sobre a antiguidade clássica, proporcionando métodos iniciais para a identificação automática de citações na literatura secundária que têm estruturas diferentes devido à sua alteração a partir da fonte original, como a mudança da ordem das palavras, omissão, inserção ou substituição de termos e as diferenças de termo, dependendo de desconsideração de caso, caracteres acentuados, mudança de pontuação ou erro de ortografia e entrada de dados[107].

Pesquisa semelhante também foi realizada na aplicação de técnicas já desenvolvidas em outros campos, como a detecção automática de plágio, semelhança de texto buscando em diferentes documentos, reutilização de texto através de paráfrase ou referência indireta e detecção de alusão automática à literatura clássica[108]. Em particular, a Biblioteca Digital Perseu desenvolveu métodos para descobrir alusões textuais imitativas em uma coleção de poesia latina clássica e reutilização de texto multilíngue em

[104] Lenfant 2007a. Em convenções tipográficas utilizadas para marcar trechos literais, paráfrases e fragmentos de liquidação duvidosa em FGrH Continuação, ver Schepens 1998, xiii.
[105] Pelling de 2000. Em Xenofonte ver Maisonneuve (2007).
[106] Kolak-Schilit (2008); Schilit-Kolak (2008).
[107] Ernst-Gerlach-Crane (2008).
[108] Ernst-Gerlach-Crane (2008, p. 79). Para o pesquisa dos dados mais significativos relativos ao plágio textual em fragmentos históricos gregos, consulte Ambaglio (2009).

textos literários[109]. Ao mesmo tempo, um novo projeto chamado eAQUA está sendo desenvolvido pela Universidade de Leipzig com vistas à aplicação de técnicas de mineração[110] em textos antigos para criar uma reconstrução semântica de obras perdidas dos historiadores atenienses e citações de Platão na literatura posterior[111].

O desenvolvimento destas técnicas apresenta perspectivas desafiadoras para identificar e representar citações na literatura antiga, ampliando nossas possibilidades e capacidades de individuação de esquemas de citação que, por sua vez, também são úteis para identificar citações de obras perdidas e apoiar as interações mais sofisticadas entre estudiosos e textos digitalizados.

Fontes Secundárias e Terciárias

Coletar fragmentos também significa procurar diversos tipos de informações, direta ou indiretamente ligadas aos autores fragmentários. Em geral, estes dados são rotulados como "fontes secundárias" e "fontes terciárias" e podem ser resumidos nas seguintes categorias fundamentais: 1) *loci paralleli*, ou seja, antigas fontes secundárias paralelas ao testemunho de um fragmento. Mesmo que o relacionamento de um *locus parallelus* com o principal citador de um texto fragmentário envolva muitos aspectos, os *loci paralleli* formam dois grupos principais: a) fontes citando ou parafraseando o mesmo fragmento (na maioria dos casos, essas fontes são cronologicamente posteriores à testemunha); b) fontes lidando com o mesmo tema do fragmento. 2) fontes terciárias, isto é, bibliografia moderna composta de monografias, artigos, enciclopédias, gramáticas, traduções e outras ferramentas bibliográficas, dando informações e comentários sobre diversos materiais relativos ao fragmento, seu autor e sua fonte de transmissão.

A representação digital de textos fragmentários deve fornecer *links* para fontes secundárias e terciárias, identificando passagens em artigos e monografias relacionadas

[109] Bamman-Crane (2008b). Além disso, foi realizado um trabalho no Projeto Perseus em busca de alusões entre o *Paraíso Perdido* de John Milton e Eneida de Virgílio, que vão desde as mais semelhantes, como traduções, até o mais oblíquo, como alusões literárias: Bamman, D.; Crane, G. "Descobrir reutilizar o texto multilíngue em Textos Literários" (white paper). Para um modelo computacional de reutilização de texto dos evangelhos em grego do Novo Testamento, ver Lee (2007).

[110] N.T. *data-mining.*

[111] No projeto eAQUA (Extraktion von strukturiertem Wissen aus Antiken Quellen für die Altertumswissenschaft), ver Büchler-Heyer-Gründer (2008).

ao fragmento e do contexto a partir do qual o fragmento foi elaborado[112]. Como dito acima, os projetos de digitalização em massa estão criando muitas coleções de fontes secundárias e terciárias úteis para classicistas. Além disso, os repositórios como JSTOR e o Projeto MUSE dão acesso aos títulos dos principais periódicos acadêmicos em várias disciplinas, bem como monografias e outros materiais fundamentais para atividades acadêmicas: estes arquivos são textos completos pesquisáveis e abrem as portas para várias possibilidades de pesquisa interdisciplinar, como imagens de alta qualidade e citações e referências interligadas[113].

Além desses recursos, há outros projetos e publicações eletrônicas para classicistas digitais desenvolvidos por organizações como The Stoa Consortium e criados com o princípio do acesso aberto. Um dos projetos mais significativos do Stoa é o Suda On Line (SOL), que é particularmente importante para aqueles interessados na construção de um acervo digital de autores gregos fragmentários, uma vez que o Suda preserva uma grande quantidade de fragmentos de autores clássicos que na maioria dos casos, pode ser classificada como *loci paralleli.* O objetivo do projeto é criar uma versão *on-line* deste léxico enciclopédico oferecendo, pela primeira vez, uma tradução e aparato interpretativo para cada entrada, graças aos esforços de cooperação internacional de muitos estudiosos[114]. Todos estes recursos representam os tipos de fontes que devem ser incluídos quando da elaboração de uma representação digital de textos fragmentários para construir um *corpus* dinâmico e interligado de fontes primárias, secundárias e terciárias.

Tradução e Comentário

Dois outros elementos fundamentais de coleções modernas de obras fragmentárias que podem receber uma grande melhoria em bibliotecas digitais são traduções e comentários[115]. A tradução de textos significa não apenas prestar um serviço para quem não tem um bom conhecimento de línguas antigas, mas é, antes de tudo, uma parte essencial da interpretação erudita produzida pelo editor. Com a coleção de várias edições

[112] Berti *et al.* (2009, p. 260).
[113] JSTOR: http://www.jstor.org. Project MUSE: http://muse.jhu.edu.
[114] On the project see Mahoney (2009).
[115] Coleções antigas de fragmentos geralmente não incluem traduções e comentários. Uma exceção notável é o FHG, que inclui a tradução dos fragmentos para o latim, mas sem comentários. A primeira edição de fragmento incluindo um comentário foi concebida por Jacoby, mesmo que não tenha tradução: ver Schepens (1997a, p. 168); (1998, p. xiv); (2000, p. 16-17).

de uma mesma obra, uma biblioteca digital permitirá que os estudiosos também consultem várias traduções para vários idiomas, comparando diferentes interpretações e restituições linguísticas da mesma passagem. Em um nível mais profundo, alinhar várias edições nos permite criar dicionários e léxicos dinâmicos das palavras gregas e latinas e seus termos correspondentes em línguas modernas lidos por máquina, criando uma ferramenta inestimável para os estudiosos e para um amplo leque de análise linguística, gramatical e sintática[116].

No que tange aos fragmentos, o comentário ao texto é constituído de duas funções fundamentais: a primeira é o esforço para "desmontar" o contexto que preserva a citação com o intuito de encontrar as características originais do fragmento e a segunda é tentar "reconstruir" o fragmento e a obra perdida à qual pertencia[117]. Quanto às variantes textuais, conjecturas e traduções, uma biblioteca digital deve ter cada passagem com links para vários comentários extraídos das edições de fragmentos e textos fonte. Um comentário digital verdadeiro, no entanto, pode ser concebido como algo mais amplo, pois pode incluir cada anotação possível identificando todos os fenômenos relativos ao texto, trazendo assim comentários tradicionais com uma ampla série de serviços, que vão desde a análise morfológica e sintática à identificação da entidade em si e diferentes 65 explicações ou disputas sobre todos os aspectos do conteúdo textual[118].

A elaboração de uma edição digital de textos fragmentários também pode ter consequências importantes na representação de coleções modernas de fontes antigas, ou seja, livros fonte que foram publicados há muitos anos, e oferece aos estudiosos e estudantes textos de referência sobre diversos assuntos relativos ao mundo clássico. O principal problema a ser enfrentado com esses tipos de coleções é a organização das fontes. Um exemplo significativo é representado pelo conjunto de fontes que pertencem ao chamado Pentekontaetia, que foi originalmente publicado por George Hill em 1897: neste texto o editor organizou extratos de fontes históricas gregas por tema (sem tradução)[119]. Cinquenta anos mais tarde, a editora Russell Meiggs e Anthony Andrewes decidiu publicar uma edição revista do livro de Hill, não apenas para adicionar novas

[116] Blackwell-Crane (2009, p. 46-47, 50, 65-71); Bamman-Crane (2008a, 2009). Ver também Bamman-Babeu-Crane (2010).
[117] Schepens (1997a, p. 168).
[118] Blackwell-Crane (2009, p. 77).
[119] Hill (1897, p. vi). Por razões de espaço e custo, o editor não incluiu trechos de fontes principais, como Heródoto, Tucídides, e a Constituição dos Atenienses de Aristóteles.

fontes epigráficas, mas também com uma nova organização das fontes. Nesta nova edição, os extratos são apresentados de acordo com a ordem alfabética dos seus autores e o livro contém índices completos que lidam com diversos temas históricos referentes à Pentekontaetia, e também antropônimos, topônimos e outros tipos de nomes próprios[120].

Mesmo que estas coleções sejam ferramentas inestimáveis para estudiosos e estudantes, esses dependem dos limites impostos pelas edições impressas, que obrigam um editor a escolher apenas um critério para organizar os textos e selecionar um número de fontes extraídas de seus contextos[121]. A representação digital de tais coleções de antigas fontes primárias permitiria aos estudiosos irem além desses limites e apresenta uma série de funções fundamentais: o acesso imediato aos textos completos em ambas as línguas antigas originais e traduções modernas; vários pontos de entrada para as informações, como fontes, eventos, nomes e geografia; itens interrelacionados em diferentes fontes; uma representação sinótica gráfica de fontes de acordo com a ordem cronológica dos eventos, geografia e categoria de informações, modelos de coleta e organização de prova antiga para outros períodos, ou temas, ou abordagens, links para informações de background sobre as fontes e, finalmente, imagens de inscrições, moedas e manuscritos, além de mapas e outros desenhos[122].

Conclusão

Elaborar um modelo e uma arquitetura para representar textos fragmentários em uma biblioteca digital é uma contribuição fundamental para uma análise sistemática e estrutural das várias camadas de produção e interpretação que constituem um fragmento textual. Em particular, as duas metas mais importantes de tal obra são: 1) Representar um fragmento textual como um hipertexto, ou seja, como um texto a partir de um texto derivado de outro texto e interligado para muitas outras tipologias diferentes de textos: isto significa idealizar e construir um conjunto expansível de links que expressam várias relações do texto do fragmento com o texto que o incorpora e transmite e com diversas

[120] Hill (1951).
[121] Outro exemplo é fornecido por um conjunto de fragmentos recente dos historiadores atenienses publicados por Harding (2008). Neste caso, o autor apresenta os fragmentos apenas em inglês, não por autor (como em FGrH), mas por tópico e data.
[122] Ver Martin (2009), para um projeto que visa representar fontes fragmentárias no Pentekontaetia em uma biblioteca digital.

fontes secundárias e terciárias (por exemplo, prova antiga, comentários e outros tipos de ferramentas bibliográficas). 2) Representar um fragmento textual como um multitexto, ou seja, como o resultado de uma obra de estratificação de variantes do manuscrito e conjecturas acadêmicas que formam o caminho através do qual o fragmento sobreviveu e sem o qual não existiria como prova.

Referências

ALBONICO, S. Sull'utilizzo della codifica TEI in filologia. In: CIOTTI, F. (Ed.). **Il manuale TEI Lite. Introduzione alla codifica elettronica dei testi letterari**. Milano: Edizioni Sylvestre Bonnard, 2005. p. 239-256.

AMBAGLIO, D. I Deipnosofisti di Ateneo e la tradizione storica frammentaria. **Athenaeum**, v. 78, n. 1, p. 51-64, 1990.

_____. Nelle pieghe dei frammenti degli storici greci, tra falsificazioni e plagi. In: LANZILLOTTA, E.; COSTA, V.; OTTONE, G. **Tivoli: Tradizione e trasmissione degli storici greci frammentari**. Edizioni Tored, 2009. p. 543-562.

BABEU, A. **Building a "FRBR-Inspired" Catalog: the Perseus Digital Library Experience.** The Perseus Digital Library. 2008. Disponível em http://www.perseus.tufts.edu/publications/PerseusFRBRExperiment.pdf.

_____. **Rome Wasn't Digitized in a Day: Building a Cyberinfrastructure for Digital Classicists.** CLIR Report. 2010. Disponível em: http://www.clir.org/activities/details/infrastructure.html.

BAMMAN, D.; CRANE, G. Building a Dynamic Lexicon from a Digital Library. In: **Proceedings of the 2008 Joint International Conference on Digital libraries (JCDL '08).** Pittsburgh, p. 11-20. New York: ACM Digital Library. 2008a. Disponível em: http://dl.tufts.edu/view_pdf.jsp?pid=tufts:PB.001.002.00003.

_____. The Logic and Discovery of Textual Allusion. In: **Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008).** Marrakech: ELRA. 2008b. Disponível em: http://dl.tufts.edu/view_pdf.jsp?pid=tufts:PB.001.002.00004.

_____. Computational Linguistics and Classical Lexicography. **Digital Humanities Quarterly**, v. 3, n. 1, 2009. Disponível em: http://www.digitalhumanities.org/dhq/vol/3/1/000033/000033.html.

BAMMAN, D.; BABEU, A.; CRANE, G. Transferring Structural Markup Across Translations Using Multilingual Alignment and Projection. In: **Proceedings of the 10th Annual Joint Conference on Digital libraries (JCDL '10). Gold Coast**, Australia, p. 11-20. New York: ACM Digital Library. 2010. Disponível em:

http://www.perseus.tufts.edu/publications/jcdl27-bamman.pdf.

BENJAMINS, V. R. *et al.* Cultural Heritage and the Semantic Web. In: **The Semantic** Web**: Research and Applications. First European Semantic** Web **Symposium**, ESWS 2004, Heraklion, Crete, Greece. Proceedings, p. 433-444. Berlin: Springer Verlag, 2004. Disponível em: http://www.springerlink.com/content/nqbe4me9rvtwfwf4/.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. Scientific American, v. 284, n. 5, p. 34-43. 2001. Disponível em: http://www.scientificamerican.com/article.cfm?id=the-semantic-web.

BERTI, M. *et al.* Collecting Fragmentary Authors in a Digital Library. In: **Proceedings of the 2009 Joint International Conference on Digital libraries (JCDL '09)**, Austin, TX, p. 259-262. New York: ACM Digital Library. 2009. Disponível em: http://www.perseus.tufts.edu/publications/JCDL09_sp.pdf.

BLACKWELL, C.; CRANE, G. Cyberinfrastructure, the Scaife Digital Library and Classics in a Digital Age. **Digital Humanities Quarterly**, v. 3, n. 1, 2009. Disponível em: http://www.digitalhumanities.org/dhq/vol/3/1/000035/000035.html.

BODARD, G. EpiDoc: Epigraphic Documents in XML for Publication and Interchange. In: FERAUDI-GRUÉNAIS, F. (Ed.). **Epigraphy on Stone. Epigraphic Research and Electronic Archives**. Lanham: Lexongton Books, 2010. p. 101-118.

BODARD, G.; MAHONY, S. **Digital Research in the Study of Classical Antiquity**. Farnham: Ashgate, 2010.

BOLTER, J. D. **Turing's Man. Western Culture in the Computer Age**. Chapel Hill: The University of North Carolina Press, 1984.

_____. The Computer, Hypertext, and Classical Studies. **AJP**, v. 112, n. 4, p. 541-545, 1991.

_____. **Writing Space. Computers, Hypertext, and the Remediation of Print**. Mahwah and London: Lawrence Erlbaum Associates Publishers, 2012.

BOSCHETTI, F. Alignment of Variant Readings for Linkage of Multiple Annotations. In: ZEMÁNEK, P. *et al.* (Ed.). **Chatre!!ar 2007. Electronic Corpora of Ancient Languages. Proceedings of the International Conference.** Prague: Charles University, 2007a. p. 11-24.

_____. Methods to Extend Greek and Latin Corpora with Variants and Conjectures. Mapping Critical Apparatuses onto Reference Text. In: **Proceedings of the *Corpus* Linguistics Conference (CL 2007).** University of Birmingham, UK, 2007b. p. 1-11. Disponível em: http://ucrel.lancs.ac.uk/publications/CL2007/paper/150_Paper.pdf.

BOSCHETTI, F. *et al.* Improving OCR Accuracy for Classical Critical Editions. In: **Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009).** Corfu, September 27 – October 2, p. 156-167. Berlin: Springer Verlag, 2009. Disponível em: http://www.springerlink.com/content/v47r4148730271m8/.

BOWERSOCK, G. W. Jacoby's Fragments and Two Greek Historians of Pre-Islamic Arabia, **Most**, p. 173-185, 1997.

BROGGIATO, M. Athenaeus, Crates and Attic Glosses. A Problem of Attribution. In: BRAUND, D.; WILKINS, J. (Ed.). **Athenaeus and his World. Reading Greek Culture in the Roman Empire.** Exeter: University of Exeter Press, 2000. p. 364-370.

BRUNT, P. A. On Historical Fragments and Epitomes. **CQ**, v. 30, n. 2, p. 477-494, 1980.

BÜCHLER, M.; HEYER, G.; GRÜNDER, S. eAQUA - Bringing Modern Text Mining Approaches to Two Thousand Years Old Ancient Texts. In: **e-Humanities - An Emerging Discipline: Workshop in the 4th IEEE International Conference on e-Science**. Indianapolis. 2008. Disponível em: http://www.clarin.eu/system/files/2008-09-05-IEEE2008-eAQUA- project.pdf.

BÜNTE, M. Text Mining with the Atthidographers. In: SCHUBERT, C.; HEYER, G. **Das Portal eAQUA – Neue Methoden in der geisteswissenschaftlichen Forschung I**. Leizpig: Universität Leipzig, p. 10-25. Disponível em: http://www.eaqua.net/public.php.

BURNARD, L.; BAUMAN, S. **TEI P5: Guidelines for Electronic Text Encoding and Interchange.** Oxford: The TEI Consortium. 2009. Disponível em http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

CAYLESS, H. *et al.* Epigraphy in 2017. **Digital Humanities Quarterly**, v. 3, n. 1. 2009. Disponível em: http://www.digitalhumanities.org/dhq/vol/3/1/000030/000030.html.

CIOTTI, F. La codifica del testo, XML e la Text Encoding Initiative. In: CIOTTI, F. (Ed.). **Il manuale TEI Lite. Introduzione alla codifica elettronica dei testi letterari**, p. 9-42. Milano: Edizioni Sylvestre Bonnard, 2005. Disponível em: http://www.tei-c.org/Guidelines/Customization/Lite/teiu5_it.xml.

CRANE, G. From the Old to the New: Integrating Hypertext into Traditional Scholarship. In: **Proceedings of the ACM Hypertext 87 Conference. Chapel Hill, NC,** p. 51-55. New York: ACM Digital Library, 1987. Disponível em: http://doi.acm.org/10.1145/317426.317432.

_____. Give us editors! Re-inventing the edition and re-thinking the humanities. In: MCGANN, J. (Ed.). **Online Humanities Scholarship: The Shape of Things to Come.**

Proceedings of the Mellon Foundation Online Humanities Conference at the University of Virginia, March 26-28, p. 137-170, 2010. Houston: Rice University Press. Disponível em: http://rup.rice.edu/cnx_content/shape/m34316.html.

CRANE, G.; BAMMAN, D. Cyberinfrastructure and the Next Generation of Ancient Corpora. In: ZEMÁNEK, P. *et al.* (Ed.). **Chatre!!ar 2007. Electronic Corpora of Ancient Languages. Proceedings of the International Conference**. Prague: Charles University, 2007. p. 43-57.

CRANE, G.; SEALES, B.; TERRAS, M. Cyberinfrastructure for Classical Philology. **Digital Humanities Quarterly**, v. 3, n. 1, 2009. Disponível em: http://www.digitalhumanities.org/dhq/vol/3/1/000023/000023.html.

CRANE, G. *et al.* Composing Culture: the Authority of an Electronic Text. **Current Anthropology**, v. 32, n. 3, p. 293-311, 1991.

_____. Beyond Digital Incunabula: Modeling the Next Generation of Digital Libraries. In: **Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006).** Alicante, Spain, p. 341-352. Berlin: Springer Verlag. 2006. Disponível em: http://dl.tufts.edu/view_pdf.jsp?pid=tufts:gcrane-2006.00002.

_____. ePhilology: When the Books Talk to their Readers. In: SIEMENS, R.; SCHREIBMAN, S. **A Companion to Digital Literary Studies.** p. 29-64. Malden: Blackwell Publishing. 2007. Disponível em: http://www.digitalhumanities.org/companionDLS/.

_____. Tools for Thinking: ePhilology and Cyberinfrastructure. In: **Working Together or Apart: Promoting the Next Generation of Digital Scholarship: Report of a Workshop Cosponsored by the Council on Library and Information Resources and The National Endowment for the Humanities.** Washington: Council on Library and Information Resources, 2009a. p. 16-26. Disponível em: http://www.clir.org/activities/digitalscholar2/crane11_11.pdf.

_____. Classics in the Million Book Library. **Digital Humanities Quarterly**, v. 3, n. 1. 2009b. Disponível em: http://www.digitalhumanities.org/dhq/vol/3/1/000034/000034.html.

DIONISOTTI, A. C. On Fragments in Classical Scholarship, **Most**, p. 1-33. 1997.

DUÉ, C.; EBBOTT, M. Digital Criticism: Editorial Standards for the Homer Multitext. **Digital Humanities Quarterly,** v. 3, n. 1, 2009. Disponível em: http://www.digitalhumanities.org/dhq/vol/3/1/000029/000029.html.

ERNST-GERLACH, A.; CRANE, G. Identifying Quotations in Reference Works and Primary Materials. In: **Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2008).** Århus, Denmark, p. 78-87. Berlin – Heidelberg: Springer Verlag. Disponível em: http://www.springerlink.com/index/7626500277722t16.pdf.

FIORMONTE, D. **Scrittura e filologia nell'era digitale**. Torino: Bollati Boringhieri, 2003.

GENETTE, G. **Palimpsests. Literature in the Second Degree**. Lincoln – London: University of Nebraska Press, 1997. [original édition G. Genette, **Palimpsestes. La littérature au second degré**. 1987. Paris: Éditions du Seuil].

GUMBRECHT, H. U. Eat Your Fragment! About Imagination and the Restitution of Texts, **Most**, 1997, p. 315-327.

HARDING, P. H. **The Story of Athens. The Fragments of the Local Chronicle of Attica.** New York: Routledge, 2008.

HARDWICK, L. Electrifying the Canon: The Impact of Computing on Classical Studies. **Computers and the Humanities**, v. 34, p. 279-295, 2000. Disponível em: http://www.springerlink.com/index/P3532154W30M541T.pdf.

HENDLER, J. *et al.* Web Science: An Interdisciplinary Approach to Understanding the Web. **Communications of the ACM**, v. 51, n. 7, p. 60-69, 2008. Disponível em: http://portal.acm.org/citation.cfm?id=1364782.1364798.

HILL, G. F. **Sources for Greek History Between the Persian and the Peloponnesian Wars**. Oxford: The Clarendon Press, 1897. Disponível em: http://www.archive.org/details/sourcesforgreek01hillgoog.

HILL, G. F. **Sources for Greek History Between the Persian and Peloponnesian Wars**. Oxford: The Clarendon Press, 1951.

IRELAND, S. The Computer and Its Role in Classical Research, **G&R**, v. 23, n. 1, p. 40-54, 1976.

KOLAK, O.; SCHILIT, B. N. Generating Links by Mining Quotations. In: **Proceedings of the nineteenth ACM conference on Hypertext and hypermedia (HT 2008).** Pittsburgh, Pennsylvania. New York: ACM Digital Library, 2008. p. 117-126. Disponível em: http://portal.acm.org/citation.cfm?id=1379117.

KRAUS, K. 2009. Conjectural Criticism: Computing Past and Future Texts. **Digital Humanities Quarterly**, v. 3, n. 4. Disponível em: http://www.digitalhumanities.org/dhq/vol/3/4/000069/000069.html.

LAKS, A. Du témoignage comme fragment, **Most**, p. 237-272, 1997.

LANDOW, G. P. **Hypertext 3.0. Critical Theory and New Media in an Era of Globalization.** Baltimore: John Hopkins University Press, 2006.

LANZILLOTTA, E. La nuova collana «I Frammenti degli Storici Greci». In: AMPOLO, C. (Ed.). **Aspetti dell'opera di Felix Jacoby.** Pisa: Scuola Normale Superiore, 2006. p. 287-292.

LEE, J. A Computational Model of Text Reuse in Ancient Literary Texts. In: **Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (June 2007).** Prague: Association for Computational Linguistics, 2007. p. 472-479. Disponível em: http://groups.csail.mit.edu/sls//publications/2007/P07-1060.pdf.

LENFANT, D. Les «fragments» d'Hérodote dans le Deipnosophistes. In: LENFANT, D. (Ed.). **Athénée et les fragments d'historiens. Actes du colloque de Strasbourg (16-18 juin 2005).** Paris: De Boccard, 2007a. p. 43-72.

LENFANT, D. Athénée: texte et systèmes de référence. In: LENFANT, D. (Ed.). **Athénée et les fragments d'historiens. Actes du colloque de Strasbourg (16-18 juin 2005).** Paris: De Boccard, 2007b.

MAHONEY, A. Tachypaedia Byzantina: The Suda On Line as Collaborative Encyclopedia. **Digital Humanities Quarterly**, v. 3, n. 1, 2009. Disponível em: http://www.digitalhumanities.org/dhq/vol/3/1/000025/000025.html.

MAISONNEUVE, C. Les «fragments» de Xénophon dans les Deipnosophistes. In: LENFANT, D. (Ed.). **Athénée et les fragments d'historiens. Actes du colloque de Strasbourg (16-18 juin 2005).** Paris: De Boccard, 2007. p. 73-106.

MARINCOLA, J. **Bonnechère, P. Die Fragmente der griechischen Historiker**. Indexes of Part I, II, and III. Indexes of Ancient Authors. Leiden, Brill, 1999. BMCR, 2000. Disponível em: http://bmcr.brynmawr.edu/2000/2000-01-09.html.

_____. F. Jacoby, Die Fragmente der griechischen Historiker. CD-ROM Edition. Leiden, Brill, 2004. BMCR. 2005. Disponível em: http://bmcr.brynmawr.edu/2005/2005-08-37.html.

MARTIN, Th. R. The Challenges for Education and Research of the Fragmentary State of the Primary Sources for the Golden Age of Ancient Greek History. In: **Proceedings of the International Symposium on the Scaife Digital Library,** Lexington, March 13, 2009. (forthcoming).

MCMANUS, B. F.; RUBINO, C. A. 2003. Classics and Internet Technology. **AJP**, v. 124, n. 4, p. 601-608.

MORDENTI, R. **Informatica e critica dei testi.** Roma: Bulzoni Editore, 2001.

MOST, G. W. **Collecting Fragments - Fragmente sammeln.** Göttingen: Vandenhoeck & Ruprecht, 1997.

NUMERICO, T.; FIORMONTE, D.; TOMASI, F. **L'umanista digitale**. Bologna: Il Mulino.

OED2 = **The Oxford English Dictionary**. SIMPSON, J. A.; WEINER, E. S. C. I-XX, Oxford, 1989.

PARODI, M.; FERRARA, A. 2002. XML, semantic web e rappresentazione della conoscenza. **Mondo Digitale** v. 3, p. 42-51, 2010. Disponível em: http://www.mondodigitale.net/Rivista/motore/parodi_p. 42-51.pdf.

PELLING, C. Fun with Fragments. Athenaeus and the Historians. In: **Athenaeus and his World. Reading Greek Culture in the Roman Empire.** BRAUND, D.; WILKINS, J. Exeter: University of Exeter Press, 2000. p. 171-190.

POLACCO, M. **L'intertestualità**. Roma-Bari: Editori Laterza, 1998.

PRITCHARD, D. Working Papers, Open Access, and Cyber-infrastructure in Classical Studies. **Literary and Linguistic Computing**, v. 23, n. 2, p. 149-162, 2008. Disponível em: http://llc.oxfordjournals.org/cgi/content/abstract/23/2/149.

RHODES, P. J. **A Commentary on the Aristotelian Athenaion Politeia**. Oxford: Oxford University Press, 1981.

ROMANELLO, M. The Digital Critical Edition of Fragments. Theoretical Problems and Technical Solutions, **Proceeding Verona**, 2010 (forthcoming).

ROMANELLO, M. *et al.* When Printed Hypertexts Go Digital: Information Extraction from the Parsing of Indices. **Hypertext 2009: Proceedings of the 20th ACM Conference on Hypertext and Hypermedia, Turin, Italy**, p. 357-358. New York: ACM Digital Library, 2009a. Disponível em: http://www.perseus.tufts.edu/publications/ht159-romanello.pdf.

_____. Rethinking Critical Editions of Fragmentary Texts by Ontologies. **ELPUB 2009: 13th International Conference on Electronic Publishing: Rethinking Electronic Publishing: Innovation in Communication Paradigms and Technologies,** Milan, Italy, p. 155-174. 2009b. Disponível em: http://www.perseus.tufts.edu/publications/elpub2009.pdf.

ROSENZWEIG, R. Should Historical Scholarship Be Free? **AHA Perspectives**, v. 43, 2005. Disponível em: http://www.historians.org/perspectives/issues/2005/0504/0504vic1.cfm.

_____. Can History Be Open Source? Wikipedia and the Future of the Past. **The Journal of American History**, v. 93, p. 117-146, 2006.

SCHEPENS, G. Jacoby's FGrHist: Problems, Methods, Prospects, **Most**, p. 144-172. 1997a.

_____. Timaeus FGrHist 566 F28 Revisited: Fragmenta or Testimonia?, **Simblos**, v. 2, p. 71-83, 1997b.

_____. Prolegomena. **Die Fragmente der Griechischen Historiker Continued, Part IV, IVA1, VII- XXI.** Leiden - Boston - Köln: Brill, 1998.

_____. Probleme der Fragmentedition. (Fragmente der griechischen Historiker). In: REITZ, C. (Ed.). **Vom Text zum Buch.** St. Katharinen: Scripta Mercaturae Verlag, 2000. p. 1-29.

SCHILIT, B. N.; KOLAK, O. Exploring a Digital Library through Key Ideas. In: **Proceedings of the 2008 Joint Conference on Digital Libraries (JCDL '08). Pittsburgh, PA.** New York: ACM Digital Library, 2008. p. 177-186. Disponível em: http://portal.acm.org/citation.cfm?id=1378920.

SCHUBERT, C. Zitationsprofile, Suchstrategien und Forschungsrichtungen. In: SCHUBERT, C.; HEYER, G. (Ed.). **Das Portal eAQUA – Neue Methoden in der geisteswissenschaftlichen Forschung I.** Leizpig: Universität Leipzig, 2010. p. 42-55. Disponível em: http://www.eaqua.net/public.php.

SHADBOLT, N. *et al.* The Semantic Web Revisited. **IEEE Intelligent Systems**, v. 21, n. 3, p. 96-101, 2006. Disponível em: http://portal.acm.org/citation.cfm?id=1155373.

SMITH, N. Citation in Classical Studies. **Digital Humanities Quarterly**, v. 3, n. 1, 2009. Disponível em: http://www.digitalhumanities.org/dhq/vol/3/1/000028/000028.html.

SMITH, N. Digital Infrastructure and the Homer Multitext Project. In: BODARD, G.; MAHONY, S. (Ed.). **Digital Research in the Study of Classical Antiquity.** Farnham: Ashgate, 2010. p. 121-137.

STEWART, G. *et al.* A New Generation of Textual Corpora. Mining Corpora from Very Large Collections. **Proceedings of the 2007 Joint Conference on Digital Libraries (JCDL '07).** New York, Vancouver: ACM Digital Library, 2007. p. 356-365. Disponível em: http://portal.acm.org/citation.cfm?id=1255175.1255247.

STRASBURGER, H. Umblick im Trümmerfeld der griechischen Geschichtsschreibung. In: **Historiographia Antiqua. Commentationes Lovanienses in honorem W. Peremans septuagenarii editae**. Leuven: Leuven University Press, 1977. p. 3-52.

TRILLINI, R. H.; QUASSDORF, S. A 'key to all quotations'? A corpus-based parameter model of intertextuality. **Literary and Linguistic Computing**, v. 25, n. 3, p. 269-286, 2010.

WELSHONS, M. **Our Cultural Commonwealth. The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences.** American Council of Learned Societies. 2006. Disponível em: www.acls.org/cyberinfrastructure/ourculturalcommonwealth.pdf.

WILLINSKY, J. **The Access Principle. The Case for Open Access to Research and Scholarhip.** Cambridge: MIT Press, 2005.

WORTHINGTON, I. Worthington on Marincola on Jacoby's FGrH. BMCR 2005.09.24. 2005. Disponível em: http://bmcr.brynmawr.edu/2005/2005-09-24.html.

WRIGHT, L. Classicists on the Via Electronica. **TAPA,** v. 124, p. 337-340, 1994.

Apêndice 1

Neste apêndice vamos publicar o texto de Plutarco, *Vida de Teseu* 24-28 de acordo com a edição de Perrin: *Plutarch's Lives*, I, ed. B. Perrin. Cambridge, MA: Harvard University Press. 1959.

Figura 1 – Plutarco, Teseu 24-28 (Perrin)

24 (1) Μετὰ δὲ τὴν Αἰγέως τελευτὴν μέγα καὶ θαυμαστὸν ἔργον εἰς νοῦν βαλόμενος συνῴκισε τοὺς τὴν Ἀττικὴν κατοικοῦντας εἰς ἓν ἄστυ, καὶ μιᾶς πόλεως ἕνα δῆμον ἀπέφηνε, τέως σποράδας ὄντας καὶ δυσανακλήτους πρὸς τὸ κοινὸν πάντων συμφέρον, ἔστι δ' ὅτε καὶ διαφερομένους ἀλλήλοις καὶ πολεμοῦντας. **(2)** ἐπιὼν οὖν ἀνέπειθε κατὰ δήμους καὶ γένη, τῶν μὲν ἰδιωτῶν καὶ πενήτων ἐνδεχομένων ταχὺ τὴν παράκλησιν αὐτοῦ, τοῖς δὲ δυνατοῖς ἀβασίλευτον πολιτείαν προτείνων καὶ δημοκρατίαν αὐτῷ μόνον ἄρχοντι πολέμου καὶ νόμων φύλακι χρησομένην, τῶν δὲ ἄλλων παρέξουσαν ἅπασιν ἰσομοιρίαν. **(3)** τοὺς μὲν ταῦτα ἔπειθεν, οἱ δὲ τὴν δύναμιν αὐτοῦ δεδιότες μεγάλην οὖσαν ἤδη καὶ τὴν τόλμαν, ἐβούλοντο πειθόμενοι μᾶλλον ἢ βιαζόμενοι ταῦτα συγχωρεῖν. καταλύσας οὖν τὰ παρ' ἑκάστοις πρυτανεῖα καὶ βουλευτήρια καὶ ἀρχάς, ἓν δὲ ποιήσας ἅπασι κοινὸν ἐνταῦθα πρυτανεῖον καὶ βουλευτήριον ὅπου νῦν ἵδρυται τὸ ἄστυ, τήν τε πόλιν Ἀθήνας προσηγόρευσε καὶ Παναθήναια θυσίαν ἐποίησε κοινήν. **(4)** ἔθυσε δὲ καὶ Μετοίκια τῇ ἕκτῃ ἐπὶ δέκα τοῦ Ἑκατομβαιῶνος, ἣν ἔτι νῦν θύουσι. καὶ τὴν βασιλείαν ἀφεὶς, ὥσπερ ὡμολόγησε, διεκόσμει τὴν πολιτείαν ἀπὸ θεῶν ἀρχόμενος· ἦκε γὰρ αὐτῷ χρησμὸς ἐκ Δελφῶν (Parke-Wormell 2.154) μαντευομένῳ περὶ τῆς πόλεως·

(5) Αἰγείδη Θησεῦ, Πιτθηΐδος ἔκγονε κούρης,
πολλαῖς τοι πολίεσσι πατὴρ ἐμὸς ἐγκατέθηκε
τέρματα καὶ κλωστῆρας ἐν ὑμετέρῳ πτολιέθρῳ.
ἀλλὰ σὺ μή τι λίην πεπονημένος ἔνδοθι θυμὸν
βουλεύειν· ἀσκὸς γὰρ ἐν οἴδματι ποντοπορεύσεις.

τοῦτο δὲ καὶ Σίβυλλαν (Hendess 23) ὕστερον ἀποστοματίσαι πρὸς τὴν πόλιν ἱστοροῦσιν, ἀναφθεγξαμένην·

Ἀσκὸς βαπτίζῃ· δῦναι δέ τοι οὐ θέμις ἐστίν.

25 (1) Ἔτι δὲ μᾶλλον αὐξῆσαι τὴν πόλιν βουλόμενος ἐκάλει πάντας ἐπὶ τοῖς ἴσοις, καὶ τὸ "Δεῦρ' ἴτε πάντες λεώ" κήρυγμα Θησέως γενέσθαι φασὶ πανδημίαν τινὰ καθιστάντος. οὐ μὴν ἄτακτον οὐδὲ μεμιγμένην περιεῖδεν ὑπὸ πλήθους ἐπιχυθέντος ἀκρίτου γενομένην τὴν δημοκρατίαν, ἀλλὰ πρῶτος ἀποκρίνας χωρὶς εὐπατρίδας καὶ γεωμόρους καὶ δημιουργούς, **(2)** εὐπατρίδαις δὲ γινώσκειν τὰ θεῖα καὶ παρέχειν ἄρχοντας ἀποδοὺς καὶ νόμων διδασκάλους εἶναι

Figura 2 – Plutarco, Teseu 24-28 (Perrin) (continuação)

καὶ ὁσίων καὶ ἱερῶν ἐξηγητάς, τοῖς ἄλλοις πολίταις ὥσπερ εἰς ἴσον κατέστησε, δόξῃ μὲν εὐπατριδῶν, χρείᾳ δὲ γεωμόρων, πλήθει δὲ δημιουργῶν ὑπερέχειν δοκούντων. ὅτι δὲ πρῶτος ἀπέκλινε πρὸς τὸν ὄχλον, ὡς Ἀριστοτέλης (*Ath.Pol.* 41.2; F 384 Rose[3]) φησί, καὶ ἀφῆκε τὸ μοναρχεῖν, ἔοικε μαρτυρεῖν καὶ Ὅμηρος (*Ilias* 2.547) ἐν νεῶν καταλόγῳ μόνους Ἀθηναίους δῆμον προσαγορεύσας. **(3)** Ἔκοψε δὲ καὶ νόμισμα, βοῦν ἐγχαράξας, ἢ διὰ τὸν Μαραθώνιον ταῦρον, ἢ διὰ τὸν Μίνω στρατηγόν, ἢ πρὸς γεωργίαν τοὺς πολίτας παρακαλῶν. ἀπ' ἐκείνου δέ φασι τὸ ἑκατόμβοιον καὶ τὸ δεκάβοιον ὀνομασθῆναι. προσκτησάμενος δὲ τῇ Ἀττικῇ τὴν Μεγαρικὴν βεβαίως, τὴν θρυλουμένην ἐν Ἰσθμῷ στήλην ἔστησεν, ἐπιγράψας τὸ διορίζον ἐπίγραμμα τὴν χώραν δυσὶ τριμέτροις, ὧν ἔφραζε τὸ μὲν πρὸς ἕω

Τάδ' οὐχὶ Πελοπόννησος, ἀλλ' Ἰωνία·

τὸ δὲ πρὸς ἑσπέραν

Τάδ' ἐστὶ Πελοπόννησος, οὐκ Ἰωνία.

(4) καὶ τὸν ἀγῶνα πρῶτος ἔθηκε κατὰ ζῆλον Ἡρακλέους, ὡς δι' ἐκεῖνον Ὀλύμπια τῷ Διΐ, καὶ δι' αὐτὸν Ἴσθμια τῷ Ποσειδῶνι φιλοτιμηθεὶς ἄγειν τοὺς Ἕλληνας. ὁ γὰρ ἐπὶ Μελικέρτῃ τεθεὶς αὐτόθι νυκτὸς ἐδρᾶτο, τελετῆς ἔχων μᾶλλον ἢ θέας καὶ πανηγυρισμοῦ τάξιν. ἔνιοι δέ φασιν ἐπὶ Σκείρωνι τὰ Ἴσθμια τεθῆναι, τοῦ Θησέως ἀφοσιουμένου τὸν φόνον διὰ τὴν συγγένειαν· Σκείρωνα γὰρ υἱὸν εἶναι Κανήθου καὶ Ἡνιόχης τῆς Πιτθέως. **(5)** οἱ δὲ Σίνιν, οὐ Σκείρωνα, καὶ τὸν ἀγῶνα τεθῆναι διὰ τοῦτον ὑπὸ Θησέως, οὐ δι' ἐκεῖνον. ἔταξεν οὖν καὶ διωρίσατο πρὸς τοὺς Κορινθίους Ἀθηναίων τοῖς ἀφικνουμένοις ἐπὶ τὰ Ἴσθμια παρέχειν προεδρίαν ὅσον ἂν τόπον ἐπίσχῃ καταπετασθὲν τὸ τῆς θεωρίδος νεὼς ἱστίον, ὡς Ἑλλάνικος (FHG I 55 fr. 76 = FGrH 4 F 165 = FGrH 323a F 15) καὶ Ἄνδρων ὁ Ἁλικαρνασεὺς (FGrH 10 F 6) ἱστορήκασιν.

26 (1) Εἰς δὲ τὸν πόντον ἔπλευσε τὸν Εὔξεινον, ὡς μὲν Φιλόχορος (FHG I 392 fr. 49 = FGrH 328 F 110) καί τινες ἄλλοι λέγουσι, μεθ' Ἡρακλέους ἐπὶ τὰς Ἀμαζόνας συστρατεύσας, καὶ γέρας Ἀντιόπην ἔλαβεν· οἱ δὲ πλείους, ὧν ἐστὶ καὶ Φερεκύδης (FGrH 3 F 151) καὶ Ἑλλάνικος (FHG I 55 fr. 76 = FGrH 4 F 166 = FGrH 323a F 16a) καὶ Ἡρόδωρος (FGrH 31 F 25a), ὕστερόν φασιν Ἡρακλέους ἰδιόστολον πλεῦσαι τὸν Θησέα καὶ τὴν Ἀμαζόνα λαβεῖν αἰχμάλωτον, πιθανώτερα λέγοντες. οὐδεὶς γὰρ ἄλλος ἱστόρηται τῶν μετ' αὐτοῦ στρατευσάντων Ἀμαζόνα λαβεῖν αἰχμάλωτον. **(2)** Βίων (FHG II 19 fr. 1 = FGrH 14 F 2 = FGrH 332 F 2) δὲ καὶ ταύτην παρακρουσάμενον οἴχεσθαι λαβόντα· φύσει γὰρ οὔσας τὰς Ἀμαζόνας φιλάνδρους οὐτε φυγεῖν τὸν Θησέα προσβάλλοντα τῇ χώρᾳ, ἀλλὰ καὶ ξένια πέμπειν· τὸν δὲ τὴν κομίζουσαν ἐμβῆναι παρακαλεῖν εἰς τὸ πλοῖον· ἐμβάσης δ' ἀναχθῆναι. Μενεκράτης (FHG II 345 fr. 8 = FGrH 701 F 1) δέ τις, ἱστορίαν περὶ Νικαίας τῆς ἐν Βιθυνίᾳ πόλεως ἐκδεδωκώς, Θησέα φησὶ τὴν Ἀντιόπην ἔχοντα διατρῖψαι περὶ τούτους τοὺς τόπους· **(3)** τυγχάνειν δὲ συστρατεύοντας αὐτῷ τρεῖς νεανίσκους ἐξ Ἀθηνῶν ἀδελφοὺς ἀλλήλων, Εὔνεων καὶ Θόαντα καὶ Σολόεντα. τοῦτον οὖν

Figura 3 – Plutarco, Teseu 24-28 (Perrin) (continuação)

ἐρῶντα τῆς Ἀντιόπης καὶ λανθάνοντα τοὺς ἄλλους, ἐξειπεῖν πρὸς ἕνα τῶν συνήθων· ἐκείνου δὲ περὶ τούτων ἐντυχόντος τῇ Ἀντιόπῃ, τὴν μὲν πεῖραν ἰσχυρῶς ἀποτρίψασθαι, τὸ δὲ πρᾶγμα σωφρόνως ἅμα καὶ πρᾴως ἐνεγκεῖν καὶ πρὸς τὸν Θησέα μὴ κατηγορῆσαι. **(4)** τοῦ δὲ Σολόεντος ὡς ἀπέγνω ῥίψαντος ἑαυτὸν εἰς ποταμόν τινα καὶ διαφθαρέντος, ἠσθημένον τότε τὴν αἰτίαν καὶ τὸ πάθος τοῦ νεανίσκου τὸν Θησέα βαρέως ἐνεγκεῖν, καὶ δυσφοροῦντα λόγιόν τι πυθόχρηστον ἀνενεγκεῖν πρὸς ἑαυτόν· εἶναι γὰρ αὐτῷ προστεταγμένον ἐν Δελφοῖς ὑπὸ τῆς Πυθίας (Parke-Wormell 2.411), ὅταν ἐπὶ ξένης ἀνιαθῇ μάλιστα καὶ περίλυπος γένηται, πόλιν ἐκεῖ κτίσαι καὶ τῶν ἀμφ' αὑτόν τινας ἡγεμόνας καταλιπεῖν. **(5)** ἐκ δὲ τούτου τὴν μὲν πόλιν, ἣν ἔκτισεν, ἀπὸ τοῦ θεοῦ Πυθόπολιν προσαγορεῦσαι, Σολόεντα δὲ τὸν πλησίον ποταμὸν ἐπὶ τιμῇ τοῦ νεανίσκου. καταλιπεῖν δὲ καὶ τοὺς ἀδελφοὺς αὐτοῦ, οἷον ἐπιστάτας καὶ νομοθέτας, καὶ σὺν αὐτοῖς Ἕρμον ἄνδρα τῶν Ἀθήνησιν εὐπατριδῶν· ἀφ' οὗ καὶ τόπον Ἑρμοῦ καλεῖν οἰκίαν τοὺς Πυθοπολίτας, οὐκ ὀρθῶς τὴν δευτέραν συλλαβὴν περισπῶντας καὶ τὴν δόξαν ἐπὶ θεὸν ἀπὸ ἥρωος μετατιθέντας.

27 (1) Πρόφασιν μὲν οὖν ταύτην ὁ τῶν Ἀμαζόνων πόλεμος ἔσχε· φαίνεται δὲ μὴ φαῦλον αὐτοῦ μηδὲ γυναικεῖον γενέσθαι τὸ ἔργον. οὐ γὰρ ἂν ἐν ἄστει κατεστρατοπέδευσαν οὐδὲ τὴν μάχην συνῆψαν ἐν χρῷ περὶ τὴν Πνύκα καὶ τὸ Μουσεῖον, εἰ μὴ κρατοῦσαι τῆς χώρας ἀδεῶς τῇ πόλει προσέμιξαν. **(2)** εἰ μὲν οὖν, ὡς Ἑλλάνικος (FGrH 4 F 167a = FGrH 323a F 17a) ἱστόρηκε, τῷ Κιμμερικῷ Βοσπόρῳ παγέντι διαβᾶσαι περιῆλθον, ἔργον ἐστὶ πιστεῦσαι· τὸ δ' ἐν τῇ πόλει σχεδὸν αὐτὰς ἐνστρατοπεδεῦσαι μαρτυρεῖται καὶ τοῖς ὀνόμασι τῶν τόπων καὶ ταῖς θήκαις τῶν πεσόντων. Πολὺν δὲ χρόνον ὄκνος ἦν καὶ μέλλησις ἀμφοτέροις τῆς ἐπιχειρήσεως· τέλος δὲ Θησεὺς κατά τι λόγιον τῷ Φόβῳ σφαγιασάμενος συνῆψεν αὐταῖς. **(3)** ἡ μὲν οὖν μάχη Βοηδρομιῶνος ἐγένετο μηνός ἐφ' ᾗ τὰ Βοηδρόμια μέχρι νῦν Ἀθηναῖοι θύουσιν. ἱστορεῖ δὲ Κλείδημος (FHG I 360 fr. 6 = FGrH 323 F 18), ἐξακριβοῦν τὰ καθ' ἕκαστα βουλόμενος, τὸ μὲν εὐώνυμον τῶν Ἀμαζόνων κέρας ἐπιστρέφειν πρὸς τὸ νῦν καλούμενον Ἀμαζόνειον, τῷ δὲ δεξιῷ πρὸς τὴν Πνύκα κατὰ τὴν Χρύσαν ἥκειν. μάχεσθαι δὲ πρὸς τοῦτο τοὺς Ἀθηναίους ἀπὸ τοῦ Μουσείου ταῖς Ἀμαζόσι συμπεσόντας, καὶ τάφους τῶν πεσόντων περὶ τὴν πλατεῖαν εἶναι τὴν φέρουσαν ἐπὶ τὰς πύλας παρὰ τὸ Χαλκώδοντος ἡρῷον, ἃς νῦν Πειραϊκὰς ὀνομάζουσι. **(4)** καὶ ταύτῃ μὲν ἐκβιασθῆναι μέχρι τῶν Εὐμενίδων καὶ ὑποχωρῆσαι ταῖς γυναιξίν, ἀπὸ δὲ Παλλαδίου καὶ Ἀρδηττοῦ καὶ Λυκείου προσβαλόντας ὥσασθαι τὸ δεξιὸν αὐτῶν ἄχρι τοῦ στρατοπέδου καὶ πολλὰς καταβαλεῖν. τετάρτῳ δὲ μηνὶ συνθήκας γενέσθαι διὰ τῆς Ἱππολύτης· Ἱππολύτην γὰρ οὗτος ὀνομάζει τὴν τῷ Θησεῖ συνοικοῦσαν, οὐκ Ἀντιόπην. Ἔνιοι δέ φασι μετὰ τοῦ Θησέως μαχομένην πεσεῖν τὴν ἄνθρωπον ὑπὸ Μολπαδίας ἀκοντισθεῖσαν, καὶ τὴν στήλην τὴν παρὰ τὸ τῆς Ὀλυμπίας ἱερὸν ἐπὶ ταύτῃ κεῖσθαι. **(5)** καὶ θαυμαστὸν οὐκ ἔστιν ἐπὶ πράγμασιν οὕτω παλαιοῖς πλανᾶσθαι τὴν ἱστορίαν, ἐπεὶ καὶ τὰς τετρωμένας φασὶ τῶν Ἀμαζόνων ὑπ' Ἀντιόπης εἰς Χαλκίδα λάθρα διαπεμφθείσας τυγχάνειν ἐπιμελείας, καὶ ταφῆναί τινας ἐκεῖ περὶ τὸ νῦν

Figura 4 – Plutarco, Teseu 24-28 (Perrin) (continuação)

Ἀμαζόνειον καλούμενον. ἀλλὰ τοῦ γε τὸν πόλεμον εἰς σπονδὰς τελευτῆσαι μαρτύριόν ἐστιν ἥ τε τοῦ τόπου κλῆσις τοῦ παρὰ τὸ Θησεῖον, ὄνπερ Ὁρκωμόσιον καλοῦσιν, ἥ τε γινομένη πάλαι θυσία ταῖς Ἀμαζόσι πρὸ τῶν Θησείων. **(6)** δεικνύουσι δὲ καὶ Μεγαρεῖς Ἀμαζόνων θήκην παρ' αὐτοῖς, ἐπὶ τὸν καλούμενον Ῥοῦν βαδίζουσιν ἐξ ἀγορᾶς, ὅπου τὸ Ῥομβοειδές. λέγεται δὲ καὶ περὶ Χαιρώνειαν ἑτέρας ἀποθανεῖν, καὶ ταφῆναι παρὰ τὸ ῥευμάτιον ὃ πάλαι μὲν, ὡς ἔοικε, Θερμώδων, Αἵμων δὲ νῦν καλεῖται· περὶ ὧν ἐν τῷ Δημοσθένους βίῳ (Plut. *Dem.* 19.2) γέγραπται. φαίνονται δὲ μηδὲ Θεσσαλίαν ἀπραγμόνως αἱ Ἀμαζόνες διελθοῦσαι· τάφοι γὰρ αὐτῶν ἔτι καὶ νῦν δείκνυνται περὶ τὴν Σκοτουσαίαν καὶ τὰς Κυνὸς κεφαλάς.

28 (1) Ταῦτα μὲν οὖν ἄξια μνήμης περὶ τῶν Ἀμαζόνων. ἦν γὰρ ὁ τῆς Θησηΐδος ποιητὴς (EGF 217 Kinkel) Ἀμαζόνων ἐπανάστασιν γέγραφε, Θησεῖ γαμοῦντι Φαίδραν τῆς Ἀντιόπης ἐπιτιθεμένης καὶ τῶν μετ' αὐτῆς Ἀμαζόνων ἀμυνομένων καὶ κτείνοντος αὐτὰς Ἡρακλέους, περιφανῶς ἔοικε μύθῳ καὶ πλάσματι. **(2)** τῆς δ' Ἀντιόπης ἀποθανούσης ἔγημε Φαίδραν, ἔχων υἱὸν Ἱππόλυτον ἐξ Ἀντιόπης, ὡς δὲ Πίνδαρός (F 176 Sn.-Mae.) φησι, Δημοφῶντα. τὰς δὲ περὶ ταύτην καὶ τὸν υἱὸν αὐτοῦ δυστυχίας, ἐπεὶ μηδὲν ἀντιπίπτει παρὰ τῶν ἱστορικῶν τοῖς τραγικοῖς, οὕτως ἔχειν θετέον ὡς ἐκεῖνοι πεποιήκασιν ἅπαντες.

Figura 5 – Plutarco, Teseu 24-28 (Perrin) (tradução)

24 (1) After the death of Aegeus, Theseus conceived a wonderful design, and settled all the residents of Attica in one city, thus making one people of one city out of those who up to that time had been scattered about and were not easily called together for the common interests of all, nay, they sometimes actually quarrelled and fought with each other. **(2)** He visited them, then, and tried to win them over to this project township by township and clan by clan. The common folk and the poor quicly answered to his summons; to the powerful he promised government without a king and a democracy, in which he should only be commander in war and guardian of the laws, while in all else everyone should be on an equal footing. **(3)** Some he readily persuaded to this course, and others, fearing his power, which was already great, and his boldness, chose to be persuaded rather than forced to agree to it. Accordingly, after doing away with the townhalls and council-chambers and magistracies in the several communities, and after building a common town-hall and council-chamber for all on the ground where the upper town of the present day stands, he named the city Athens, and instituted a Panathenaic festival. **(4)** He instituted the Metoecia, or Festival of Settlement, on the sixteenth day of the month Hecatombaeon, and this is still celebrated. Then, laying aside the royal power, as he had agreed, he proceeded to arrange the government, and that too with the sanction of the gods. For an oracle came to him from Delphi (Parke-Wormell 2.154), in answer to his enquiries about the city, as follows:

Figura 6 – Plutarco, Teseu 24-28 (Perrin) (tradução) (continuação)

(5) "Theuses, offspring of Aegeus, son of the daughter of Pittheus,
Many indeed the cities to which my father has given Bounds and future fates within your citadel's confines.
Therefores be not dismayed, but with firm and confident spirit
Counsel only; the bladder will traverse the sea and its surges."

And this oracle they say the Sibyl (Hendess 23) afterwards repeated to the city, when she cried:

"Bladder may be submerged; but its sinking will not be permitted."

25 (1) Desiring still further to enlarge the city, he invited all men thither on equal terms, and the phrase "Come hither all ye people," they say was a proclamation of Theseus when he established a people, as it were, of all sorts and conditions. However, he did not suffer his democracy to become disordered or confused from an indiscriminate multitude streaming into it, but was the first to separate the people into noblemen and husbandmen and handicraftsmen. **(2)** To the noblemen he committed the care of religious rites, the supply of magistrates, the teaching of the laws, and the interpretation of the will of Heaven, and for the rest of the citizens he established a balance of privilege, the noblemen being thought to excel in dignity, the husbandmen in usefulness, and the handicraftsmen in numbers. And that he was the first to show a leaning towards the multitude, as Aristotle (*Ath.Pol.* 41.2; F 384 Rose[3]) says, and gave up his absolute rule, seems to be the testimony of Homer (*Ilias* 2.547) also, in the Catalogue of Ships, where he speaks of the Athenians alone as a "people." **(3)** He also coined money, and stamped it with the effigy of an ox, either in remembrance of the Marathonian bull, or of Taurus, the general of Minos, or because he would invite the citizens to agriculture. From this coinage, they say, "ten oxen" and "a hundred oxen" came to be used as terms of valauation. Having attached the territory of Megara securely to Attica, he set up that famous pillar on the Isthmus, and carved upon it the inscription giving the territorial boundaries. It consisted of two trimeters, of which the one towards the east declared: –

"Here is not Peloponnesus, but Ionia;"

and the one towards the west: –

"Here is the Peloponnesus, not Ionia."

(4) He also instituted the games here, in emulation of Heracles, being ambitious that as the Hellenes, by that hero's appointment, celebrated Olympian games in honour of Zeus, so by his own appointment they should celebrate Isthmian games in honour of Poseidon. For the games already instituted there in honour of Melicertes were celebrated in the night, and had the form of a religious rite rather than of a spectacle and public assembly. But some say that the Isthmian games were instituted in memory of Sciron, and that Theseus thus made expiation for his murder, because of the relationship between them; for Sciron was a son of Canethus

Figura 7 – Plutarco, Teseu 24-28 (Perrin) (tradução) (continuação)

and Henioche, who was the daughter of Pittheus. **(5)** And others have it that Sinis, not Sciron, was their son, and that it was in his honour rather that the games were instituted by Theseus. However that may be, Theseus made a formal agreement with the Corinthians that they should furnish Athenian visitors to the Isthmian games with a place of honour as large as could be covered by the sail of the state galley which brought them thither, when it was stretched to its full extent. So Hellanicus (FHG I 55 fr. 76 = FGrH 4 F 165 = FGrH 323a F 15) and Andron of Halicarnassus (FGrH 10 F 6) tell us.

26 (1) He also made a voyage into the Euxine Sea, as Philochorus (FHG I 392 fr. 49 = FGrH 328 F 110) and sundry others say, on a campaign with Heracles against the Amazons, and received Antiope as a reward of his valour; but the majority of writers, including Pherecydes (FGrH 3 F 151), Hellanicus (FHG I 55 fr. 76 = FGrH 4 F 166 = FGrH 323a F 16a), and Herodorus (FGrH 31 F 25a), say that Theseus made this voyage on his own account, after the time of Heracles, and took the Amazon captive; and this is the more probable story. For it is not recorded that any one else among those who shared his expedition took an Amazon captive. **(2)** And Bion (FHG II 19 fr. 1 = FGrH 14 F 2 = FGrH 332 F 2) says that even this Amazon he took and carried off by means of a stratagem. The Amazons, he says, were naturally friendly to men, and did not fly from Theseus when he touched upon their coasts, but actually sent him presents, and he invited the one who brought them to come on board his ship; she came on board, and he put out to sea. And a certain Menecrates (FHG II 345 fr. 8 = FGrH 701 F 1), who published a history of the Bythinian city of Nicaea, says that Theseus, with Antiope on board his ship, spent some time in those parts, **(3)** and that there chanced to be with him on this expedition three young men of Athens who were brothers, Euneos, Thoas, and Soloïs. This last, he says, fell in love with Antiope unbeknown to the rest, and revealed his secret to one of his intimate friends. That friend made overtures to Antiope, who positively repulsed the attempt upon her, but treated the matter with discretion and gentleness, and made no denunciation to Theseus. **(4)** Then Soloïs, in despair, threw himself into a river and drowned himself, and Theseus, when he learned the fate of the young man, and what had caused it, was grievously disturbed, and in his distress called to mind a certain oracle which he had once received at Delphi (Parke-Wormell 2.411). For it had there been enjoined upon him by the Pythian priestess that when, in a strange land, he should be sorest vexed and full of sorrow, he should found a city there, and leave some of his followers to govern it. **(5)** For this cause he founded a city there, and called it, from the Pythian god, Pythopolis, and the adjacent river, Soloïs, in honour of the young man. And he left there the brothers of Soloïs, to be the city's presidents and law-givers, and with them Hermus, one of the noblemen of Athens. From him also the

for his murder, because of the relationship between them; for Sciron was a son of Canethus

Figura 8 – Plutarco, Teseu 24-28 (Perrin) (tradução) (continuação)

Pythopolitans call a place in the city the House of Hermes, incorrectly changing the second syllable, and transferring the honour from a hero to a god.

27 (1) Well, then, such were the grounds for the war of the Amazons, which seems to have been no trivial nor womanish enterprise for Theseus. For they would not have pitched their camp within the city, nor fought hand to hand battles in the neighbourhood of the Pnyx and the Museum, had they not mastered the sorrounding country and approached the city with impunity. **(2)** Whether, now, as Hellanicus (FGrH 4 F 167a = FGrH 323a F 17a) writes, they came round by the Cimmerian Bosporus, which they crossed on the ice, may be doubted; but the fact that they encamped almost in the heart of the city is attested both by the names of the localities there and by the graves of those who fell in battle. Now for a long time there was hesitation and delay on both sides in making the attack, but finally Thesesus, after sacrificing to Fear, in obedience to an oracle, joined battle with the women. **(3)** This battle, then, was fought on the day of the month Boëdromion on which, down to the present time, the Athenians celebrate the Boëdromia. Cleidemus (FHG I 360 fr. 6 = FGrH 323 F 18), who wishes to be minute, writes that the left wing of the Amazons extended to what is now called the Amazoneum, and that with their right they touched the Pnyx at Chrysa; that with this left wing the Athenians fought, engaging the Amazons from the Museum, and that the graves of those who fell are on either side of the street which leads to the gate by the chapel of Chalcodon, which is now called the Peiraïe gate. **(4)** Here, he says, the Athenians were routed and driven back by the women as far as the shrine of the Eumenides, but those who attacked the invaders from the Palladium and Ardettus and the Lyceum, drove their right wing back as far as their camp, and slew many of them. And after three months, he says, a treaty of peace was made through the agency of Hippolyta; for Hippolyta is the name which Cleidemus gives to the Amazon whom Theseus married, not Antiope. But some say that the woman was slain with a javelin by Molpadia, while fighting at Theseus' side, and that the pillar which stands by the sanctuary of Olympian Earth was set up in her memory. **(5)** And it is not astonishing that history, when dealing with events of such great antiquity, should wander in uncertainty, indeed, we are also told that the wounded Amazons were secretly sent away to Chalcis by Antiope, and were nursed there, and some were buried there, near what is now called the Amazoneum. But that the war ended in a solemn treaty is attested not only by the naming of the place adjoining the Theseum, which is called Horcomosium, but also by the sacrifice which, in ancient times, was offered to the Amazons before the festival of Theseus. **(6)** And the Megarians, too, show a place in their country where Amazons were buried, on the way from the market-place to the place called Rhus, where the Rhomboid stands. And it is said, likewise, that others of them died near Chaeroneia, and were buried on the banks of the little stream

Figura 9 – Plutarco, Teseu 24-28 (Perrin) (tradução) (continuação)

which, in ancient times, as it seems, was called Thermodon, but nowadays, Haemon; concerning which names I have written in my Life of Demosthenes (Plut. *Dem.* 19.2). It appears also that not even Thessaly was traversed by the Amazons without opposition, for Amazonian graves are to this day shown in the vicinity of Scotussa and Cynoschephalae.

28 (1) So much, then, is worthy of mention regarding the Amazons. For the "Insurrection of the Amazons," written by the author of the Theseid (EGF 217 Kinkel), telling how, when Theseus married Phaedra, Antiope and the Amazons who fought to avenge her attacked him, and were slain by Heracles, has every appearance of fable and invention. **(2)** Thesues did, indeed, marry Phaedra, but this was after the death of Antiope, and he had a son by Antiope, Hippolytus, or, as Pindar (F 176 Sn.-Mae.) says, Demophoön. As for the calamities which befell Phaedra and the son of Theseus by Antiope, since there is no conflict here between historians and tragic poets, we must suppose that they happened as represented by the poets uniformly.

(trans. Perrin)

Figura 10 – Tópicos e fontes citadas por Plutarco em *Teseu* 24-25

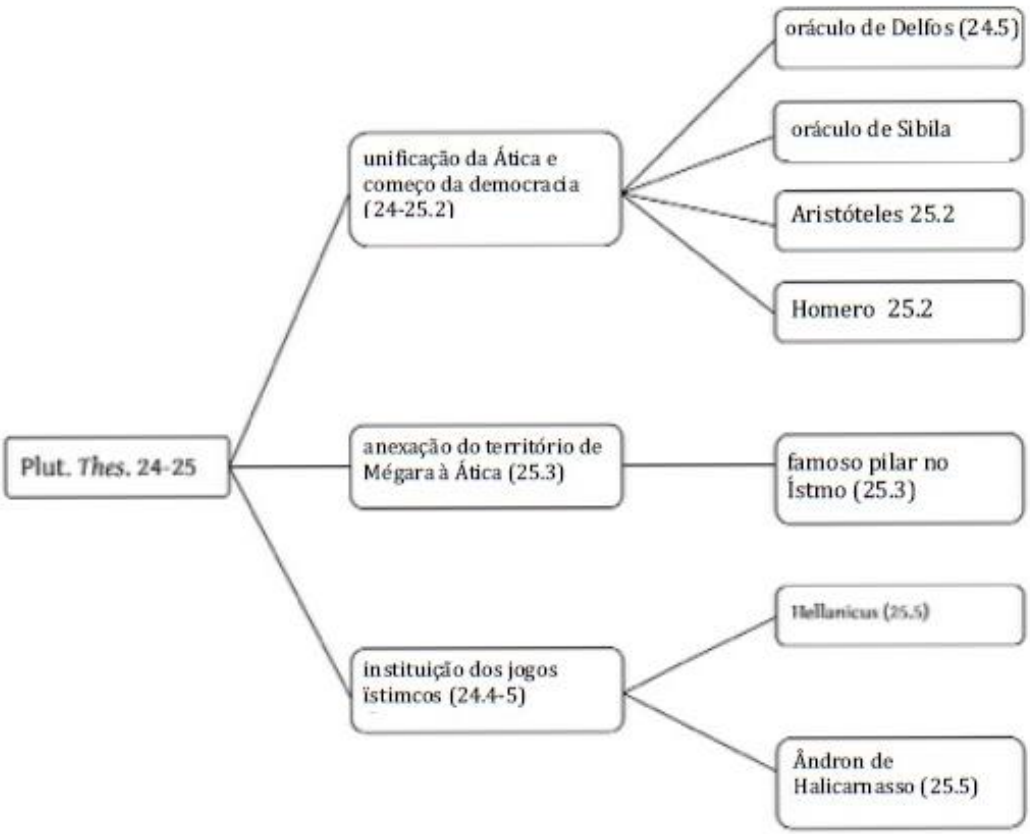
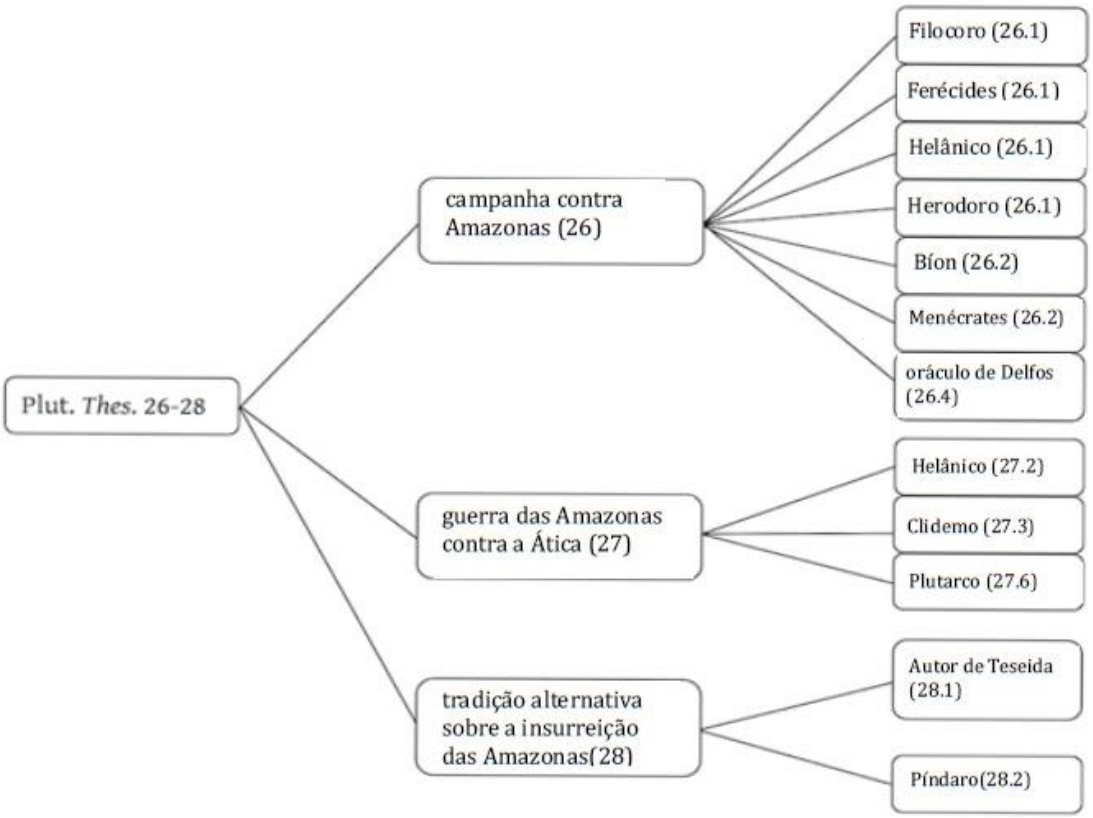


Figura 11 – Tópicos e fontes citadas por Plutarco em *Teseu* 26-28



Apêndice 2

Neste apêndice, publicamos três exemplos do *Deipnosophistae,* onde Ateneu cita fragmentos de textos sobreviventes de Heródoto, Tucídides e Xenofonte. Os textos são publicados de acordo com as seguintes edições: *Athenaei Naucratitae Dipnosophistarum Libri XV*, rec. G. Kaibel. Vol. I. Lipsiae 1887 e vol. III Lipsiae 1890; *Herodotus IV* (Livros VIII-IX), ed. AD Godley. Cambridge, Ma 1969; Tucídides. *History of the Peloponnesian War II* (Books III-IV), ed. C. F. Smith. Cambridge, Ma 1958; *Xenophon. Memorabilia. Oeconomicus. Symposium. Apology*, ed. E. C. Marchant. Cambridge, Ma 1923.

Figura 12 – Ateneu, Deipnosophistae 4.15 (138b-d) & Heródoto 9.82

Ath. Deipn. 4.15 (138b-d) ἐξῆς δὲ λεκτέον καὶ περὶ τῶν Λακωνικῶν συμποσίων. Ἡρόδοτος μὲν οὖν ἐν τῇ ἑνάτῃ τῶν ἱστοριῶν (9.82) περὶ τῆς Μαρδονίου παρασκευῆς λέγων καὶ μνημονεύσας Λακωνικῶν συμποσίων φησί· 'Ξέρξης φεύγων ἐκ τῆς Ἑλλάδος Μαρδονίῳ τὴν παρασκευὴν κατέλιπε τὴν αὑτοῦ. Παυσανίαν οὖν ἰδόντα τὴν τοῦ Μαρδονίου παρασκευὴν χρυσῷ καὶ ἀργύρῳ καὶ παραπετάσμασι ποικίλοις κατεσκευασμένην κελεῦσαι τοὺς ἀρτοποιοὺς καὶ ὀψοποιοὺς κατὰ ταὐτὰ καθὼς Μαρδονίῳ δεῖπνον παρασκευάσαι. ποιησάντων δὲ τούτων τὰ κελευσθέντα τὸν Παυσανίαν ἰδόντα κλίνας χρυσᾶς καὶ ἀργυρᾶς ἐστρωμένας καὶ τραπέζας ἀργυρᾶς καὶ παρασκευὴν μεγαλοπρεπῆ δείπνου ἐκπλαγέντα τὰ προκείμενα κελεῦσαι ἐπὶ γέλωτι τοῖς ἑαυτοῦ διακόνοις παρασκευάσαι Λακωνικὸν δεῖπνον. καὶ παρασκευασθέντος γελάσας ὁ Παυσανίας μετεπέμψατο τῶν Ἑλλήνων τοὺς στρατηγοὺς καὶ ἐλθόντων ἐπιδείξας ἑκατέρου τῶν δείπνων τὴν παρασκευὴν εἶπεν· 'ἄνδρες Ἕλληνες,	**Hdt. 9.82 (1)** Λέγεται δὲ καὶ τάδε γενέσθαι, ὡς Ξέρξης φεύγων ἐκ τῆς Ἑλλάδος Μαρδονίῳ τὴν κατασκευὴν καταλίποι τὴν ἑωυτοῦ· Παυσανίην ὦν ὁρῶντα τὴν Μαρδονίου κατασκευὴν χρυσῷ τε καὶ ἀργύρῳ καὶ παραπετάσμασι ποικίλοισι κατεσκευασμένην, κελεῦσαι τούς τε ἀρτοκόπους καὶ τοὺς ὀψοποιοὺς κατὰ ταὐτὰ καθὼς Μαρδονίῳ δεῖπνον παρασκευάζειν. **(2)** ὡς δὲ κελευόμενοι οὗτοι ἐποίευν ταῦτα, ἐνθαῦτα τὸν Παυσανίην ἰδόντα κλίνας τε χρυσέας καὶ ἀργυρέας εὖ ἐστρωμένας καὶ τραπέζας τε χρυσέας καὶ ἀργυρέας καὶ παρασκευὴν μεγαλοπρεπέα τοῦ δείπνου, ἐκπλαγέντα τὰ προκείμενα ἀγαθὰ κελεῦσαι ἐπὶ γέλωτι τοὺς ἑωυτοῦ διηκόνους παρασκευάσαι Λακωνικὸν δεῖπνον. **(3)** ὡς δὲ τῆς θοίνης ποιηθείσης ἦν πολλὸν τὸ μέσον, τὸν Παυσανίην γελάσαντα μεταπέμψασθαι τῶν Ἑλλήνων τοὺς στρατηγούς, συνελθόντων δὲ τούτων εἰπεῖν τὸν Παυσανίην, δεικνύντα ἐς ἑκατέρην τοῦ δείπνου τὴν παρασκευήν, "Ἄνδρες Ἕλληνες, τῶνδε εἵνεκα ἐγὼ ὑμέας συνήγαγον,
συνήγαγον ὑμᾶς βουλόμενος ἐπιδεῖξαι τοῦ Μήδων ἡγεμόνος τὴν ἀφροσύνην, ὃς τοιαύτην δίαιταν ἔχων ἦλθεν ὡς ἡμᾶς οὕτω ταλαίπωρον ἔχοντας.' φασὶ δέ τινες καὶ ἄνδρα Συβαρίτην ἐπιδημήσαντα τῇ Σπάρτῃ καὶ συνεστιαθέντα ἐν τοῖς φιδιτίοις εἰπεῖν· 'εἰκότως ἀνδρειότατοι ἁπάντων εἰσὶ Λακεδαιμόνιοι· ἕλοιτο γὰρ ⟨ἄν⟩ τις εὖ φρονῶν μυριάκις ἀποθανεῖν ἢ οὕτως εὐτελοῦς διαίτης μεταλαβεῖν.'	βουλόμενος ὑμῖν τοῦδε τοῦ Μήδων ἡγεμόνος τὴν ἀφροσύνην δέξαι, ὃς τοιήνδε δίαιταν ἔχων ἦλθε ἐς ἡμέας οὕτω ὀϊζυρὴν ἔχοντας ἀπαιρησόμενος." ταῦτα μὲν Παυσανίην λέγεται εἰπεῖν πρὸς τοὺς στρατηγοὺς τῶν Ἑλλήνων.

Figura 13 – Ateneu, Deipnosophistae 4.15 (138b-d) & Heródoto 9.82 (tradução)

Ath. Deipn. 4.15 (138b-d) Next we must speak also of Spartan symposia. Now Herodotus, in the ninth book of his *Histories* (9.82), speaking of Mardonius' tent and mentioning by the way the Spartan banquests, says: "When Xerxes fled from Greece he left behind the royal pavilion for Mardonius. Pausanias, therefore, when he saw the tent of Mardonius adorned with gold and silver and embroidered tapestries, commanded the bakers and fancy cooks to prepare a dinner exactly as they would for Mardonius. When they had done his bidding, Pausanias, seeing the gold and silver divans spread with coverings, and silver tables and a magnificent outlay for the dinner, in amazement at what was set before him, ordered in jest his own servants to prepare a Spartan dinner. And when it was ready, Pausanias laughed and sent for the Greek generals. On their arrival he pointed to the	**Hdt. 9.82 (1)** This other story is told. Xerxes in his fight from Hellas, having left to Mardonius his own establishment, Pausanias, seeing Mardonius' establishment with its display of gold and silver and gaily-coloured tapestry, bade the bakers and the cooks to prepare a dinner in such wise as they were wont to do for Mardonius. **(2)** They did his bidding; whereat Pausanias, when he saw golden and silvern couches richly covered, and tables of gold and silver, and all the magnificent service of the banquet, was amazed at the splendour before him, and for a jest bade his own servants prepare a dinner after Laconian fashion. **(3)** When that meal was ready and was far different from the other, Pausanias fell a-laughing, and sent for the generals of the Greeks. They being assembled, Pausanias pointed to the fashion after which either dinner was served, and said: "Men of Hellas, I have brought you
preparations made for each of the dinners and said: 'Men of Greece, I have gathered you together because I wish to show you the folly of the Median commander who, with all his luxury of living, came to attack us who are so poor.'" And some say that a Sybarite who had sojourned in Sparta and had been entertained among them at their public mess remarked: 'It is no wonder that Spartans are the bravest men in the world; for anyone in his right mind would prefer to die ten thousand times rather than share in such poor living.' (trans. Gulick)	hither because I desired to show you the foolishness of the leader of the Medes; who, with such provision for life as you see, came hither to take away from us ours, that is so pitiful." Thus, it is said, Pausanias spoke to the generals of the Greeks. (trans. Godley)

Figura 14 – Ateneu, *Deipnosophistae* 5.15 (189c) & Tucídides 4.103.1

Ath. Deipn. 5.15 (189c) ἔτι δὲ αὐλὸς μὲν τὸ ὄργανον, ὅτι διέρχεται τὸ πνεῦμα, καὶ πᾶν τὸ διατεταμένον εἰς εὐθύτητα σχῆμα αὐλὸν καλοῦμεν ὥσπερ τὸ στάδιον καὶ τὸν κρουνὸν τοῦ αἵματος· αὐτίκα δ' αὐλὸς ἀνὰ ῥῖνας παχὺς ἦλθε, καὶ τὴν περικεφαλαίαν ὅταν ἐκ τοῦ μέσου πρὸς ὀρθὸν ἀνατείνῃ αὐλῶπιν. λέγονται δὲ Ἀθήνησι καὶ ἱεροί τινες αὐλῶνες, ὧν μέμνηται Φιλόχορος (FHG I 409 fr. 147 = FGrH 328 F 68) ἐν τῇ ἐνάτῃ. καλοῦσι δ' ἀρσενικῶς τοὺς αὐλῶνας, ὥσπερ Θουκυδίδης (4.103.1) ἐν τῇ δ' καὶ πάντες οἱ καταλογάδην συγγραφεῖς, οἱ δὲ ποιηταὶ θηλυκῶς.	**Thuc. 4.103 (1)** Ἐπὶ ταύτην οὖν ὁ Βρασίδας ἄρας ἐξ Ἀρνῶν τῆς Χαλκιδικῆς ἐπορεύετο τῷ στρατῷ. καὶ ἀφικόμενος περὶ δείλην ἐπὶ τὸν Αὐλῶνα καὶ Βορμίσκον, ᾗ ἡ Βόλβη λίμνη ἐξίησιν ἐς θάλασσαν, καὶ δειπνοποιησάμενος ἐχώρει τὴν νύκτα (...)

Figura 15 – Ateneu, *Deipnosophistae* 5.15 (189c) & Tucídides 4.103.1 (tradução)

Ath. Deipn. 5.15 (189c) Again there is the instrument called *aulos*, because the air goes through it, and any figure prolonged in a straight line we call *aulos*, like a stadium, or a gush of blood: "Forthwith a thick gush came from his nostrils;" or of the helmet when it extends straight up from the middle we say that it is "tube-like." At Athens there are certain "sacred hollows" (*aulones*), as they are called, which Philochorus (FHG I 409 fr. 147 = FGrH 328 F 68) mentions in the ninth book. The noun meaning "hollows" is masculine, as in Thucydides, Book iv. (4.103.1), and all the historians who write in prose; but in the poets it is feminine. (trans. Gulick)	**Thuc. 4.103 (1)** Against this place Brasidas marched with his army, setting out from Arnae in Chalcidice. Arriving about dusk at Aulon and Bromiscus, where the lake Bolbe has its oulet into the sea, he took supper and then proceeded by night (...) (trans. Smith)

Figura 16 – Ateneu, *Deipnosophistae* 13.54 (588d) & Xenophon, Memorabilia 3.11.1

Ath. Deipn. 13.54 (588d) τὸ δ' αὐτὸ καὶ Σωκράτης ἐμαντεύσατο περὶ Θεοδότης τῆς Ἀθηναίας, ὥς φησι Ξενοφῶν ἐν Ἀπομνημονεύμασιν (3.11.1)· "ὅτι δὲ καλλίστη εἴη καὶ στέρνα κρείττω λόγου παντὸς ἔχοι λέγοντός τινος, 'ἰτέον ἡμῖν, ἔφη, θεασομένοις τὴν γυναῖκα· οὐ γὰρ δὴ ἀκούουσιν ἔστιν κρῖναι τὸ κάλλος.' "	**Xen. Mem. 3.11 (1)** Γυναικὸς δέ ποτε οὔσης ἐν τῇ πόλει καλῆς, ᾗ ὄνομα ἦν Θεοδότη, καὶ οἵας συνεῖναι τῷ πείθοντι, μνησθέντος αὐτῆς τῶν παρόντων τινὸς καὶ εἰπόντος ὅτι κρεῖττον εἴη λόγου τὸ κάλλος τῆς γυναικός, καὶ ζωγράφους φήσαντος εἰσιέναι πρὸς αὐτὴν ἀπεικασομένους, οἷς ἐκείνην ἐπιδεικνύειν ἑαυτῆς ὅσα καλῶς ἔχοι, Ἰτέον ἂν εἴη θεασομένους, ἔφη ὁ Σωκράτης· οὐ γὰρ δὴ ἀκούσασί γε τὸ λόγου κρεῖττον ἔστι καταμαθεῖν. καὶ ὁ διηγησάμενος, Οὐκ ἄν φθάνοιτ', ἔφη, ἀκολουθοῦντες.

Figura 17 – Ateneu, *Deipnosophistae* 13.54 (588d) & Xenophon, Memorabilia 3.11.1 (tradução)

Ath. Deipn. 13.54 (588d) Socrates, also, divined the same promise in the case of Theodote of Athens, as Xenophon says in his *Memorabilia* (3.11.1): "When someone remarked that she was very beautiful and had a bosom beyond the power of any tongue to describe Socrates said, 'We must go to see the woman; for it is not possible to judge her beauty by hearsay.' " (trans. Gulick)	**Xen. Mem. 3.11 (1)** At one time there was in Athens a beautiful woman named Theodote, who was ready to keep company with anyone who pleased her. One of the bystanders mentioned her name, declaring that words failed him to describe the lady's beauty, and adding that artists visited her to paint her portrait, and she showed them as much as decency allowed. "We had better go and see her," cried Socrates; "of course what beggars description can't very well be learned by hearsay." (trans. Marchant)

Apêndice 3

Neste apêndice, publicamos quatro exemplos de visualização de fragmentos em seu contexto original: para o texto integral consulte o Apêndice 1 e 2. Para mais informações, ver: http://demo.fragmentarytexts.org.

Figura 18 – Plutarco, *Teseu* 26 (as cores significam a extensão do fragmento de acordo com uma edição especial, tanto no texto grego quanto na tradução em inglês. O ícone do PDF é um link para a edição impressa armazenada no Google ou o Internet Archive)

26 (1) Εἰς δὲ τὸν πόντον ἔπλευσε τὸν Εὔξεινον, ὡς μὲν Φιλόχορος (FHG I 392 fr. 49 = FGrH 328 F 110) καί τινες ἄλλοι λέγουσι, μεθ' Ἡρακλέους ἐπὶ τὰς Ἀμαζόνας συστρατεύσας, καὶ γέρας Ἀντιόπην ἔλαβεν· οἱ δὲ πλείους, ὧν ἐστὶ καὶ Φερεκύδης (FGrH 3 F 151) καὶ Ἑλλάνικος (FHG I 55 fr. 76 = FGrH 4 F 166 = FGrH 323a F 16a) καὶ Ἡρόδωρος (FGrH 31 F 25a), ὕστερόν φασιν Ἡρακλέους ἰδιόστολον πλεῦσαι τὸν Θησέα καὶ τὴν Ἀμαζόνα λαβεῖν αἰχμάλωτον, πιθανώτερα λέγοντες. οὐδεὶς γὰρ ἄλλος ἱστόρηται τῶν μετ' αὐτοῦ στρατευσάντων Ἀμαζόνα λαβεῖν αἰχμάλωτον. **(2)** Βίων (FHG II 19 fr. 1 = FGrH 14 F 2 = FGrH 332 F 2) δὲ καὶ ταύτην παρακρουσάμενον οἴχεσθαι λαβόντα· φύσει γὰρ οὔσας τὰς Ἀμαζόνας φιλάνδρους οὔτε φυγεῖν τὸν Θησέα προσβάλλοντα τῇ χώρᾳ, ἀλλὰ καὶ ξένια πέμπειν· τὸν δὲ τὴν κομίζουσαν ἐμβῆναι παρακαλεῖν εἰς τὸ πλοῖον· ἐμβάσης δὲ ἀναχθῆναι. Μενεκράτης (FHG II 345 fr. 8 = FGrH 701 F 1) δέ τις, ἱστορίαν περὶ Νικαίας τῆς ἐν Βιθυνίᾳ πόλεως ἐκδεδωκώς, Θησέα φησὶ τὴν Ἀντιόπην ἔχοντα διατρῖψαι περὶ τούτους τοὺς τόπους· **(3)** τυγχάνειν δὲ συστρατεύοντας αὐτῷ τρεῖς νεανίσκους ἐξ Ἀθηνῶν ἀδελφοὺς ἀλλήλων, Εὔνεων καὶ Θόαντα καὶ Σολόεντα. τοῦτον οὖν ἐρῶντα τῆς Ἀντιόπης καὶ λανθάνοντα τοὺς ἄλλους ἐξειπεῖν πρὸς ἕνα τῶν συνήθων· ἐκείνου δὲ περὶ τούτων ἐντυχόντος τῇ Ἀντιόπῃ, τὴν μὲν πεῖραν ἰσχυρῶς ἀποτρίψασθαι, τὸ δὲ πρᾶγμα σωφρόνως ἅμα καὶ πράως ἐνεγκεῖν καὶ πρὸς τὸν Θησέα μὴ κατηγορῆσαι. **(4)** τοῦ δὲ Σολόεντος ὡς ἀπέγνω ῥίψαντος ἑαυτὸν εἰς ποταμόν τινα καὶ διαφθαρέντος, ᾐσθημένον τότε τὴν αἰτίαν καὶ τὸ πάθος τοῦ νεανίσκου τὸν Θησέα βαρέως ἐνεγκεῖν, καὶ δυσφοροῦντα λόγιόν τι πυθόχρηστον ἀνενεγκεῖν πρὸς ἑαυτόν· εἶναι γὰρ αὐτῷ προστεταγμένον ἐν Δελφοῖς ὑπὸ τῆς Πυθίας (Parke-Wormell 2.411), ὅταν ἐπὶ ξένης ἀνιαθῇ μάλιστα καὶ περίλυπος γένηται, πόλιν ἐκεῖ κτίσαι καὶ τῶν ἀμφ' αὐτόν τινας ἡγεμόνας καταλιπεῖν. **(5)** ἐκ δὲ τούτου τὴν μὲν πόλιν, ἣν ἔκτισεν, ἀπὸ τοῦ θεοῦ Πυθόπολιν προσαγορεῦσαι, Σολόεντα δὲ τὸν πλησίον ποταμὸν ἐπὶ τιμῇ τοῦ νεανίσκου. καταλιπεῖν δὲ καὶ τοὺς ἀδελφοὺς αὐτοῦ, οἷον ἐπιστάτας καὶ νομοθέτας, καὶ σὺν αὐτοῖς Ἕρμον ἄνδρα τῶν Ἀθήνησιν εὐπατριδῶν· ἀφ' οὗ καὶ τόπον Ἑρμοῦ καλεῖν οἰκίαν τοὺς Πυθοπολίτας, οὐκ ὀρθῶς τὴν δευτέραν συλλαβὴν περισπῶντας καὶ τὴν δόξαν ἐπὶ θεὸν ἀπὸ ἥρωος μετατιθέντας.

26 (1) He also made a voyage into the Euxine Sea, as Philochorus (FHG I 392 fr. 49 = FGrH 328 F 110) and sundry others say, on a campaign with Heracles against the Amazons, and received Antiope as a reward of his valour; but the majority of writers, including Pherecydes (FGrH 3 F 151), Hellanicus (FHG I 55 fr. 76 = FGrH 4 F 166 = FGrH 323a F 16a), and Herodorus (FGrH 31 F 25a), say that Theseus made this voyage on his own account, after the time of Heracles, and took the Amazon captive; and this is the more probable story. For it is not recorded that any one else among those who shared his expedition took an Amazon captive. **(2)** And Bion (FHG II 19 fr. 1 = FGrH 14 F 2 = FGrH 332 F 2) says that even this Amazon he took and carried off by means of a stratagem. The Amazons, he says, were naturally friendly to men, and did not fly from Theseus when he touched upon their coasts, but actually sent him presents, and he invited the one who brought them to come on board his ship; she came on board, and he put out to sea. And a certain Menecrates (FHG II 345 fr. 8 = FGrH 701 F 1), who published a history of the Bythinian city of Nicaea, says that Theseus, with Antiope on board his ship, spent some time in those parts, **(3)** and that there chanced to be with him on this expedition three young men of Athens who were brothers, Euneos, Thoas, and Soloïs. This last, he says, fell in love with Antiope unbeknown to the rest, and revealed his secret to one of his intimate friends. That friend made overtures to Antiope, who positively repulsed the attempt upon her, but treated the matter with discretion and gentleness, and made no denunciation to Theseus. **(4)** Then Soloïs, in despair, threw himself into a river and drowned himself, and Theseus, when he learned the fate of the young man, and what had caused it, was grievously disturbed, and in his distress called to mind a certain oracle which he had once received at Delphi (Parke-Wormell 2.411). For it had there been enjoined upon him by the Pythian priestess that when, in a strange land, he should be sorest vexed and full of sorrow, he should found a city there, and leave some of his followers to govern it. **(5)** For this cause he founded a city there, and called it, from the Pythian god, Pythopolis, and the adjacent river, Soloïs, in honour of the young man. And he left there the brothers of Soloïs, to be the city's presidents and law-givers, and with them Hermus, one of the noblemen of Athens. From him also the Pythopolitans call a place in the city the House of Hermes, incorrectly changing the second syllable, and transferring the honour from a hero to a god.

Os seguintes exemplos são as comparações entre *Deipnosophistae* de Ateneu e os textos preservados de Heródoto, Tucídides e Xenofonte. As cores representam as palavras correspondentes em Ateneu e na fonte citada.

Figura 19 – Ateneu, *Deipnosophistae* 4.15 (138b - d) & Heródoto 9.82

Athenaei Naucratitae Dipnosophistarum Libri XV, rec. G. Kaibel. Vol. I. Lipsiae 1887

Ath. Deipn. 4.15 (138b-d) ἐξῆς δὲ λεκτέον καὶ περὶ τῶν Λακωνικῶν συμποσίων. Ἡρόδοτος μὲν οὖν ἐν τῇ ἑνάτῃ τῶν ἱστοριῶν (9.82) περὶ τῆς Μαρδονίου παρασκευῆς λέγων καὶ μνημονεύσας Λακωνικῶν συμποσίων φησί· ‘Ξέρξης φεύγων ἐκ τῆς Ἑλλάδος Μαρδονίῳ τὴν παρασκευὴν κατέλιπε τὴν αὐτοῦ. Παυσανίαν οὖν ἰδόντα τὴν τοῦ Μαρδονίου παρασκευὴν χρυσῷ καὶ ἀργύρῳ καὶ παραπετάσμασι ποικίλοις κατεσκευασμένην κελεῦσαι τοὺς ἀρτοποιοὺς καὶ ὀψοποιοὺς κατὰ ταὐτὰ καθὼς Μαρδονίῳ δεῖπνον παρασκευάσαι. ποιησάντων δὲ τούτων τὰ κελευσθέντα τὸν Παυσανίαν ἰδόντα κλίνας χρυσᾶς καὶ ἀργυρᾶς ἐστρωμένας καὶ τραπέζας ἀργυρᾶς καὶ παρασκευὴν μεγαλοπρεπῆ δείπνου ἐκπλαγέντα τὰ προκείμενα κελεῦσαι ἐπὶ γέλωτι τοῖς ἑαυτοῦ διακόνοις παρασκευάσαι Λακωνικὸν δεῖπνον. καὶ παρασκευασθέντος γελάσας ὁ Παυσανίας μετεπέμψατο τῶν Ἑλλήνων τοὺς στρατηγοὺς καὶ ἐλθόντων ἐπιδείξας ἑκατέρου τῶν δείπνων τὴν παρασκευὴν εἶπεν· ‘ἄνδρες Ἕλληνες, συνήγαγον ὑμᾶς βουλόμενος ἐπιδεῖξαι τοῦ Μήδων ἡγεμόνος τὴν ἀφροσύνην, ὃς τοιαύτην δίαιταν ἔχων ἦλθεν ὡς ἡμᾶς οὕτω ταλαίπωρον ἔχοντας.’ φασὶ δέ τινες καὶ ἄνδρα Συβαρίτην ἐπιδημήσαντα τῇ Σπάρτῃ καὶ συνεστιαθέντα ἐν τοῖς φιδιτίοις εἰπεῖν· ‘εἰκότως ἀνδρειότατοι ἁπάντων εἰσὶ Λακεδαιμόνιοι· ἕλοιτο γὰρ ⟨ἄν⟩ τις εὖ φρονῶν μυριάκις ἀποθανεῖν ἢ οὕτως εὐτελοῦς διαίτης μεταλαβεῖν.’

Herodotus IV (Books VIII-IX), ed. A.D. Godley. Cambridge, Ma 1969

Hdt. 9.82 (1) Λέγεται δὲ καὶ τάδε γενέσθαι, ὡς Ξέρξης φεύγων ἐκ τῆς Ἑλλάδος Μαρδονίῳ τὴν κατασκευὴν καταλίποι τὴν ἑωυτοῦ· Παυσανίην ὦν ὁρῶντα τὴν Μαρδονίου κατασκευὴν χρυσῷ τε καὶ ἀργύρῳ καὶ παραπετάσμασι ποικίλοισι κατεσκευασμένην, κελεῦσαι τούς τε ἀρτοκόπους καὶ τοὺς ὀψοποιοὺς κατὰ ταὐτὰ καθὼς Μαρδονίῳ δεῖπνον παρασκευάζειν. **(2)** ὡς δὲ κελευόμενοι οὗτοι ἐποίευν ταῦτα, ἐνθαῦτα τὸν Παυσανίην ἰδόντα κλίνας τε χρυσέας καὶ ἀργυρέας εὖ ἐστρωμένας καὶ τραπέζας τε χρυσέας καὶ ἀργυρέας καὶ παρασκευὴν μεγαλοπρεπέα τοῦ δείπνου, ἐκπλαγέντα τὰ προκείμενα ἀγαθὰ κελεῦσαι ἐπὶ γέλωτι τοὺς ἑωυτοῦ διηκόνους παρασκευάσαι Λακωνικὸν δεῖπνον. **(3)** ὡς δὲ τῆς θοίνης ποιηθείσης ἦν πολλὸν τὸ μέσον, τὸν Παυσανίην γελάσαντα μεταπέμψασθαι τῶν Ἑλλήνων τοὺς στρατηγούς, συνελθόντων δὲ τούτων εἰπεῖν τὸν Παυσανίην, δεικνύντα ἐς ἑκατέρην τοῦ δείπνου τὴν παρασκευήν,“‘Ἀνδρες Ἕλληνες, τῶνδε εἵνεκα ἐγὼ ὑμέας συνήγαγον, βουλόμενος ὑμῖν τοῦδε τοῦ Μήδων ἡγεμόνος τὴν ἀφροσύνην δέξαι, ὃς τοιήνδε δίαιταν ἔχων ἦλθε ἐς ἡμέας οὕτω ὀϊζυρὴν ἔχοντας ἀπαιρησόμενος.” ταῦτα μὲν Παυσανίην λέγεται εἰπεῖν πρὸς τοὺς στρατηγοὺς τῶν Ἑλλήνων.

Figura 20 – Ateneu, *Deipnosophistae* 5.15 (189c) & Tucídides 4.103 .1

Athenaei Naucratitae Dipnosophistarum Libri XV, rec. G. Kaibel. Vol. I. Lipsiae 1887

Ath. Deipn. 5.15 (189c) ἔτι δὲ αὐλὸς μὲν τὸ ὄργανον, ὅτι διέρχεται τὸ πνεῦμα, καὶ πᾶν τὸ διατεταμένον εἰς εὐθύτητα σχῆμα αὐλὸν καλοῦμεν ὥσπερ τὸ στάδιον καὶ τὸν κρουνὸν τοῦ αἵματος· αὐτίκα δ’ αὐλὸς ἀνὰ ῥῖνας παχὺς ἦλθε, καὶ τὴν περικεφαλαίαν ὅταν ἐκ τοῦ μέσου πρὸς ὀρθὸν ἀνατείνῃ αὐλῶπιν. λέγονται δὲ Ἀθήνησι καὶ ἱεροί τινες αὐλῶνες, ὧν μέμνηται Φιλόχορος ἐν τῇ ἑνάτῃ (FHG I 409 fr. 147 = FGrH 328 F 68). καλοῦσι δ’ ἀρσενικῶς τοὺς αὐλῶνας, ὥσπερ Θουκυδίδης ἐν τῇ δ’ (4.103.1) καὶ πάντες οἱ καταλογάδην συγγραφεῖς, οἱ δὲ ποιηταὶ θηλυκῶς.

Thucydides. History of the Peloponnesian War II (Books III-IV), ed. C.F. Smith. Cambridge, Ma 1958

Thuc. 4.103 (1) Ἐπὶ ταύτην οὖν ὁ Βρασίδας ἄρας ἐξ Ἀρνῶν τῆς Χαλκιδικῆς ἐπορεύετο τῷ στρατῷ. καὶ ἀφικόμενος περὶ δείλην ἐπὶ τὸν Αὐλῶνα καὶ Βορμίσκον, ᾗ ἡ Βόλβη λίμνη ἐξίησιν ἐς θάλασσαν, καὶ δειπνοποιησάμενος ἐχώρει τὴν νύκτα. (...)

Figura 21 – Ateneu, *Deipnosophistae.* 13.54 (588d) & Xenofonte, *Memorabilia* 3.11.1

Athenaei Naucratitae Dipnosophistarum Libri XV, rec. G. Kaibel. Vol. III Lipsiae 1890

Ath. Deipn. 13.54 (588d) τὸ δ’ αὐτὸ καὶ Σωκράτης ἐμαντεύσατο περὶ Θεοδότης τῆς Ἀθηναίας, ὡς φησι Ξενοφῶν ἐν Ἀπομνημονεύμασιν (3.11.1)· “ὅτι δὲ καλλίστη εἴη καὶ στέρνα κρείττω λόγου παντὸς ἔχοι λέγοντός τινος, ‘ἰτέον ἡμῖν, ἔφη, θεασομένοις τὴν γυναῖκα· οὐ γὰρ δὴ ἀκούουσιν ἔστιν κρῖναι τὸ κάλλος.’ ”

Xenophon. Memorabilia. Oeconomicus. Symposium. Apology, ed. E.C. Marchant. Cambridge, Ma 1923

Xen. Mem. 3.11 (1) Γυναικὸς δέ ποτε οὔσης ἐν τῇ πόλει καλῆς, ᾗ ὄνομα ἦν Θεοδότη, καὶ οἵας συνεῖναι τῷ πείθοντι, μνησθέντος αὐτῆς τῶν παρόντων τινὸς καὶ εἰπόντος ὅτι κρεῖττον εἴη λόγου τὸ κάλλος τῆς γυναικός, καὶ ζωγράφους φήσαντος εἰσιέναι πρὸς αὐτὴν ἀπεικασομένους, οἷς ἐκείνην ἐπιδεικνύειν ἑαυτῆς ὅσα καλῶς ἔχοι, Ἰτέον ἂν εἴη θεασομένους, ἔφη ὁ Σωκράτης· οὐ γὰρ δὴ ἀκούσασί γε τὸ λόγου κρεῖττον ἔστι καταμαθεῖν. καὶ ὁ διηγησάμενος, Οὐκ ἂν φθάνοιτ’, ἔφη, ἀκολουθοῦντες.

CIÊNCIAS HUMANAS DIGITAIS NA SALA DE AULA – UMA ABORDAGEM TÉCNICA PARA INTEGRAÇÃO DE PLATAFORMA[123]

Bridget Almas[124]
Marie Claire Beaulieu[125]

SoSOL e CITE são duas estruturas diferentes, desenvolvidas de forma independente, para trabalhar com representações digitais de fontes antigas. Cada uma delas aborda o conjunto de problemas a partir de diferentes direções, resultando em pouca sobreposição entre o que os dois oferecem e um grande potencial para a integração.

A plataforma SoSOL foi concebida como apoio para a edição colaborativa de diferentes tipos de dados em XML sendo integrados a partir de múltiplas fontes sob a plataforma Papyri.info. Os tipos de dados suportados incluem transcrições, traduções, metadados, comentários e bibliografias, cada um aderindo ao esquema TEI/EpiDoc, mas com diferentes convenções e restrições aplicadas. As publicações compostas de um ou mais destes tipos de dados são conduzidas através de um ciclo de edição por um motor de fluxo de trabalho construído sobre um repositório *git*. O suporte para um modelo de usuário baseado em simples papel é fornecido, aproveitando a especificação OpenID delegando autenticação para Provedores de Identidade Social. Os editores podem pesquisar um catálogo de identificadores de publicação pré-estabelecidos para selecionar os itens para editar ou podem criar sua publicação. Cada usuário trabalha com as publicações em seu próprio clone do repositório *git* fonte subjacente até que estejam prontos para enviar uma publicação revista para aprovação, quando suas propostas são passadas para um conselho editorial para revisão; qualquer um deles pode ser devolvido ao editor para futuros trabalhos e correções ou finalizado e atualizado no *master branch* do repositório.

A arquitetura CITE (Collections, Indexes, and Texts, with Extensions) oferece um quadro para a digitalização de fontes textuais e para a criação de mapeamentos entre as

[123] Publicado originalmente em 2012, com o título *Digital Humanities in the Classroom – Technical Approach to Platform Integration*, por *Perseus Project Updates.* Em português, distribuído sem fins comerciais, sob permissão das autoras.

[124] Perseus Project – Department of Classics Tufts University – Medford MA – 02140 – USA – balmas@gmail.com.
[125] Perseus Project – Department of Classics Tufts University – Medford MA – 02140 – USA – marie-claire.beaulieu@tufts.edu.

fontes e os fac-símiles digitais no nível da citação. Trata-se de esquemas URN detectáveis por máquina e independentes de tecnologia para citação canônica, APIS para serviços de rede que identificam e recuperam objetos identificados por URN canônica e implementações destas API em diversas plataformas. Esta arquitetura foi desenvolvida pelo Centro de Estudos Helênicos (CHS), em parte, para permitir o trabalho do Projeto Homer Multitext (HMT). No desenvolvimento da arquitetura, a equipe do CHS pretende sustentar uma ampla gama de material fonte antigo, além de manuscritos e, com a sintaxe URN do CTS (Serviços de Textos Canônicos), pudemos expressar em um único identificador tanto a posição do trabalho em uma hierarquia FRBR semelhante quanto a posição de uma faixa contínua de nó ou nós dentro de uma obra. A sintaxe CITE URN aplica a mesma teoria para objetos não documento e suporta um esquema de citação para imagens, permitindo, em um único identificador, identificar a própria imagem e suas coordenadas específicas.

Temos várias necessidades distintas, mas relacionadas a direcionar nosso trabalho na integração destas duas plataformas de Perseus. A maior parte de nosso trabalho se concentra nos dois primeiros destes com o intuito de suporta a terceira e quarta metas em trabalho posterior.

1. Apoiar o trabalho colaborativo dos alunos, junto com o modelo do projeto HMT, permitindo assim que os alunos realizem uma pesquisa linguística concreta com um resultado tangível – a publicação de uma edição digital de seu trabalho.
2. Trabalhar não apenas com inscrições e papiros, mas com fontes textuais mais gerais, como as coleções de grego, latim e árabe da Biblioteca Digital Perseu, para os quais subconjuntos de Diretrizes de TEI, como o subconjunto TEI-Analytics (que está sendo desenvolvido pelo Projeto Abbott), são mais adequados.
3. Apoiar o trabalho em uma gama crescente de fontes históricas em vários formatos e idiomas, que incluem mais de 1.200 manuscritos medievais para os quais a Walter Art Gallery (250 MSS) e o projeto de *e-codices* suíço (900 MSS) publicaram produtos digitalizados de alta resolução sob uma licença Creative Commons.
4. Apoiar uma comunidade grande e internacional de editores digitais, incluindo estudantes, pesquisadores avançados e acadêmicos. A base de usuários

> da Biblioteca Digital Perseu nos primeiros meses de 2012 ultrapassou 300 mil usuários, com aproximadamente 10% (30.000) trabalhando diretamente com fontes gregas e latinas. A regra 90-9-1 prevê que 9% de uma comunidade on-line fará contribuições ocasionais e 1% fará a maioria das novas contribuições. Isto significaria comunidades ativas de 30.000 para Perseus como um todo e 3.000 para as coleções grega e latina.

O projeto da Professora Beaulieu de envolver os alunos no trabalho sobre antigas inscrições funerárias proporciona uma excelente oportunidade para explorar este trabalho. O trabalho de mapeamento de sua coleção de imagens para as transcrições com o intuito de produzir edições digitais aproveitando estes mapeamentos se compara, de muitas maneiras, com o trabalho do projeto HMT e é uma boa opção para os serviços CITE e APIs. Além disso, o padrão EpiDoc XML baseado em TEI a ser usado para digitalizar as inscrições já é bem suportado pela plataforma SoSOL. Podemos reutilizar grande parte da validação do XML e exibir o código do suporte de publicação de papiros em SoSOL, porém, com foco na adição de suporte para os CTS identificadores. Com esta abordagem gradual, podemos lançar as bases para o suporte final de toda a coleção de integração de textos do Perseus ao mesmo tempo em que produzimos algo com aplicação mais imediata e disponível para uso por uma comunidade de estudantes menor e controlada que pode agir, efetivamente, como testadores beta para a plataforma.

Ao seguir as metodologias ágeis de desenvolvimento, estamos adotando uma abordagem iterativa para a integração. Começamos com as seguintes bases de código:

1. um clone bifurcado do repositório *git* da base da código da plataforma SoSOL JRuby;
2. a implementação de referência de Groovy/Java/Google App Engine de CTS e CITE APIs do Projeto HMT

O primeiro resultado foi criar uma implementação protótipo que reutilizou o código SoSOL existente para as transcrições EpiDoc quase em sua totalidade por subclassificá-la e mudar apenas a estrutura dos identificadores de documentos para melhor correspondência com a sintaxe URN CTS. Também substituímos um inventário de texto CTS pelo catálogo Papyri.info. A codificação do protótipo nos abriu um meio para explorar o projeto do código da plataforma SoSOL e avaliar sua viabilidade para reutilização. O produto concreto de uma interface de usuário operacional deu aos

professores Beaulieu e Crane um meio de explorar a viabilidade do ponto de vista do usuário (estudante e revisor).

O próximo passo foi analisar se podemos também ampliar este trabalho para apoiar o *corpus* Perseus maior, que usará o esquema TEI-Analytics XML em vez de EpiDoc, e para o qual teremos de apoiar a edição colaborativa, não apenas em nível do texto completo mas também ao de uma citação ou passagem. Este último otimiza a API CTS por completo. No entanto, como o CTS é uma API de leitura, precisávamos desenvolver um conjunto de funcionalidade paralela para escrever/atualizar/excluir que poderia ser usada para atualizar e criar novas edições de textos compatíveis com o CTS. Para experimentar com isso, aumentamos a implementação baseada em XQuery das APIS do CTS do projeto Alpheios, que foi escrito pelo desenvolvedor trabalhando neste projeto. Também codificamos protótipos de extensões adicionais ao código SoSOL para trabalhar com textos e passagens que usam esquema XML TEI-A ao invés de EpiDoc e mostrar uma interface de seleção de passagem.

A conclusão destes dois produtos nos muniu com a confiança de que a integração era de fato viável e o financiamento como um projeto start-up de NEH nos permite encaminhar o trabalho além da fase de protótipo para a implementação real.

Com trabalho no protótipo, pudemos identificar alguns desafios de interoperabilidade para as duas plataformas.

Para o SoSOL, o foco tem sido em identificar e isolar os pressupostos específicos da plataforma Papyri, principalmente nas seguintes áreas:

- esquema identificador
- sistema de catalogação
- folhas de estilo para exibição
- diferentes conceitos do que se compõe uma "publicação"

Para o CTS, o desafio da integração primária até agora tem sido aumentá-lo com um sistema de Criar/Atualizar/Excluir compatível.

Os desafios também incluem a necessidade de identificar ou definir um esquema de citação canônico para as inscrições, embora isso não seja especificamente uma questão de integração plataforma, mas uma questão mais geral, relacionada com a criação de edições digitais.

O primeiro resultado da fase de implementação do projeto foi integrar o código

do protótipo com o *branch master* do repositório SoSOL, que continuou a evoluir durante os nossos esforços de criação de protótipo e com o qual o nosso clone bifurcado não estava sincronizado. Através deste processo, pudemos nos beneficiar de várias melhorias feitas no código SoSOL nesse ínterim e reduzir a quantidade de mudanças necessárias para a base do código principal para apoiar os novos dados e tipos de identificadores. Este processo também exigiu reescrever significativamente o código do protótipo, o que não surpreendeu, pois a criação de um código de qualidade de produção não era o objetivo principal do protótipo. Atualmente, estamos trabalhando em um *branch* do repositório mestre SoSOL, ao invés de um fork, e esperamos poder reintegrar o código ramificado ao *branch master* muito em breve.

Após a conclusão do projeto acima, o próximo objetivo foi implantar o SoSOL e serviços CTS em um servidor Perseus com uma interface de funcionamento que a Professora Beaulieu e seus assistentes pudessem usar para selecionar uma inscrição na qual trabalhar e, em seguida, inserir no XML para a transcrição, tradução e comentário da inscrição. Este objetivo foi cumprido e eles concluíram a criação de uma transcrição digital e tradução do epigrama Nedymos através da interface SoSOL.

Inicialmente, havíamos planejado incluir também a integração com a ferramenta ImageJ nesta iteração, porém, o desenvolvimento no período interveniente, pela HMT, de uma ferramenta Image Citation melhor para trabalhar com imagens, junto com a adoção crescente da especificação de Modelo de Dados Open Annotation Core (OAC) para anotações, nos fez mudar de curso nesta parte do projeto. Começamos o trabalho de integração da ferramenta Image Citation à interface SoSOL e que agora pode ser usada a partir desta interface para selecionar uma área de interesse em uma imagem e criar um CITE URN para a seleção ao se editar ou ver a transcrição. No momento, estamos usando uma planilha compartilhada no Google Drive para gravar essas urnas e as urnas CTS correspondentes para o texto mapeado em um índice. O próximo passo será a ferramenta SoSOL para gravar e armazenar automaticamente esses mapeamentos como anotações no texto na forma de triplos de OAC RDF.

A implantação e o uso da interface SoSOL para a inscrição nos deram uma melhor compreensão do fluxo do trabalho real de que precisaremos apoiar para o trabalho nas inscrições e revelaram algumas diferenças entre este fluxo de trabalho e aquele atualmente suportado pela plataforma SoSOL para o trabalho Papyrological. Entre outros,

identificamos a necessidade de tomar algumas decisões sobre a forma como queremos lidar com o comentário e bibliografia para as inscrições e também reconhecemos a necessidade de algumas alterações de projeto na interface introduzida pela abordagem CTS de manter as traduções em documentos separados das edições de origem. Estas alterações serão incluídas na próxima iteração, durante a qual também começaremos a trabalhar em adicionar suporte a imagem de armazenamento para a mapeamentos de texto como anotações OAC e continuar a avançar com o apoio para TEI-Analytics e edição baseada em citação, que será necessária para a *corpus* Perseu maior.

Após usar estas ferramentas para produzir o XML e os dados de mapeamento de imagem para a inscrição Nedymos, agora podemos avaliar os requisitos para a exibição final da edição digital. Usamos a implementação de referência baseada em Groovy de um navegador cópia a partir do projeto HMT e os *plug-ins* do navegador Alpheios para experimentar com as opções e produzir capturas de tela através das quais podemos rever e discutir os requisitos de uma maneira concreta. Na próxima iteração, decidiremos quanto a uma abordagem de implementação para o código de exibição e de suporte à integração automática da tela e ambientes de edição.

UMA ABORDAGEM DO VIZINHO MAIS PRÓXIMO PARA A ANÁLISE AUTOMÁTICA DA MORFOLOGIA DO GREGO ANTIGO[126]

John Lee[127]

Introdução

A civilização grega antiga, da qual o mundo ocidental recebeu muito da sua herança, tem recebido devidamente uma quantidade significativa de atenção no meio acadêmico. A fim de obter-se uma compreensão mais aprofundada da civilização, é indispensável o acesso a dissertações, poemas e outros documentos gregos na língua original.

O grego antigo é uma língua indo-europeia[128] altamente flexionada. Um verbo, por exemplo, é flexionado de acordo com sua pessoa, número, voz, tempo/aspecto e modo. De acordo com Crane (1991), "um único verbo poderia ter, aproximadamente, 1000 formas e, se considerarmos que nenhum verbo possa ser precedido por até três prefixos distintos, o número de formas explode para, aproximadamente, 5.000.000". As flexões são realizadas pelos prefixos e sufixos inseridos no radical, e às vezes, promovendo mudanças na grafia dentro do radical. Essas formas numerosas podem tornar-se mais complicadas pelos acentos e por mudanças na grafia nos morfemas finais por questões fonológicas. O efeito total pode produzir uma forma flexionada em que a raiz[129] mal pode ser reconhecida.

[126] Publicado originalmente em 2008, com o título *A Nearest-Neighbor Approach to the Automatic Analysis of Ancient Greek Morphology*, na *12th CoNLL*. Em português, distribuído sem fins comerciais, sob permissão do autor e licença *Creative Commons* BY-NC-SA 3.0.

[127] Spoken Language Systems – Computer Science and Artificial Intelligence Laboratory Cambridge – Universidade de Harvard – MIT – MA 02139 – Cambridge – Massachusetts – USA – jsylee@csail.mit.edu

[128] Todas as palavras gregas estão transcritas no alfabeto romano neste artigo. Os acentos agudo, grave e circunflexo são representados pelos diacríticos como em *ó*, *ò* e *õ*, respectivamente. O espírito doce é omitido, enquanto o rude é assinalado pelo *h*. Barras subscritas usadas em *e* e *o* representam *eta* e *ômega*.

[129] A raiz é também chamada de "base" ou forma de "pesquisa lexical", uma vez que é a forma convencionalmente usada como entrada no dicionário. No caso de verbos do grego antigo, a forma da raiz é a forma da primeira pessoa do singular, do indicativo presente ativo (cf. Para o inglês é o infinitivo). Para substantivos, a forma usada é o nominativo singular. No caso de adjetivo, é o nominativo, singular, masculino.

De fato, um exercício básico para estudantes de grego antigo é identificar a forma raiz de um verbo flexionado. Essa habilidade é essencial; sem conhecer a forma da raiz, não se pode compreender o significado da palavra ou sequer procurá-la no dicionário.

Para estudiosos das clássicas, essas inúmeras formas também colocam desafios formidáveis. A fim de procurar ocorrências de uma palavra em um *corpus,* todas as suas formas devem ser enumeradas, uma vez que as palavras não aparecem frequentemente nas suas formas de raiz. Esse procedimento torna-se extremamente laborioso para pequenas palavras que coincidem com outras mais comuns (CRANE, 1991).

A análise morfológica automática do grego antigo seria útil tanto para propósitos educacionais quanto de pesquisa. De fato, um dos primeiros analisadores foi desenvolvido como ferramenta pedagógica (PACARD, 1973). Hoje, um analisador amplamente usado está embutido na *Perseus Digital Library* (CRANE, 1996), um recurso da internet usado tanto por estudantes quanto por pesquisadores.

Este artigo apresenta um analisador do grego antigo que infere a forma da raiz de uma palavra. Ele traz duas inovações. Primeiramente, utiliza uma **abordagem do vizinho mais próximo**, que não requer regras manuais e fornece analogias para facilitar a aprendizagem da máquina.

Tabela 1 – Tabela paradigma para o verbo no presente indicativo ativo que usa como exemplo o verbo *lúo* (soltar), mostrando as flexões de acordo com a pessoa e número

Pessoa/Número	Forma	Pessoa/Número	Forma
1 sing	*lú<u>o</u>*	1 plural	*lúomen*
2 sing	*lúeis*	2 plural	*lúete*
3 sing	*lúei*	3 plural	*lúousi(n)*

Em segundo lugar e talvez de modo mais significativo, **ele explora um corpus amplo e não etiquetado** para melhor a previsão de suas raízes novas.

O restante do artigo está organizado do seguinte modo. Fornecemos, primeiramente, a motivação para essas inovações (§2) e resumimos pesquisas anteriores sobre análise morfológica (§3). Então, descrevemos os dados e nossas adaptações para a **abordagem do vizinho mais próximo** seguidos dos resultados das avaliações (§7).

Inovações

Uso da Analogia e do Vizinho mais Próximo

Normalmente, espera-se que um estudante de grego antigo memorize uma série de "paradigmas", como aquele mostrado na tabela 1, que podem ocupar várias páginas de um livro de gramática. Embora a tabela do paradigma mostre a flexão de um verbo em particular, *lúo* (soltar), o aluno precisa aplicar os padrões a outros verbos. Na prática, ao invés de abstrair os padrões, muitos estudantes simplesmente memorizam esses verbos "paradigmáticos" para serem usados como analogias, a fim de identificar a forma da raiz de um verbo desconhecido. Suponhamos que este seja *phéreis* ("carregas"); o raciocínio seria: "eu sei que *lúeis* é a segunda pessoa do singular da raiz *lúo*; de modo similar, *phéreis* deve ser a segunda pessoa do singular de *phéro*".

O uso da analogia pode ser especialmente útil quando lidamos com um grande número de regras, por exemplo, com os chamados "verbos contratos". O radical de um verbo contrato termina em vogal; quando um sufixo com vogal inicial é incorporado ao radical, ocorrem mudanças na grafia. Por exemplo, o radical *pl<u>e</u>ro-* (encher) combinado com o sufixo *–omen* torna-se *pler-ou-men,* devido à interacão entre dois *ômicrons* nos extremos das palavras. Embora seja possível deduzir essas mudanças a partir de seus primeiros princípios ou memorizar as regras para todas as permutações de vogais (por exemplo, "o" + "o" = "ou"), deve ser mais fácil recordar as mudanças gráficas vistas em um verbo familiar (por exemplo, *pl<u>e</u>ro o → pleroumen)* e então usar a analogia para inferir o radical de um verbo não visto ainda. A abordagem de aprendizagem de máquina do vizinho mais próximo é utilizada para fornecer essas analogias. Dada uma palavra numa forma flexionada (e.g., *phéreis*) o algoritmo procura pela forma do radical (*phér<u>o</u>)* entre seus "vizinhos", fazendo substituições em seus prefixos e sufixos. Substituições válidas são extraídas de pares de formas flexionadas e de raiz/radicais (e.g., *lúeis, lúo*) no grupo de treinamento; esses pares, então, podem servir como analogias para reforçar a aprendizagem de máquina. Ademais, essas substituições de afixos podem ser aprendidas automaticamente, reduzindo a quantidade de esforços de engenharia. Elas também aumentam a transparência do analisador, mostrando explicitamente como ele deduz a raiz.

Raízes novas

O grego antigo, em seus vários dialetos, foi usado desde a época de Homero até a Idade Média, em textos de grande variedade de gêneros. Mesmo os mais abrangentes dicionários não abarcam completamente seu extenso vocabulário. Pelo que sabemos, todos os atuais analisadores do grego antigo requerem uma base de dados de radicais pré-definida; assim, provavelmente encontrarão palavras com radicais novos ou desconhecidos, que eles não são designados a analisar. Ao invés de expandir uma base de dados para aumentar a cobertura, criamos um mecanismo para lidar com todas as novas raízes. Uma vez que as palavras não aparecem com frequência em suas formas de raiz/formas radicais, inferir uma raiz nova a partir de uma forma superfície não é tarefa fácil (LINDE, 2008). Propomos o uso de dados não etiquetados para guiar a determinação de uma raiz nova.

Trabalhos prévios

Depois de uma breve discussão a respeito da análise morfológica em geral, revisaremos analisadores existentes específicos para o grego antigo.

Análise morfológica

Uma tarefa fundamental na análise morfológica é a segmentação da palavra em morfemas, isto é, na menor unidade portadora de significado de uma palavra. Métodos não supervisionados[130] mostraram uma boa atuação nessa tarefa. No recente desafio PASCAL, os melhores resultados foram conquistados por Keshava e Pitler (2006). Seu algoritmo descobre afixos, considerando palavras que aparecem como subcadeias de outras palavras, e estimando probabilidades para fronteiras de morfema. Outra abordagem bem sucedida é a do uso do Comprimento Mínimo de Descrição, que interativamente reduz a extensão da gramática morfológica (GOLDSMITH, 2001). Mudanças na grafia nas fronteiras de morfema (e.g., *deny*, mas *deni-al*) podem ser capturadas por regras

[130] N.E.P. Métodos não supervisionados são aqueles em que as amostras não têm uma identificação ou classificação prévia.

ortográficas como "mude *y*- para *i*- quando o sufixo é *–al*". Tais regras são especificadas manualmente no modelo morfológico de dois níveis (KOSKENNIEMI, 1983), mas elas também podem ser induzidas (DASGUPTA, 2006). Alomorfes (e.g., *deni* e *deny*) também são automaticamente identificados em Dasgupta (2007), mas o problema geral de reconhecer formas altamente irregulares é examinado mais extensivamente em Yarowsky e Wicentowski (2000). Eles tentam alinhar cada verbo com sua forma de raiz/forma radical, explorando a combinação de similaridade de frequência, similaridade de contexto, distância de edição e probabilidades de transformação morfológica, todas estimadas a partir de um *corpus* não anotado. Uma precisão de 80,4% foi alcançada para palavras altamente irregulares no conjunto de testes.

Desafios para o Grego Antigo

O grego antigo apresenta algumas dificuldades que impedem uma ingênua aplicação da abordagem minimamente supervisionada em Yarowsky e Wicentowski (2000). Primeiramente, análises de frequência e contexto são sensíveis à escassez de dados, que é mais pronunciada em línguas altamente flexionadas, como o grego, do que no inglês. Muitas das formas flexionadas não aparecem mais que algumas vezes. Em segundo lugar, muitas formas de raiz não aparecem[131] no *corpus.* Em finlandês e suaíl, também línguas altamente flexionadas, apenas de 40 a 50% das palavras aparecem em suas formas de raiz/radicais (LINDE'N, 2008). O mesmo deve ser esperado do grego antigo.

De fato, para essas línguas, prever raízes novas é um problema desafiador. Essa tarefa foi tratada em (ALDER *et al.*, 2008) para o hebraico moderno, e em Linde'n (2008), para o finlandês. No primeiro, características como letra n-*grams n-grams* de letra e padrões de formação de palavras são usados para predizer a morfologia das palavras hebraicas desconhecidas a um analisador existente. Na segunda, uma abordagem probabilística é usada para extrair prefixos e sufixos em palavras finlandesas, favorecendo as mais longas. Entretanto, nenhuma estratégia foi proposta para grafias irregulares nos radicais.

[131] As formas raízes dos verbos contratos, por exemplo *pheró<u>o</u>*, não são sequer formas flexionadas.

Tabela 2 – Dados da amostra de partes do Gênesis 1:2 ("e o Espírito de Deus pairava no ar…"). A anotação original é mais extensa e somente a porção usada nesta pesquisa está mostrada aqui.

Forma de superfície	Anotação morfológica	Forma radical ou de raiz
kaí (e)	Conjunção	*kaí*
pneuma (espírito)	Subst. de 3ª decl.	*pneuma*
theou (deus)	Subst. de 2ª decl.	*theós*
epephéreto (pairar)	Verbo	*phéro*

Análise Morfológica do Grego Antigo

Os dois analisadores mais conhecidos para o grego antigo são ambos sistemas baseados em regras, requerendo *a priori* um conhecimento sobre os possíveis radicais e afixos, que são compilados manualmente. Para se ter uma ideia básica, por volta de 40.000 radicais e 13.000 flexões são conhecidos pelo sistema MORPHEUS, que será descrito abaixo.

O algoritmo em MORPH (PACKARD, 1973) procura por terminações possíveis que resultariam em um radical em sua base de dados. Caso não obtenha sucesso, ele então tenta remover preposições e prefixos do início da palavra. Acentos, essenciais para desambiguização em alguns casos, são ignorados. O analisador foi usado na *Apologia,* de Platão, para estudar a distribuição das terminações das palavras, com o propósito de otimizar a origem dos tópicos gramaticais para serem abarcados em um curso introdutório. A avaliação do analisador salientou essa perspectiva pedagógica e a precisão das análises não é relatada.

Morpheus (CRANE, 1991) ampliou o MORPH com um componente generativo que, dado um radical, enumera todas as possibilidades de flexão em diferentes dialetos, incluindo acentos. Quando os acentos são considerados durante a análise, a precisão do analisador melhora em um quarto. No entanto, a real precisão e o conjunto de testes não estão especificados.

Neste artigo, optamos por uma abordagem **orientada por dados** para determinar automaticamente os radicais e os afixos dos dados de treino.

Dados

Dados morfológicos

Usamos o *corpus* da Septuaginta[132] preparado pelo Centro de Análise Computacional de Textos na Universidade da Pensilvânia. A Septuaginta, datando do terceiro ao primeiro séculos a. C, é uma tradução grega da Bíblia hebraica. O *corpus* está analisado morfologicamente e a Tabela 2 mostra alguns desses dados.

Tabela 3 – Estatísticas das classes gramaticais (*parts-of-speech*) das palavras no conjunto de testes, considerando apenas palavras de única ocorrência

Parte do discurso	Percentual
Verbos	68,6%
Adjetivos	10,4%
Nomes (1ª declinação)	5,6%
Nomes (2ª declinação masculino)	4,3%
Nomes (2ª declinação neutro)	2,8%
Nomes (3ª declinação)	7,6%
Outros	0,7%

O *corpus* está dividido entre conjuntos de treinamento e de testes. O conjunto de treinamento é composto por toda a *Septuaginta* com exceção dos primeiros cinco livros. Ele consiste em cerca de 470K palavras, com 37.842 palavras de ocorrência única. Os primeiros cinco livros, também conhecidos como o Torá ou Pentateuco, Torah ou Netateuch, constituem o conjunto de testes. Este contém 120K palavras, dentre as quais há 3.437 palavras de ocorrência única não vistas no conjunto de treinamento, e 7.381 palavras de ocorrência única vistas no conjunto de treinamento. Uma especificação das

[132] http://ccat.sas.upenn.edu/gopher/text/religion/biblical/.

categorias gramaticais (*parts-of-speech*) do conjunto de testes é fornecida na Tabela 3. Nomes próprios, muitos dos quais não declinam, estão excluídos de nossa avaliação.

Dados não etiquetados

A fim de conduzir a previsão de novas raízes, utilizamos o *corpus* do *Thesaurus Linguae Graecae* (BERKOWITZ; SQUITTER, 1986), que contém mais de um milhão de palavras únicas, retiradas de uma grande variedade de textos em grego antigo.

Avaliação

Muitas palavras comuns no conjunto de testes também são vistas no conjunto de treinamentos. Em vez de aumentar artificialmente a taxa de precisão, avaliaremos a *performance* em palavras únicas ao invés de todas as palavras, individualmente.

Algumas formas superficiais possuem mais de uma forma de raiz possível. Por exemplo, a palavra *purõn* pode ser flexionada do substantivo *purá* ("altar"), ou *purós* ("trigo") ou *pũr* ("fogo"). Seria necessário examinar o contexto para selecionar o substantivo apropriado, mas a desambiguização morfológica (HAKKANU-TÜR *et al.*, 2002) está além do objetivo deste artigo. Nesses casos (tirar/legitimar) formas de raiz legítimas propostas pelo nosso analisador podem ser rejeitadas, mas arcamos com o preço em troca de um procedimento de avaliação automática.

Abordagem do Vizinho mais Próximo

A abordagem de aprendizagem de máquina baseada em memória opera bem em marca de medida de tarefas de aprendizagem de línguas (DAELEMANS, 1999), incluindo segmentação morfológica do holandês (VAN DEN BOSCH, 1999). Nesse quadro teórico, vetores de características são extraídos do conjunto de treinamento e armazenados numa base de dados de ocorrências, chamada **base de ocorrências**. Uma métrica de distância é então definida. Para cada ocorrência do teste, seu conjunto de **vizinhos mais próximos** é recuperado da base de ocorrências, e a etiqueta em maioria do conjunto é retornada.

Agora, adaptamos esse quadro teórico para nossa tarefa, primeiro definindo a métrica de distância (seção atual), então descrevendo o algoritmo de pesquisa para o vizinho mais próximo (§6).

Métrica de distância

Toda palavra consiste em um radical, um prefixo (possivelmente vazio) e um sufixo (possivelmente vazio). Se duas palavras compartilham um mesmo radical, uma pode ser transformada na outra, substituindo seu prefixo e sufixo com suas contrapartes na outra palavra. Chamaremos essas substituições de **transformação de prefixo** e de **transformação de sufixo**.

A "distância" entre as palavras deve ser definida em termos dessas transformações. Seria desejável a duas palavras flexionadas do mesmo radical serem vizinhas próximas. A métrica da distância pode obter esse efeito, favorecendo transformações de prefixo e sufixo que são frequentemente observadas entre palavras flexionadas de um mesmo radical. Nós, portanto, provisoriamente definimos "distância" como a soma das contas de frequência das transformações sufixais e prefixais requeridas para alternar uma palavra na outra.

Radicais e afixos

Definindo "radical". Para contar as frequências de transformações de prefixo e sufixo, o radical de cada palavra no conjunto de treinamento deve ser determinado. De modo ideal, todas as palavras flexionadas de uma mesma raiz deveriam compartilhar um mesmo radical. Infelizmente, no caso do grego antigo, é difícil insistir em um mesmo radical comum. Em alguns casos, os radicais são completamente diferentes[133]; em outros, o radical comum é ofuscado em formas superficiais devido às mudanças na grafia[134]. Recorremos a uma definição funcional de "radical" – a subcadeia comum mais longa de

[133] Cada verbo pode ter até seis diferentes radicais, conhecidos como as "partes principais". Em casos extremos, um radical pode aparecer sem qualquer relação com a raiz na superfície. Por exemplo, o *íso* e *énegkon* são ambos radicais da raiz *phér<u>o</u>* (levar). Um exemplo comparável em inglês é a forma verbal flexionada *went* e sua forma raiz *go.*

[134] Por exemplo, a raiz *oz* na forma raiz *óz<u>o</u>* (cheirar) muda para *os* em *exósthesan*, uma forma do aoristo passivo.

um par de palavras. Alguns exemplos são mostrados na tabela 4. A raiz do verbo *lú*o (soltar) e três de suas formas flexionadas são mostradas. Cada forma flexionada é comparada à forma raiz, bem como as outras formas flexionadas. O radical, definido como a subcadeia comum mais longa, é determinado para cada par. As transformações de prefixo e sufixo são então extraídas. *O* representa a cadeia vazia.

Tabela 4

Palavra	Prefixo	Radical	Sufixo	Transformação Prefixo		Transformação Sufixo
(raiz) *lú*o	-	*lú*	o	(raiz, 1)	ǫ ↔e	o ↔ eto
(1) *elúeto*	*e*	*lú*	*eto*	(raiz, 2)	ǫ ↔ para	o ↔ sai
(2) *paralũsai*	*para*	*lũ*	*sai*	(raiz, 3)	ǫ ↔ ek	o ↔ thésontai
(3) *ekluthésontai*	*ek*	*lu*	*thésontai*	(1,2)	e ↔ para	eto ↔ sai
				(1,3)	e ↔ ek	eto ↔ thésontai
				(2,3)	para ↔ ek	sai ↔ thésontai

Refinamentos para a definição. Três refinamentos adicionais para a definição de "radical" foram consideradas úteis. Em primeiro lugar, acentos são ignorados quando estão determinando a subcadeia comum mais longa. Acentos nos radicais frequentemente mudam no processo de flexão. Essas mudanças são ilustradas na tabela 4 pelo radical *lu*, cuja letra *u* possui um acento agudo, um acento circunflexo e nenhum acento em três formas flexionadas.

Em segundo lugar, uma extensão mínima é requerida ao radical. Por um lado, alguns pares como *ágo* (conduzir) e *áxo* possuem um radical de extensão um (1) ("*a*"). Por outro lado, permitir radicais tão pequenos pode prejudicar a performance, uma vez que muitos radicais falsos podem ser construídos por engano, tal como "*e*" entre *phéro* e *énegkon.* Para este artigo, a extensão mínima do radical é empiricamente definida como dois.

A extensão por si só não pode filtrar todos os radicais falsos. Por exemplo, para o par *patéo* (andar) e a forma flexionada *katepástesan*, há dois radicais candidatos

igualmente longos, **ate* e *pat.* Esse último produz afixos como *"-éo"* e "*-esan*" que são relativamente frequentes[135]. Baseado nisso, o segundo radical é escolhido.

Alguns meios adicionais de reduzir o ruído são requerer uma transformação de afixo, para ocorrer por pelo menos um número mínimo de vezes no conjunto de treinamento, e restringir o contexto fonológico no qual a transformação possa ser aplicada[136]. Embora reduzam significativamente a rechamada/o reprocessamento (*recall*), essas restrições adicionais produzem apenas uma melhora limitada na precisão.

Algoritmo

Na etapa de treinamento, um conjunto de transformações de prefixos e sufixos, juntamente com a contagem deles, é compilada para cada classe gramatical (*part-of-speech*). Essas contagens nos permitem computar a distância entre quaisquer duas palavras e, portanto, determinar o "vizinho mais próximo" de cada palavra. No teste, dada uma forma flexionada, seu vizinho é qualquer palavra para a qual pode ser transformada usando as transformações de afixos. Primeiro tentamos encontrar seu vizinho mais próximo no conjunto de treinamento (§6.2); se nenhum é encontrado, uma nova raiz é prevista (§6.2).

Encontrando raízes conhecidas

Se a palavra de entrada (*input*) nos dados aparece no conjunto de treinamento, simplesmente olhamos sua análise morfológica. Se a palavra de entrada não é vista no conjunto de treinamento, a forma de sua raiz ou outra forma flexionada pode ainda ser encontrada. Tentamos transformar a palavra inserida no vocábulo mais próximo, i.e., usando as transformações de prefixais e sufixais mais frequentes, de acordo com a métrica de distância (§5.1).

[135] A frequência de cada afixo é contada em um turno preliminar, em que cada afixo recebe metade da contagem em casos de extensão de radical unido.
[136] Por exemplo, uma certa transformação de sufixo pode ser válida apenas quando o radical terminar com certas letras.

Grafia de radicais irregulares. Tipicamente, se não há mudanças na grafia em um radical, a palavra testada pode ser transformada diretamente à raiz, por exemplo, de *phéreis* para *phéro.* Se a grafia do radical é substancialmente diferente, é provável que seja transformada para outra forma flexionada da raiz que contém o mesmo radical irregular. Por exemplo, a palavra *prosexénegken* carrega uma pequena semelhança com sua raiz *phéro,* mas pode ser mapeada à palavra *énegken* no conjunto de treinamento, do qual recuperamos sua raiz de *phéro.*

Ordem de pesquisa. Alguns afixos são circunfixos, isto é, tanto o prefixo quanto o sufixo devem ocorrer juntos. Por exemplo, o sufixo, *-eto* não pode ser aplicado sozinho, mas deve sempre ser usado em conjunção com o prefixo *e-*, para formar palavras como *elúeto,* conforme mostrado na tabela 4.

Outros afixos, entretanto, podem misturar-se livremente uns com os outros e nem todas as combinações estão certificadas no conjunto de treinamento. Isso é particularmente comum quando o prefixo contém duas ou mais preposições. Por exemplo, a combinação *dia-kata-* ocorre apenas duas vezes no conjunto de treinamento, mas ela pode potencialmente emparelhar com um grande número de sufixos diferentes.

A procura por vizinhos, portanto, prossegue em dois estágios. No primeiro (denominado CIRCUNFIXO) a pesquisa é restrita aos circunfixos, isto é, requer que ao menos um par de palavras no conjunto de treinamento contenha ambas transformações de prefixo e sufixo. Essa restrição está inclinada à variação de dados; se nenhum vizinho é encontrado, as transformações prefixais e sufixais são então permitidas a serem aplicadas separadamente no segundo estágio (denominado PREFIXO/SUFIXO).

Propondo novas raízes

Uma palavra pode ser derivada de uma raiz da qual nenhuma forma flexionada é vista no conjunto de treinamento. Naturalmente, nenhum vizinho seria encontrado na etapa anterior, e uma nova raiz deve ser proposta. Aplicamos as transformações de prefixo e sufixo aprendidas em § 5.2, usando apenas circunfixos observados entre uma forma flexionada e uma forma raiz. Por razões óbvias, a cadeia resultante não é mais requerida para ser um vizinho, por exemplo, uma palavra vista no conjunto de treinamento.

Normalmente, as várias transformações produzem muitas raízes candidatas. Por exemplo, a palavra *homom**e**tríou* (nascido da mesma mãe), um adjetivo masculino genitivo, pode ser transformado em sua raiz adjetiva *homom**é**trios*, porém não poderia ser igualmente bem transformada em um hipotético nome neutro, **homométrion.* Ambas as formas são raízes perfeitamente plausíveis.

As transformações de afixo descobertas automaticamente inevitavelmente contêm algum ruído. Quando lidamos com raízes conhecidas, muitos dos ruídos são suprimidos, porque aplicações equivocadas dessas transformações raramente transformam a palavra de entrada em uma palavra real encontrada no conjunto de dados de treinamento. Quando propomos novas raízes, não mais tiramos proveito dessa limitação. Embora a métrica de distância ainda ajude a discriminar candidatos inválidos, a elevada ambiguidade nos leva a uma menor precisão. Abordamos essa questão explorando um *corpus* maior, não etiquetado.

Uso de um *corpus* não etiquetado. Se uma forma de raiz proposta estiver correta, ela deverá ser capaz de gerar formas flexionadas atestadas em um amplo *corpus.* Intuitivamente, a "produtividade" da forma de raiz deve ter correlação com sua exatidão.

Para gerar formas flexionadas de uma raiz, nós simplesmente pegamos o conjunto de transformações de afixos observadas das formas flexionadas para as raízes e revertemos as transformações. Continuando com o exemplo acima, geramos formas flexionadas para ambas as raízes candidatas: o adjetivo *homom<u>é</u>trios,* e o hipotético substantivo neutro **homom<u>é</u>trion.* Enquanto algumas formas flexionadas são geradas por ambos os candidatos, três são únicas ao adjetivo – *homométrios, homométrioi* e *homom<u>é</u>trian* – o nominativo masculino singular e plural, e o acusativo feminino singular, respectivamente. Nenhuma dessas poderia ter sido flexionada a partir de um substantivo neutro.

Uma noção direta de produtividade de uma raiz seria simplesmente o número de formas flexionadas atestadas em um grande *corpus*. Ela pode ser refinada mais, contudo, considerando a prevalência das formas flexionadas; isto é, a uma forma gerada com mais transformações de afixos compartilhados deve ser dada maior importância do que a uma gerada com transformações menos compartilhadas. Suponhamos que as duas raízes candidatas, o adjetivo *telesphóros* (que chega ao final) e o verbo hopotético **telesphoró<u>o</u>*,

sejam consideradas. Ambas podem gerar a forma flexionada *telesphórou*, a primeira como adjetivo masculino, genitivo e a segunda, tanto como indicativo imperfeito, como presente do imperativo de um verbo contrato. Uma vez que a flexão do adjetivo é mais frequente no conjunto de dados de treinamento do que a relativamente rara classe de verbos contratos, a existência de *telesphórou* deveria conceder mais peso ao adjetivo.

Portanto, a métrica "de produtividade" de uma nova raiz é o número de palavras em um grande *corpus* que pode gerar transformações de afixo, ponderado pelas frequências dessas transformações.

Experimentos

Algumas estatísticas do conjunto de testes são apresentadas na Tabela 3. Dentre as 7.831 palavras que são vistas no conjunto de dados de treinamento, 98,2% receberam a forma raiz correta. Depois de excluir palavras conhecidas, que atingem uma precisão de 98,2%, a performance das 3437 palavras únicas restantes no conjunto de testes é mostrada acima. Por favor, veja §7 para discussões. Resultados para novas raízes são apresentadas com mais detalhes na Tabela 6.

Tabela 5

Tipo de transformação	Proporção	Precisão
CIRCUNFIXO	77,5%	94,5%
PREXIFO/SUFIXO	10,8%	61,2%
Raízes novas	11,7%	50,0%
Total	100%	85,7%

O 1,8% restante tinha múltiplas raízes possíveis; um exame do contexto correto seria necessário para desambiguação (ver comentários em §4.3).

A Tabela 5 apresenta a precisão das raízes previstas, depois de excluir 7381 palavras vistas. O resultado é dividido de acordo com o tipo de transformação: para o tipo "Raízes Novas" mais detalhes dos resultados são apresentados na tabela 6.

Tal qual discutido em §6.1, o algoritmo primeiro pesquisou com CIRCUNFIXO. Para 77,5% das palavras, um vizinho foi encontrado usando o subconjunto de transformações de afixos. O restante foi então processado usando procedimentos de *back-up,* PREFIXO/SUFIXO, permitindo transformações de prefixo e sufixo selecionadas de diferentes pares de palavras. Esse procedimento encontrou vizinhos para 10.8% das palavras; novas raízes foram hipotetizadas para as restantes.

Não surpreendentemente, raízes conhecidas foram previstas com mais confiabilidade (94,5%) com mais circunfixos do que com prefixos e sufixos separados (61,2%), mas ambas categorias ainda alcançaram uma precisão maior do que a desafiadora tarefa de propor novas raízes (50,0%). Agora olharemos mais de perto os erros para as raízes conhecidas e também para as novas.

Raízes conhecidas

Há três fontes de erros principais. O primeiro é o ruído nas transformações de afixo. Por exemplo, a transformação de prefixo p↔ph foi derivada do par *phér<u>o</u>* e *perien<u>é</u>gkasan.* Quando aplicada em *pasát<u>o</u>,* junto com a transformação de sufixo, ela produziu a falsa raiz de *phásk<u>o</u>.*

Uma segunda fonte pode ser atribuída às fronteiras incorretas de afixo. Por exemplo, *ektéinantes* foi erroneamente construído com "*e-*" ao invés da preposição *ek* como prefixo. Esse prefixo é por si só perfeitamente viável, mas "*e-*" e "*-antes*" não podem ocorrer juntos como circunfixo. A cadeia resultante ocorreu para fazer pareamento com a raiz *ktéin<u>o</u>,* ao invés da verdadeira raiz *teín<u>o</u>*

Uma terceira fonte é a confusão entre classes gramaticais (*parts-of-speech*), mais comumente substantivos e verbos. Por exemplo, o vizinho mais próximo do substantivo genitivo *lup<u>õ</u>n* foi o verbo *lup<u>é</u>sei,* produzindo a raiz verbal *lupé<u>o</u>* ao invés do nome *lúp<u>e</u>.*

Tabela 6

Método de avaliação	Precisão
BASELINE	45.0%
Rerank TLG	50.0%
+ *Ignorar acentos*	55.2%
+ *Oracle POS*	65.5%

Resultados para prever raízes novas, para as 402 palavras para as quais nenhum vizinho foi encontrado. BASELINE usa a métrica da distância (§5.1) como antes: TLG RERANK explora o *corpus* não etiquetado do *Thesahurus Linguae Graecae* para reposicionar os candidatos mais esperados (§6.2) propostos pela BASELINE.

Novas raízes

Como uma linha de base, a métrica de distância foi usada isoladamente para ranquear as candidatas a novas raízes. Conforme visto na tabela 6, o desempenho caiu para 45.0%.

Quando o *corpus* do *Thesaurus Linguae Graecae* foi utilizado para reposicionar novas raízes candidatas propostas pela linha de base, um ganho absoluto de 5% foi alcançado[137]. Fora desses, os outros 5.5% dos erros foram devido à colocação incorreta do acento, tal como *kt<u>e</u>nótophos* ao invés de *kt<u>e</u>notróphos*, a maioria em substantivos e adjetivos. Esses erros são difíceis de retificar, uma vez que múltiplas posições são com frequências possíveis[138].

Finalmente, para medir a extensão por quais as confusões das classes gramaticais (*parts-of-speech* POS) são responsáveis, realizamos um experimento em que o padrão ouro POS de cada palavra foi fornecido ao analisador (ver "Oracle POS" na tabela 6). Ao

[137] O nível de significância é para $p = 0{,}11$, de acordo com o teste de McNemar. O desempenho melhor não é estatisticamente significativo e pode ser um reflexo do relativamente pequeno conjunto de dados de teste.
[138] O acento em um substantivo flexionado retém sua posição na raiz, a menos que certas regras fonológicas violem essa posição. Em muitos casos, não há meio confiável de predizer a posição do acento no substantivo-raiz a partir da posição da forma flexionada.

derivar raízes novas, apenas as transformações de afixo pertencentes ao oráculo POS foram consideradas. Com essa restrição, a precisão aumentou para 65.5%.

Conclusão

Propusemos uma abordagem de aprendizagem de máquina do vizinho mais próximo para analisar a morfologia do grego antigo. Esse quadro é orientado por dados, com a descoberta automática de radicais e afixos. O analisador é capaz de prever novas raízes. Uma novidade significativa é a exploração de um grande *corpus* não etiquetado *corpus* para melhorar o desempenho.

Planejamos melhorar mais a derivação de raízes novas prevendo suas classes gramaticais do contexto e incorporando informação distribucional (YAROWSKY; WICENTOWSKI, 2000).

Agradecimentos

O autor gostaria de agradecer Stephanie Seneff, Kalliroi Georgila, Konstantinos Katsiapis e Steven Lulich por seus comentários criteriosos.

Referências

ADLER, M.; GOLDBERG, Y.; GABAY, D.; ELHADAD, M. Unsupervised Lexicon-based Resolution of Unknown Words for Full Morphological Analysis. **Proc. ACL.** Columbus: OH, 2008.

BERKOWITZ, L.; SQUITTER, K. A. **Thesaurus Linguae Graecae**. UK: Oxford University Press, 1986.

BOSCH, A. van den; DAELEMANS, W. Memory-based Morphological Analysis. **Proc. ACL.** MD: College Park, 1999.

CRANE, G. Generating and Parsing Classical Greek. **Literary and Linguistic Computing**, v. 6, n. 4, p. 243-245, 1991.

_____. **Perseus 2.0: Interactive Sources and Studies on Ancient Greece.** New Haven: Yale University Press, 1996.

DAELEMANS, W.; BOSCH, A. van den; ZAVREL, J. Forgetting Exceptions is Harmful in Language Learning. **Machine Learning**, v. 34, p. 11-41, 1999.

DASGUPTA, S.; NG, V. High-Performance, Language-Independent Morphological Segmentation. **Proc. HLT-NAACL**. NY: Rochester, 2007.

GOLDSMITH, J. Unsupervised Learning of the Morphology of a Natural Language. **Computational Linguistics**, v. 27, n. 2, p. 153-198, 2001.

HAKKANI-TÜR, D. Z.; OFLAZER, K.; TU˙R, G. Statistical Morphological Disambiguation for Agglutinative Languages. **Computers and the Humanities**, v. 36, n. 4, p. 381-410, 2002.

KESHAVA, S.; PITLER, E. A Simpler, Intuitive Approach to Morpheme Induction. **Proc. 2nd PASCAL Challenges Workshop.** Venice, 2006.

KOSKENNIEMI, K. Two-level morphology: a general computation model for word-form recognition and production. **Publication No. 11, Department of General Linguistics,** University of Helsinki. Helsinki, Finland. 1983.

LINDEN, K. A Probabilistic Model for Guessing Base Forms of New Words by Analogy. **Proc. CICLing**, Haifa, Israel. 2008.

PACKARD, D. W. Computer-assisted Morphological Analysis of Ancient Greek. **Proc. 5th Conference on Computational Linguistics**. Pisa, Italy. 1973.

YAROWSKY, D.; WICENTOWSKI, R. Minimally Supervised Morphological Analysis by Multimodal Alignment. **Proc. ACL.** Hong Kong, China. 2000.

REDES SOCIAIS E A LINGUAGEM DA TRAGÉDIA GREGA[139]

Jeff Rydberg-Cox[140]

Introdução

Usando os *treebanks* de dependência linguística e textos digitalizados criados pela Biblioteca Digital Perseu, estamos criando as redes sociais para uma coleção de tragédias gregas que permitem aos usuários visualizar as interações entre os personagens nas peças[141]. Como o número de personagens que aparecem no palco em uma tragédia grega é limitado, a maioria destes diagramas de rede sociais se enquadram em alguns tipos básicos. O aspecto mais interessante destas redes é, portanto, as arestas que conectam os nós nos gráficos. Os dados linguísticos usados para rotular ou mesmo criar essas arestas torna-se o ponto de partida para visualizar e explorar a linguagem da tragédia grega.

Estes gráficos de redes sociais são concebidos para traçar um meio-termo entre a abordagem de leitura distância emergente adotada por muitos humanistas digitais e uma abordagem leitura atenta tradicionalmente adotada por alunos e estudiosos das ciências humanas. À medida que as grandes coleções de textos são colocadas on-line, uma das

[139] Publicado originalmente em 2011, com o título *Social Networks and the Language of Greek Tragedy,* no *Journal of the Chicago Colloquium on Digital Humanities and Computer Science.* Em português, distribuído sem fins comerciais, sob licença *Creative Commons* BY-NC-SA.

[140] Department of English, Language & Literature – University of Missouri-Kansas City – UMKC – 64110 – Kansas City – Missouri – Estados Unidos – rydbergcoxj@umkc.edu.

[141] Nenhum trabalho foi realizado sobre as redes sociais na tragédia grega, mas outros estudiosos têm feito este tipo de trabalho sobre as peças de Shakespeare: J. Stiller, Dr. Nettle e R. I. M. Dunbar, "The Small World of Shakespeare's Plays," *Human Nature* 14, n. 4 (2003): 397-408; P. Mutton, "Inferring and Visualizing Social Networks on Internet Relay Chat," em *Proceedings Eighth International Conference on Information Visualisation IV* (2004); J. Stiller and M. Hudson, "Weak Links and Scene Cliques Within the Small World of Shakespeare," *Journal of Cultural and Evolutionary Psychology* 3, nº 1 (2005): 57-73, e os círculos literários da literatura do século XVIII e XIX. Gillian Russell and Clara Tuite, *Romantic Sociability:Social Networks and Literary Culture in Britain, 1770-1840* (Cambridge, U.K.; New York: Cambridge University Press, 2002). O recente artigo de Franco Moretti (Franco Moretti, "Network Theory, Plot Analysis," *New Left Review*, no. 68 (2011): 80-102) e a obra de David Elson em Columbia explorando a extração automática de redes sociais de textos de romances do século XIX é a obra atual mais intrigante nesta área (see D. Elson and K. McKeown, "Extending and Evaluating a Platform for Story Understanding," in *Proceedings of the AAAI 2009 Spring Symposium on Intelligent Narrative Technologies II* (2009); D. Elson and K. McKeown, "A Tool for Deep Semantic Encoding of Narrative Texts," in *Proceedings of the ACL- IJCNLP 2009 Software Demonstrations* (2009); D. Elson, Nicholas Dames, and K. McKeown, "Extracting Social Networks From Literary Fiction," in *Proceedings of the 48Th Annual Meeting of the Association for Computational Linguistics.* ACL '10. Association for Computational Linguistics (2010); and D. Elson and K.McKeown, "Automatic Attribution of Quoted Speech in Literary Narrative," in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence* (2010).

questões mais prementes para os estudiosos de ciências humanas é o que exatamente fazer com os vastos corpora de fontes primárias disponíveis em formato digital[142]. Uma abordagem para os vastos corpora que vêm surgindo é usar várias técnicas de leitura a distância na qual dados quantificáveis como padrões de publicação ou palavras-chave são extraídas e visualizadas. Franco Moretti tem seguido essa abordagem em seu trabalho sobre os padrões de publicação em torno do surgimento do romance como gênero coerente e o surgimento posterior de gêneros[143]. Abordagens como estas são extremamente interessantes e valiosas, contudo não tratam de questões que o leitor pode perguntar ao ler um determinado texto literário[144].

O intuito deste projeto é encontrar o espaço entre a leitura a distância e a leitura próxima; da mesma forma que a abordagem de leitura a distância, ele tenta descobrir amplos padrões quantificáveis dentro de textos literários; da mesma forma que a abordagem de leitura próxima, ele tenta focar ou em obras literárias individuais ou em coleções de textos literários relativamente pequenas. Esperamos que com os métodos quantitativos os leitores possam se orientar dentro de uma obra literária e fazer conexões entre os personagens. Ao mesmo tempo, esperamos que uma abordagem baseada em visualização tornará os dados quantitativos sobre o texto mais acessíveis aos leitores que não são especialistas em métodos estatísticos.

Dados da fonte

Estes gráficos de redes sociais têm por base textos digitais e *treebanks* que foram criados na Biblioteca Digital Perseu e divulgados nos termos da licença Creative Commons[145]. Os próprios textos há anos pertencem à Biblioteca Digital Perseu e são

[142] Ver Gregory Crane, "What Do You Do with a Million Books?" *D-Lib Magazine* 12, no. 3 (2006) e a questão que acompanha o D-Lib.
[143] Moretti, *Graphs, Maps, Trees: Abstract Models for a Literary History.* Moretti, *The Novel.* Moretti, "Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740-1850)".
[144] See D. Bamman and G. Crane, "The Design and Use of a Latin Dependency Treebank," em *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories* (TLT2006) (2006); D. Bamman, M. Passarotti, G. Crane, and S. Raynaud, "A Collaborative Model of Treebank Development," em P*roceedings of the Sixth International Workshop on Treebanks and Linguistic Theories* (December 2007); D. Bamman, F. Mambrini, and G. Crane, "An Ownership Model of Annotation: The Ancient Greek Dependency Treebank," em P*roceedings of the Eighth International Workshop on Treebanks and Linguistic Theories* (TLT8) (2009); and J. Lee and D. Haug, "Porting An Ancient Greek and Latin Treebank". In: *Proc. LREC* (2010).
[145] Ver D. Bamman e G. Crane, "The Design and Use of a Latin Dependency Treebank," in *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT2006)* (2006); D. Bamman, M. Passarotti,

codificados em XML conforme a TEI. Os *treebanks* também foram criados por equipes que trabalham com a Biblioteca Digital Perseu desde 2007. *Treebanks* são conjuntos de dados que contêm análises sintáticas das relações gramaticais entre cada palavra em uma coleção de textos, juntamente com informações sobre qual palavra depende de qual. Desde 2007, equipes de estudiosos e pesquisadores de graduação têm trabalhado nestes *treebanks* e etiquetaram cerca de 53 mil palavras do latim clássico e 192.000 palavras do grego antigo. Nota: Para versões maiores e de melhor maior qualidade das figuras reproduzidas aqui, consulte a seção *Supplementary Data* que acompanha este artigo *on-line* em: http://jdhcs.uchicago.edu.

Tipos de redes sociais em tragédia grega

Como o número de personagens que aparecem no palco ao mesmo tempo é limitado na tragédia grega, suas redes sociais tendem se enquadrar em um dos quatro tipos essenciais. Um tipo aparece em peças onde um personagem central ocupa o palco e uma sequência de personagens entra no palco para falar com aquela pessoa, como em *Prometeu Acorrentado* de Ésquilo.

O segundo tipo ocorre quando todos os personagens ocupam o palco essencialmente ao mesmo tempo e todas se comunicam, como em *Suplicantes* de Ésquilo.

O terceiro tipo surge quando grupos de personagens aparecem no palco por sua vez e falam uns com os outros, sem um personagem central no palco todo, como em *Ajax* de Sófocles.

O quarto tipo aparece quando há dificuldades ou anomalias textuais, como o fim espúrio *Sete Contra Tebas* de Ésquilo, onde Antígona e Ismene não falam com os outros personagens da peça.

G. Crane e S. Raynaud, "A Collaborative Model of Treebank Development," in *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories* (December 2007); D. Bamman, F. Mambrini e G. Crane, "An Ownership Model of Annotation: The Ancient Greek Dependency Treebank," in *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT8)* (2009); e J. Lee e D. Haug, "Porting An Ancient Greek and Latin Treebank". Em *Proc.LREC* (2010).

Figura 1 – Diagrama de rede social para *Prometeu Acorrentado* de Ésquilo

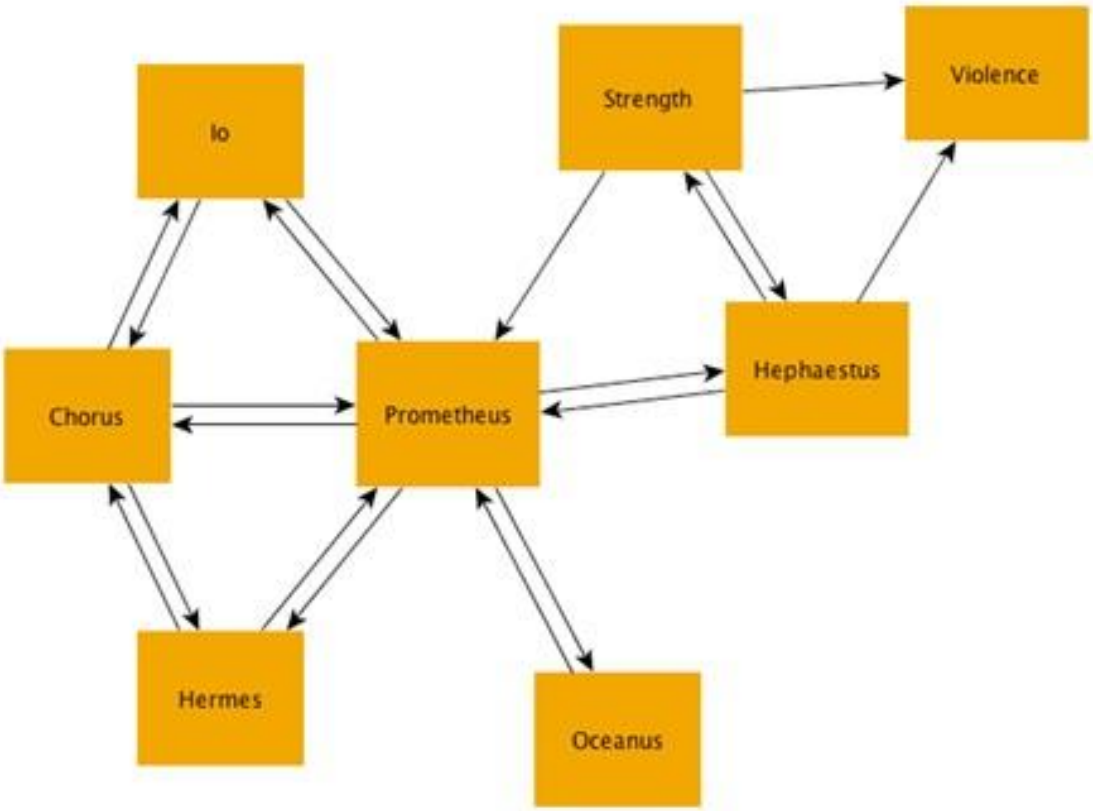


Figura 2 – Diagrama de rede social para *Suplicantes* de Ésquilo

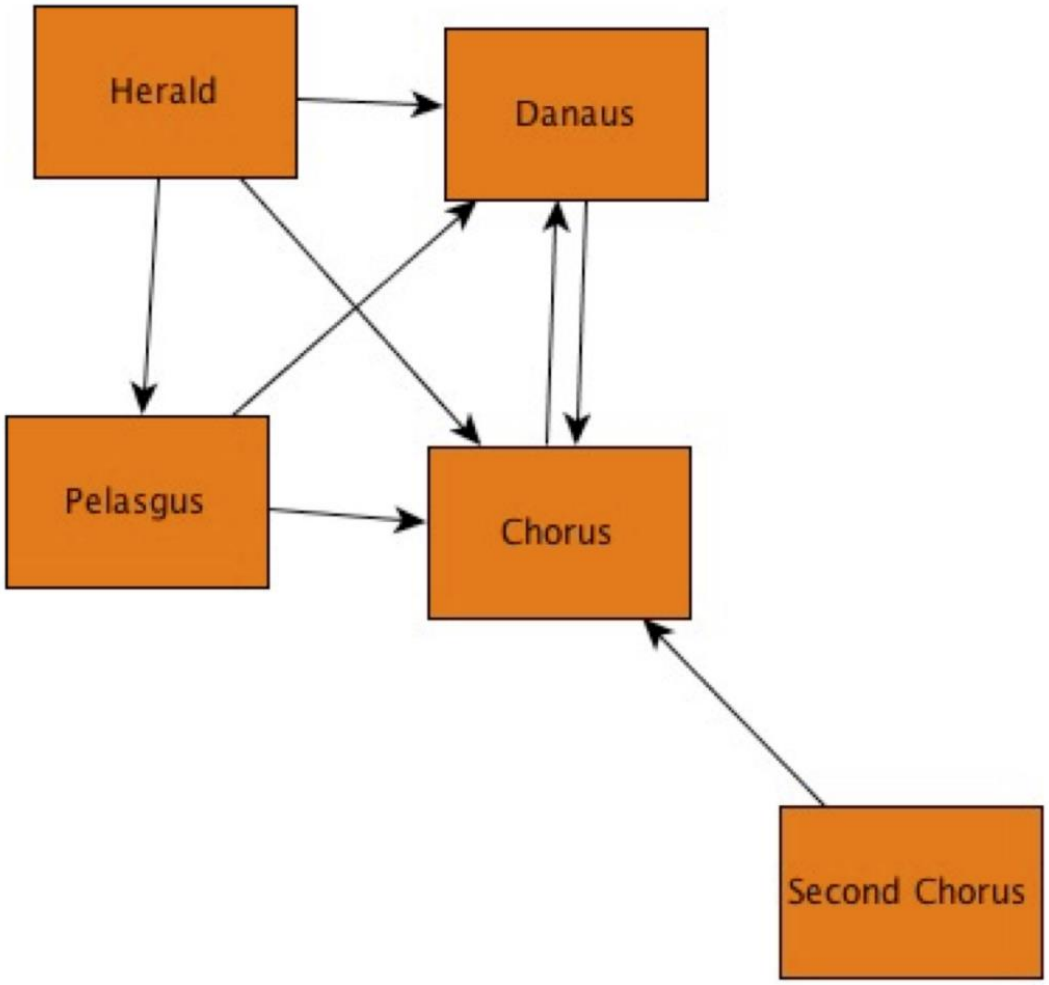


Figura 3 – Rede social em *Ajax* de Sófocles

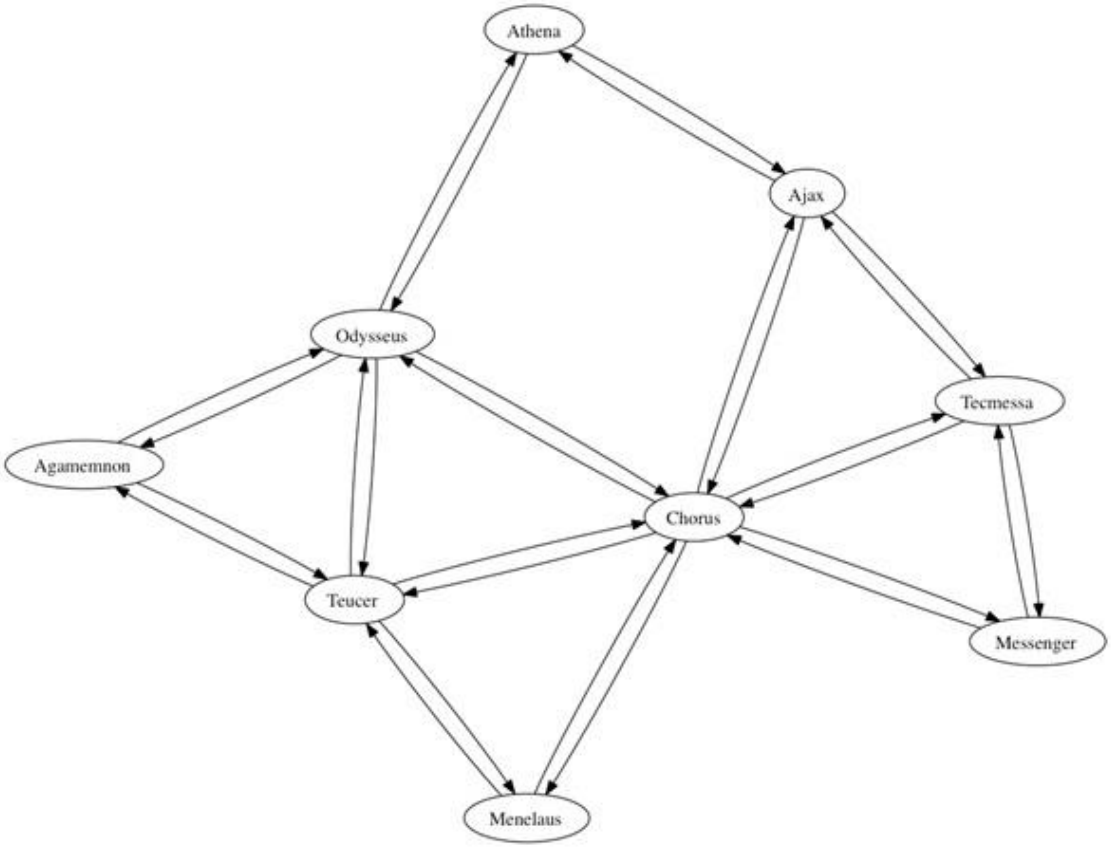
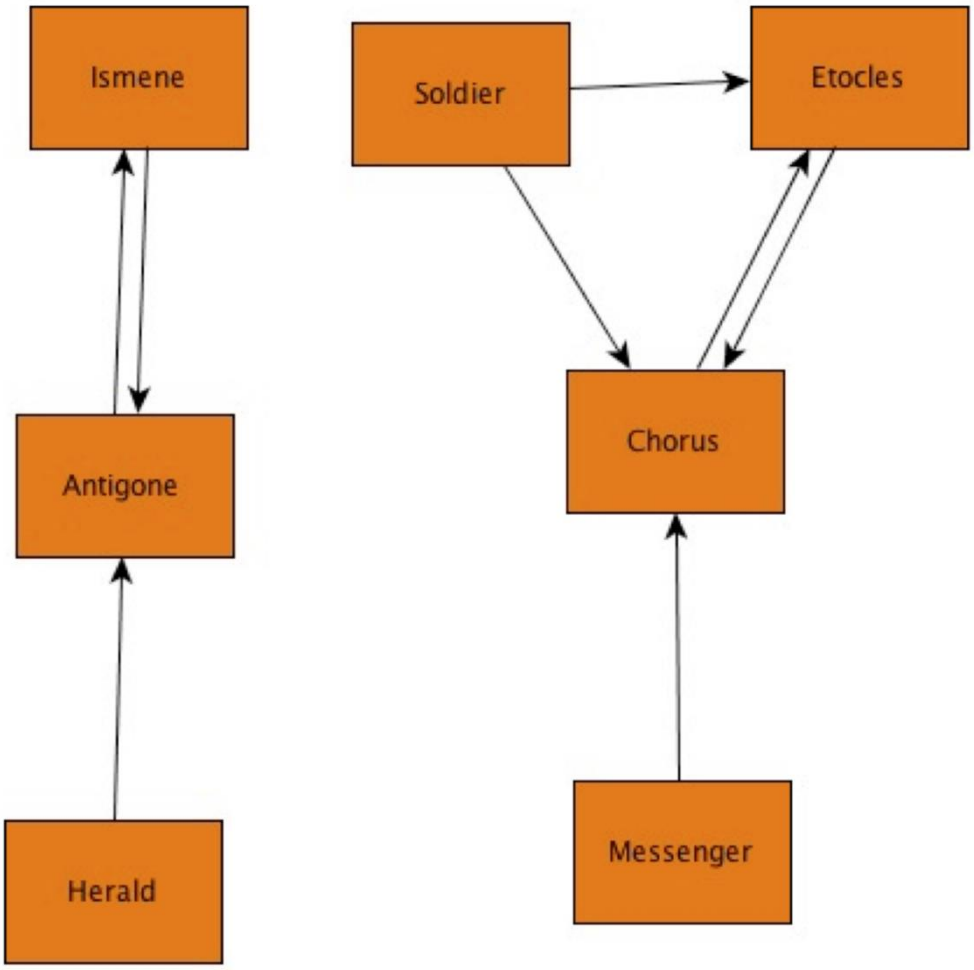


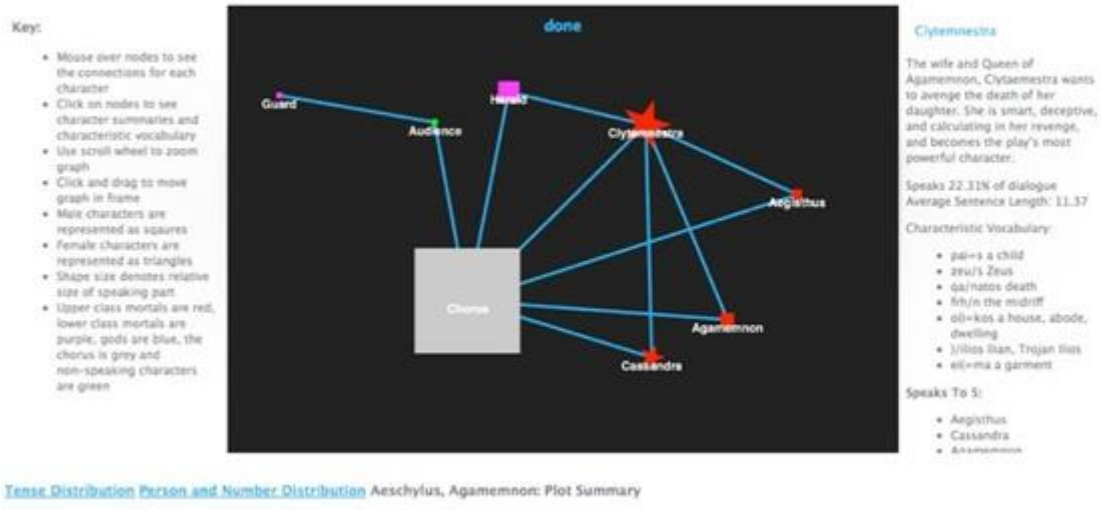
Figura 4 – Rede social em *Sete Contra Tebas* de Ésquilo



Adicionando dados linguísticos

Uma vez que estes gráficos se enquadram em alguns tipos básicos, o método correto para definir e etiquetar as arestas entre os nós é o aspecto mais interessante destas visualizações. O gráfico da rede social se torna um gancho facilmente compreensível para transmitir outras informações sobre o texto com base em como etiquetamos os nós e as arestas. Vários protótipos em evolução destes gráficos de redes sociais para as tragédias de Ésquilo, Sófocles e Eurípides estão disponíveis *on-line* em http://daedalus.umkc.edu/ VisualExplorer. Nestes gráficos, cada página da web tem um resumo do enredo da peça como um cabeçalho com um diagrama de rede social, como mostrado abaixo.

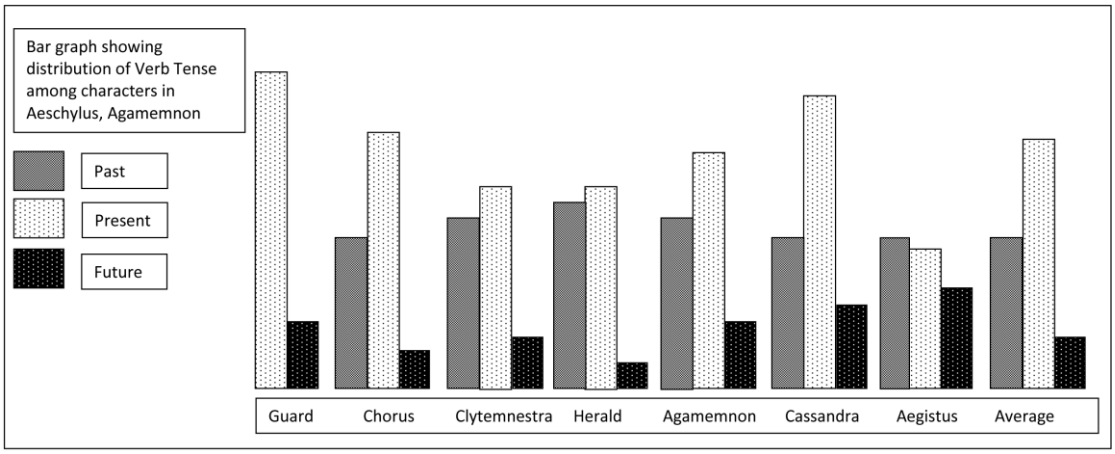
Figura 5 – Gráfico da rede social para *Agamêmnon* de Ésquilo



Nesta rede social, cada personagem é representado como um nó neste gráfico; o tamanho do nó indica a proporção relativa do diálogo recitado por cada personagem e a cor a forma do nó indicam o sexo e a classe social de cada personagem (os mortais de classe alta são vermelhos, os mortais de classe baixa, violetas, os deuses são azuis, o coro é cinza e os personagens sem fala, verdes). Quando um usuário clica em um nó dentro do

gráfico, os dados específicos do personagem aparecem na coluna à direita. Estes dados incluem uma descrição escrita especialmente para o personagem, dados sobre o comprimento médio da oração recitada pelo personagem, termos-chave associados com o personagem calculado usando a pontuação TF x IDF[146] e uma lista de outros personagens a quem o personagem se dirige. *Links* adicionais na parte inferior de cada página dão acesso a um mapa que mostra a distribuição relativa dos verbos do passado, presente e futuro entre os personagens da peça.

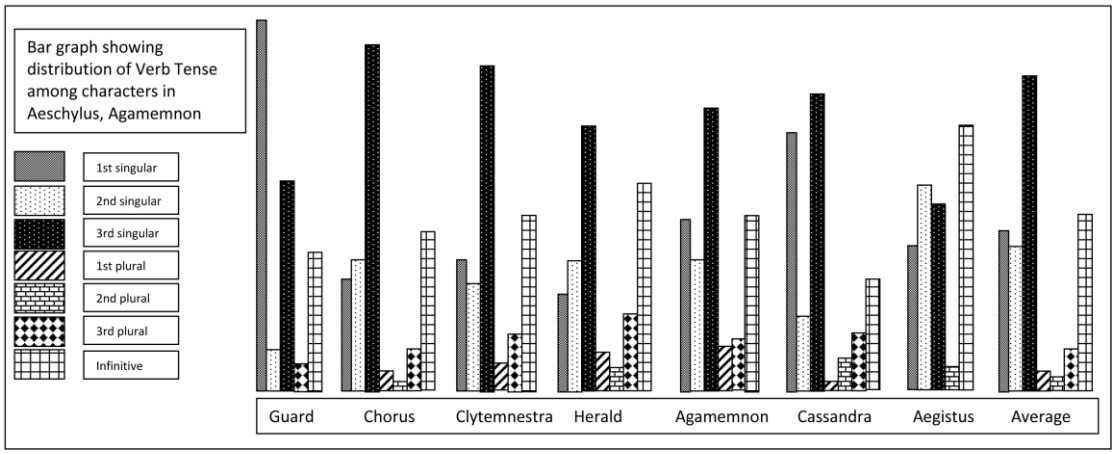
Figura 6 – Gráfico de barras mostrando a distribuição de tempo verbal entre os personagens de *Agamêmnon* de Ésquilo



[146] Para uma discussão da técnica usada para extrair estas frases-chave, consulte Jeffrey A. Rydberg-Cox, "Keyword Extraction From Ancient Greek Literary Texts". *Literary and Linguistic Computing* 17, n. 2 (2002): 231-244.

Um segundo gráfico que mostra a distribuição de número e pessoas verbais entre os personagens da peça.

Figura 7 – Gráfico de barras mostrando a distribuição da pessoa verbal entre os personagens de Ésquilo



Gráficos como estes permitem aos leitores começar a ver e considerar como características gramaticais quantificáveis como estas acompanham aspectos literários ou tramas na peça.

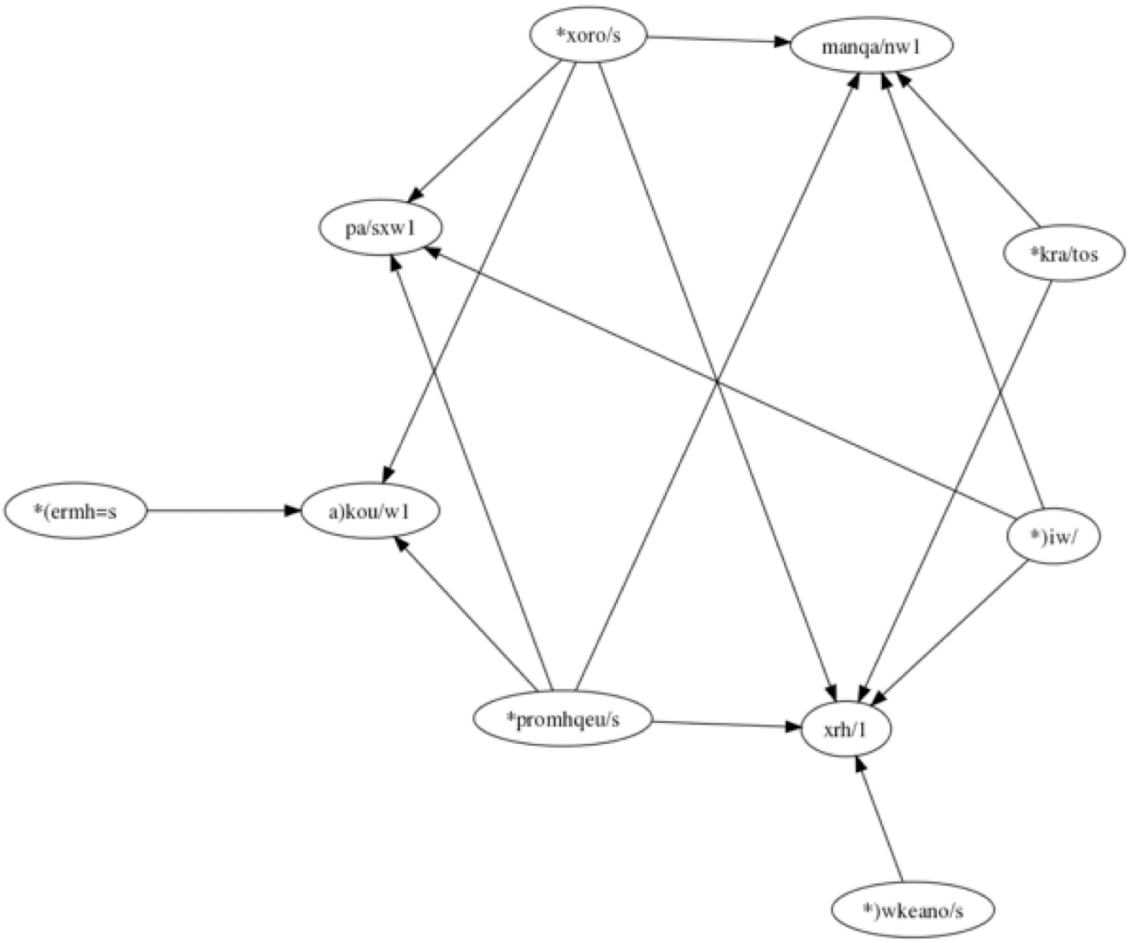
Direções futuras

Estas visualizações continuam a evoluir e mudar à medida que experimentamos com outras visualizações que podem ser úteis para os leitores que trabalham com esses textos. Este trabalho está-se movendo em várias direções diferentes. Em primeiro lugar, os gráficos como mostrado acima fornecem apenas percentagens brutas de frequência sem indicação da significância estatística das variações entre os diferentes falantes. Estamos trabalhando nas visualizações para integrar *t-scores* para este gráfico para que os usuários possam ver o que está fora da faixa esperada. Em segundo lugar, estamos explorando outros tipos de dados que podem ser introduzidos nestes gráficos, como uma métrica de correlação de vocabulário que expresse o grau de sobreposição entre as palavras usadas pelos dois personagens e um gráfico que organize as palavras recitadas por cada par de falantes permitindo aos leitores ver quais palavras têm uma relação mais

próxima com o personagem[147]. Também estamos trabalhando em visualizações que incorporam as palavras como se fossem atores dentro da rede social. Para esta visualização, as palavras com associação mais próxima com cada personagem são calculadas como uma pontuação TF x IDF com as cinco palavras principais para cada personagem incluídas na rede social como o objeto de um relacionamento social com seu falante servindo, portanto, como intermediários entre os personagens nas peças.

[147] Ver J. F. Burrows, *Computation Into Criticism: A Study of Jane Austen's Novels and An Experiment in Method* (Oxford [Oxfordshire]; New York: Clarendon Press, Oxford University Press, 1987), que constrói este tipo de gráficos para as palavras muito comuns associadas com personagens dos romances de Jane Austen. Há muitos modelos para os tipos de dados linguísticos que podem ser representados graficamente nesta interface. Além do trabalho de base Burrow, estamos examinando abordagens baseadas em *corpus* para a variação linguística em Douglas Biber, *Variation Across Speech and Writing* (Cambridge [England]; New York: Cambridge University Press, 1988); Douglas Biber, *Dimensions of Register Variation: A Cross-Linguistic Comparison* (Cambridge; New York: Cambridge University Press, 1995); Douglas Biber, Susan Conrad, and Randi Reppen, *Corpus Linguistics: Investigating Language Structure and Use* (Cambridge; New York: Cambridge University Press, 1998); Susan Conrad and Douglas Biber, *Variation in English: Multi-Dimensional Studies* (Harlow, England; New York: Longman, 2001); Randi Reppen, Susan M. Fitzmaurice, and Douglas Biber, *Using Corpora to Explore Linguistic Variation* (Amsterdam; Philadelphia: J. Benjamins, 2002); and Douglas Biber, Ulla Connor, and Thomas A. Upton, *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure* (Amsterdam; Philadelphia: John Benjamins Pub. Co., 2007).

Figura 8 – Gráfico mostrando as conexões entre os personagens de *Prometeu Acorrentado* Ésquilo com os verbos

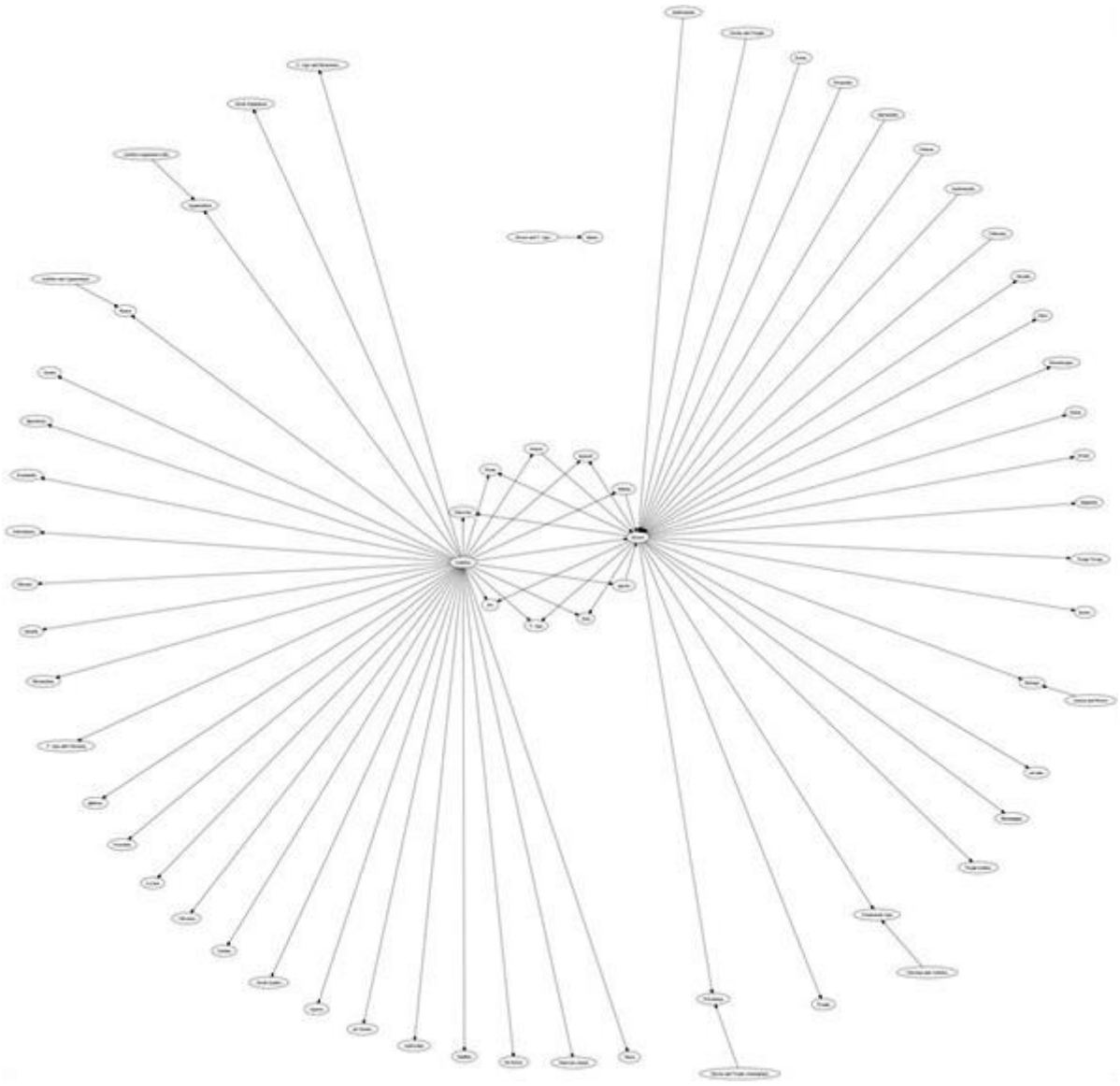


Esta é uma possibilidade tentadora que enfrenta algumas dificuldades práticas. Se incluirmos todas as palavras recitadas por um personagem particular, obtemos um gráfico ininteligível porque há muitos nós, mas se nos concentrarmos no vocabulário característico de cada personagem, como definido pela pontuação TF x IDF, há poucas sobreposições entre os personagens. Listas de palavras que são interessantes selecionadas manualmente para um leitor específico parecem apresentar os resultados mais promissores, mas esta abordagem não quantifica amplamente a menos que fosse construída como um mecanismo de busca interativa que sugira palavras candidatas e forneça um sistema de navegação interativa aos leitores envolvidos em leitura atenta com auxílio de computador.

Finalmente, também estamos trabalhando em expandir esta abordagem para outras obras em outras línguas e outros gêneros. Este trabalho também é muito preliminar, mas as primeiras visualizações são muito intrigantes. Se, por exemplo, examinarmos a *Ilíada* e a *Odisseia*, onde encontramos uma ampla gama de personagens, os gráficos dos relacionamentos de personagens, até mesmo em um nível mais amplo, nos dão uma ideia da natureza destes textos. Por exemplo, se estamos construindo uma rede social para a *Ilíada* e nos concentramos nos personagens que falam com Aquiles ou Heitor, o gráfico inicial é muito interessante e instiga perguntas sobre as possíveis diferenças na linguagem usada entre os personagens no centro deste gráfico comparado com a linguagem usada pelos personagens nas extremidades[148].

[148] Hilary Susan Mackie, *Talking Trojan: Speech and Community in the Iliad* (Lanham: Rowman & Littlefield Publishers, 1996).

Figura 9 – Gráfico mostrando aqueles que falam de Aquiles e Heitor na *Ilíada* de Homero



Referências

BAMMAN, D.; CRANE, G. The Design and Use of a Latin Dependency Treebank. **Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT2006)**, 2006.

BAMMAN, D.; MAMBRINI, F.; CRANE, G. An Ownership Model of Annotation: The Ancient Greek Dependency Treebank. **Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT8)**, 2009.

BAMMAN, D.; M., PASSAROTTI; CRANE, G.; RAYNAUD, S. A Collaborative Model of Treebank Development. **Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories,** 2007.

BIBER, D. **Dimensions of Register Variation: A Cross-Linguistic Comparison**. Cambridge; New York: Cambridge University Press, 1995.

_____. **Variation Across Speech and Writing**. Cambridge; New York: Cambridge University Press, 1988.

BIBER, D.; CONRAD, S.; REPPEN, R. **Corpus Linguistics: Investigating Language Structure and Use**. Cambridge; New York: Cambridge University Press, 1998.

BIBER, D.; CONNOR, U.; UPTON, T. A. **Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure.** Amsterdam; Philadelphia: John Benjamins Pub. Co., 2007.

BURROWS, J. F. **Computation Into Criticism: A Study of Jane Austen's Novels and An Experiment in Method.** Oxford [Oxfordshire]; New York: Clarendon Press; Oxford University Press, 1987.

CONRAD, S.; BIBER, D. **Variation in English: Multi-Dimensional Studies.** Harlow; New York: Longman, 2001.

CRANE, G. What Do You Do with a Million Books? **D-Lib Magazine D-Lib Magazine**, v. 12, n. 3, 2006.

ELSON, D. K.; MCKEOWN, K. R. A Tool for Deep Semantic Encoding of Narrative Texts. **Proceedings of the ACL-IJCNLP 2009 Software Demonstrations**, 2009.

_____. Automatic Attribution of Quoted Speech in Literary Narrative. **Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence**, 2010.

ELSON, D. K.; MCKEOWN, K. R. Extending and Evaluating a Platform for Story Understanding. **Proceedings of the AAAI 2009 Spring Symposium on Intelligent Narrative Technologies II**, 2009.

ELSON, D. K.; DAMES, N.; MCKEOWN, K. R. Extracting Social Networks From Literary Fiction. **Proceedings of the 48Th Annual Meeting of the Association for Computational Linguistics**, 2010.

LEE, J.; HAUG, D. Porting An Ancient Greek and Latin Treebank. **Proc. LREC**, 2010.

MACKIE, H. S. **Talking Trojan: Speech and Community in the Iliad**. Lanham: Rowman & Littlefield Publishers, 1996.

MORETTI, F. **Graphs, Maps, Trees: Abstract Models for a Literary History**. London; New York: Verso, 2005.

_____. Network Theory, Plot Analysis. **New Left Review**, n. 68, p. 80-102, 2011.

_____. Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740-1850). **Critical Inquiry**, v. 36, n. 1, p. 134-158, 2009.

_____. **The Novel**. Princeton: Princeton University Press, 2006.

MUTTON, P. Inferring and Visualizing Social Networks on Internet Relay Chat. **Proceedings Eighth International Conference on Information Visualisation IV,** 2004.

REPPEN, R.; FITZMAURICE, S. M.; BIBER, D. **Using Corpora to Explore Linguistic Variation.** Amsterdam; Philadelphia: J. Benjamins, 2002.

RUSSELL, G.; TUITE, C. **Romantic Sociability: Social Networks and Literary Culture in Britain**. Cambridge; New York: Cambridge University Press, 2002. p. 1770-1840.

RYDBERG-COX, J. A. Keyword Extraction From Ancient Greek Literary Texts. **Literary and Linguistic Computing**, v. 17, n. 2, p. 231-244, 2002.

STILLER, J.; HUDSON, M. Weak Links and Scene Cliques Within the Small World of Shakespeare. **Journal of Cultural and Evolutionary Psychology**, v. 3, n. 1, p. 57-73, 2005.

STILLER, J.; NETTLE, D.; DUNBAR, R. I. M. The Small World of Shakespeare's Plays, **Human Nature**, v. 14, n. 4, p. 397-408, 2003.

Agência Brasileira do ISBN
ISBN 978-85-69395-01-0