PESQUISA EM EDUCAÇÃO: O WORDSMITH COMO FERRAMENTA DE EXPLORAÇÃO DE CORPORA

Maria Zuleide da Costa Pereira[1]
Samara Wanderley Xavier Barbosa[2]
Rafael Ferreira de Souza Honorato[3]
Nathália Fernandes Egito Rocha[3]

RESUMO

Este texto constitui-se a partir da implementação das ações de um projeto do Programa Institucional de Bolsas de Iniciação Científica (PIBIC) da UFPB, intitulado "Os Sentidos do Currículo nas Escolas da Rede Municipal de Ensino de João Pessoa/PB", e desenvolvido no período de 2013 a 2014. O objetivo do plano/projeto é destacar o papel do software *Worsmith Tools 6*, como ferramenta de análise de corpora, na exploração dos sentidos de educação, currículo e ensino nos documentos curriculares analisados, que são os documentos de políticas curriculares nacionais (Lei de Diretrizes e Bases de nº 9394/96, Parâmetros Curriculares Nacionais de 1ª a 4ª série, Diretrizes Curriculares Gerais para Educação Básica e Diretrizes Curriculares para o Ensino Fundamental de Nove Anos) e os locais (Projetos Político-Pedagógicos de nove escolas da Rede Municipal de Ensino). Dessa forma, mostramos os recursos do conjunto de ferramentas utilizadas, exemplificando de que modo elas contribuíram para uma análise documental mais exata e confiável do que outras perspectivas de análise linguística permitiriam. De fato, ao mesmo tempo em que facilitaram às possíveis articulações entre educação, currículo e ensino, o conjunto de ferramentas em questão, como argumenta Sardinha (2004), nos deu a possibilidade de analisar vários aspectos da linguagem, tais como: a composição lexical, o tema dos textos selecionados e a organização da retórica e composicional dos gêneros discursivos. Metodologicamente, organizamos os documentos em análise num conjunto de textos informatizados, de tal forma que se tornaram adequados para o pesquisador analisar, sempre tendo em vista a autenticidade, a legibilidade e a extensão dos textos, e a seleção criteriosa dos enunciados que comporiam o corpus. Para empreender a análise propriamente dita, decidimos utilizar o *Worsmith Tools 6* e suas três ferramentas o *Wordlist,* o *Concord* e o *Keywords*, cada uma delas com suas especificidades. A partir das primeiras análises, elaboramos gráficos e tabelas, destacando os sentidos de currículo mais recorrentes, cruzando essas informações com articulação com os preceitos de educação e ensino, de forma a verificar as relações entre os sentidos associados a essas diferentes categorias e suas implicações para o contexto educacional a que elas se referem, levando em conta aspectos culturais, políticos e ideológicos que transparecem ao longo dos documentos. Assim, a pesquisa realizada mostrou que no campo da Educação e do Currículo, e mais especificamente a análise documental nesses campos, a aproximação com a Linguística de Corpus é fundamental à medida que permitiu uma visão mais abrangente dos textos estudados e sob uma multiplicidade de perspectivas, o que contribuiu para tornar os achados mais consistentes e para compreender os significados mais profundos dos textos, abarcando aspectos culturais, ideológicos, políticos e identitários.

Palavras-chave: Linguística. Sentidos. Educação

[1] Orientadora.
[2] Colaboradora.
[3] Bolsista PIBIC/CNPq.

INTRODUÇÃO

Este artigo tem como objetivo refletir sobre as contribuições da Linguagem de Corpus e do *Wordsmith Tools* como ferramenta de análise de corpora, na exploração dos sentidos de currículo e suas articulações com a educação e o ensino nos documentos curriculares analisados, que são os documentos de políticas curriculares nacionais (Lei de Diretrizes e Bases de nº 9394/96, Parâmetros Curriculares Nacionais de 1ª a 4ª série, Diretrizes Curriculares Gerais para Educação Básica e Diretrizes Curriculares para o Ensino Fundamental de Nove Anos) e os locais (Projetos Político-Pedagógicos de nove escolas da Rede Municipal de Ensino) no que concerne, especificamente, à negociação de marcas identitárias produzidas pelo discurso dos professores que estão arcados nos documentos que nos propomos a analisar.

A nossa proposta de investigação é considerável quando olhamos para o currículo como resultado das articulações discursivas dos seus vários lugares – prescrição, ação e dimensão oculta – assim partimos de uma concepção de currículo como prática de significação (SILVA, 1995). Logo, "(...) inextricavelmente, centralmente, vitalmente, envolvido naquilo que somos, naquilo que nos tornamos: na nossa identidade, na nossa subjetividade" (SILVA, 2007, p. 15)

Assim, entendemos que nos documentos oficiais nacionais e locais existem representações identitária construídas pelos sujeitos que exercem as práticas pedagógicas em consonância com os contextos onde os sujeitos estão inseridos. O nosso interesse pelo estudo dos sentidos do currículo pode se dizer que vai além dos documentos e propostas curriculares quando existe uma proposta de investigar como essas propostas estão mergulhadas em construções políticas existentes nos espaços sociais. Elencamos que essa proposta de investigação política está inserida no plano dois desse projeto.

A partir dos estudos de Silva (1995), observamos que as marcas de poder no currículo passam pela investigação do campo de representações, ou melhor, expresso por Silva como regimes de representações. O que mais interessa é o fato de que esses regimes são constituídos por diferentes discursos que são imbricados de sinais e vestígios das relações de poder instituídas.

Para a nossa investigação, traçamos como procedimento metodológico que é contemplado com a coleta de dados através dos documentos que já citamos acima, sendo os nacionais disponibilizados online e os locais, no caso dos Projetos Político-Pedagógicos, precisou-se ir até as escolas. Esses procedimentos caracterizam o que Gonsalves (2003) chama de uma pesquisa de campo. Após essa coleta, começamos o processo de preparação dos documentos através da utilização da Linguística de Corpus, para uma futura inserção da ferramenta *Wordsmith Tools* 6. Ainda a partir dos estudos de Gonsalves (2003), a pesquisa ainda caracterizou-se como descritiva e quanto-qualitativa.

Dentre esses procedimentos, coube a esse relatório fazer a análise do uso da ferramenta e relatar o processo que foi feito para chegar aos dados. Após, passando a outra bolsista para a devida análise a luz da Teoria do Discurso de Laclau (2005) e Laclau & Mouffe (2004).

Adiantamos que a utilização da ferramenta e da Linguística de Corpus foi bem promissora e nos forneceu elementos importantes para as nossas considerações, principalmente por termos como proposta a análise de tantos documentos o que tornaria impossível a leitura de forma manual. Nesse caso, o processo foi um esforço dos/as pesquisadores/as e a ferramenta para conseguirmos dar conta de estudarmos essa quantidade de documentos.

METODOLOGIA

A Linguística de corpus

A Linguística de *Corpus* é um tema emergente que vem assumindo uma posição de destaque entre as pesquisas que buscam sentidos e percepções a partir do discurso de certa demanda social. Um dos pesquisadores que é referencia na Linguística de *Corpus* é o Tony Berber Sardinha que, entre os pesquisadores brasileiros, vem a algum tempo utilizando a linguística de *corpus* em suas pesquisas. No primeiro momento de aproximação com esse campo de estudos, nos debruçamos em sua obra "Linguística de *Corpus*", publicada em 2004, fomos aos poucos conhecendo outros estudiosos, tais como: Stela Tagninda Universidade de São Paulo (USP) e Guilherme Fromm da Universidade Federal de Uberlândia (UFU), entre outros que se embrenham nesse campo com produções significativas.

O *Corpus* não é um elemento posterior ao computador. Antes desse evento tecnológico, a análise desse conjunto de elementos era feito através de fichas que iam sendo construídas e analisadas posteriormente. A chegada do computador fez com que a análise dos corpora[4] se tornasse mais rápida e possibilitou um volume maior de documentos. *Corpus* significa "corpo ou conjunto de documentos, dados e informações sobre determinada matéria", assim foi descrito por Ferreira (2004, p. 557).

A chegada do computador como citada anteriormente causou uma maior tranquilidade no que se refere ao tratamento dos corpora com um grande volume de informações, possibilitando uma maior confiabilidade nas pesquisas que se utilizam desses métodos, por nos permitir desenvolver atividades mais complexas e elaboração e organização de informações que contemplam o método quantitativo, principalmente nas pesquisas que são da área de humanas. Com o avanço do computador e a sua capacidade de armazenamento melhorada, essas pesquisas foram cada vez mais ocupando espaço e ganhando méritos pelos resultados.

A influência da Linguística de Corpus teve seu limiar na década de 80, pois nessa década ela foi popularizada dando acesso amplo aos pesquisadores, exercendo certa influência nas pesquisas em Linguística. Essa popularização acarretou a utilização dessa modalidade em espaços não acadêmicos, ou seja, além das pesquisas tiveram espaços empresariais que se utilizaram desses conhecimentos. No Brasil, temos um alargamento das possibilidades com essa área, podendo observar seu uso de forma tímida nas Ciências Humanas e Sociais em cursos como Pedagogia, que contempla a grande área que é a educação, mas vemos o uso dessa modalidade principalmente em estudos voltados para o processamento da linguagem natural, ciências do léxico e linguística computacional.

Beber Sardinha (2004) defende a Linguística de corpus como possibilidade de exploração da linguagem, indagando que

> [...] a Linguística de Corpus ocupa-se da coleta e da exploração e corpora, ou conjunto de dados linguísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-se à exploração da linguagem por meio de evidências empíricas, extraídas por computador. (p. 3).

[4] Plural de corpus.

Uma das concepções que superamos durante os nossos estudos é que essa forma de organização das informações não é apenas um conjunto de textos informatizados e aglomerados em forma de arquivos computacionais para a análise do pesquisador. Esses arquivos dos documentos a serem analisados devem ter autenticidade dos dados, cumprir um propósito de estudo, um conteúdo escolhido criteriosamente, ser legível pelo computador, ter representatividade de uma língua e ser possuidor de uma vasta extensão, ou seja, nem todo conjunto de informações é um corpus.

Beber Sardinha (2004) nos apresentou alguns critérios e uma tipologia de corpus descrita nas literaturas com propósito de definição do conteúdo e do propósito dos corpora, são eles:

a. Modos: falados (transcrição da fala) ou escritos;

b. Tempo: sincrônico, diacrônico, contemporâneo, histórico;

c. Seleção: de amostragem (amostra finita da linguagem, é estático), monitor (dinâmico, reciclável), equilibrado (balanceado, distribuição de textos em quantidades semelhantes);

d. Conteúdo: especializado (gêneros), regionais ou dialetais, multilíngue;

e. Autoria: de aprendiz (falantes não-nativos), de língua nativa (falantes nativos);

f. Disposição interna: paralelo (original e tradução), alinhado/comparável;

g. Finalidade: de estudo (descrição de corpus), de referência (para contrastar com o corpus de estudo), de treinamento (para desenvolvimento de aplicações e ferramentas de análise).

O *Wordsmith tools* 6 no contexto da pesquisa

Criado e escrito por *Mike Scott* e publicado pela *Oxford University Press* há cinco anos, o *software Wordsmith Tools* está disponível na *web*[5] no próprio site de seu criador, a versão 6.0 (demo), sendo que para ter acesso a todos os recursos, o usuário deverá adquirir uma licença, ou seja, é um programa de caráter privado.

Nossa primeira aproximação com o programa nos colocou frente a algumas dificuldades, entre elas a configuração que contempla a língua inglesa e nós operadores do programa enquanto iniciantes na pesquisa ligados a uma graduação não temos domínio de outro código linguístico ou estamos em processo de estudos para o alcance desse domínio de outra língua. A operacionalização do programa também não é tão simples e exige do operador/pesquisador um conhecimento médio sobre informática, mas a necessidade e curiosidade foram grandes impulsionadores para a superação dos entraves que se apresentaram durante o caminho.

A partir dos estudos de Beber Sardinha (2004), vimos que a utilização da ferramenta na pesquisa potencializa a "análise de vários aspectos da linguagem, como a composição lexical, a temática de textos selecionados e a organização retórica e composicional de gêneros discursivos." (p.86)

[5] http://www.lexically.net/index.html - neste site encontra-se também o tutorial do programa e um link direcionado ao Wordsmith Tools Groups (grupo de discussão em que os usuários postam suas dúvidas sobre o programa, sendo estas respondidas pelo próprio Mike Scott e/ou usuários do programa), dentre outros.

Usufruímos da sua empregabilidade em nossa pesquisa, que se constitui no campo da educação, especificamente no campo do currículo de forma não obstante, assim pretendemos ampliar e dar maior confiabilidade aos nossos resultados, como também disseminar o uso da linguística computacional e de corpus nas pesquisas em educação. Compreendemos que novos métodos abrem novas possibilidades e abarcam novas demandas.

O *Wordsmith tools* é um conjunto de ferramentas, utilitários, instrumentos e funções. Apropriamo-nos nesta pesquisa apenas das ferramentas básicas do programa, que mediante os objetivos propostos nos foram muito eficazes. Nos tópicos à frente nos limitaremos a destacar apenas o que foi aplica nessa pesquisa.

Sardinha (2009, p.9), divulga as três ferramentas do programa e dá as suas características.

> ***Wordlist*:** produz lista de palavra contendo todas as palavras do arquivou arquivos selecionados, elencadas em conjunto com suas frequências absolutas e percentuais. Também compara listas, criando listas de consistência, onde é informado em quantas listas cada palavra aparece.
>
> ***Concord*:** realiza concordâncias, ou listagens de uma palavra específica (o 'nódulo', *node word ou search word)* juntamente com parte do texto onde ocorreu. Oferece também lista de colocados, isto é, palavras que ocorreram perto do nódulo.
>
> ***KeyWords*:** extrai palavras de uma lista cujas freqüências são estatisticamente diferentes (maiores ou menores) de que as frequências das mesmas palavras num outro corpus (de referência). Calcula também palavras-chaves chave, que são chaves em vários textos.

Essas ferramentas apresentam outros utilitários que foram utilizados na pesquisa que hora descrevemos. Elas se dividem assim:

Wordlist

a. Lista de palavras individuais (wordlist)

b. Lista de multipalavras (wordlist, clusters activated)

c. Lista de palavras de consistência individuais (detailed consistency)

d. Lista de multipalavras de consistência (detailed consistency, clusters activated)

e. Lista de dimensões e densidade lexical (statistics)

Keywords

a. Lista de palavras-chave (keywords)

b. Banco de dados de listas de palavras-chave (database)

c. Lista de palavras-chave chave (key keywords)

d. Lista de palavras-chave associadas (associates)

e. Lista de agrupamentos textuais (chunks)

f. Gráfico de distribuição de palavra-chave (keyword plot)

g. Listagem de elos entre palavras-chave (keywords plot links)

Concord

a. Concordância (concordance)

b. Lista de colocados (collocates)

c. Lista de agrupamentos lexicais (clusters)

d. Lista de padrões de colocados (patterns)

e. Gráfico de distribuição de palavra de busca (plot)

RESULTADOS E DISCURSÕES

Para construirmos essa descrição da pesquisa, fomos cautelosos em realizar um procedimento para coleta e tratamento dos dados e, com o intuito de contribuir para a compreensão da metodologia utilizada no trabalho, faz-se necessária uma descrição detalhada dos procedimentos adotados.

A abordagem metodológica que decidimos utilizar - a Linguística de Corpus - tem como pré-requisito a existência de dois corpora, um de estudo e outro de referência. Como já apontamos em nossa metodologia, existe a classificação para o tamanho dos corpora, assim apresentamos o quadro abaixo organizado por Beber Sardinha.

TABELA 1 - Classificação relativa ao tamanho dos corpora

Tamanho em palavras	Classificação
Menos de 80 mil	Pequeno
80 e 250 mil	Pequeno-Médio
250 mil e 1 milhão	Médio
1 milhão e 10 milhões	Médio- Grande
10 milhões ou mais	Grande

FONTE: Beber Sardinha 2004

Nos nossos estudos, vimos que a representatividade de um corpus está ligada ao seu tamanho, assim, quanto maior a extensão do corpus maior sua representatividade. A partir da classificação de Sardinha (2004), classificamos o corpus dessa pesquisa como Médio-Grande, por ter 3.496.732 palavras.

Os nossos corpora de estudos é formado por textos dos Projetos Políticos Pedagógicos (PPPs), das nove escolas da Rede Municipal de João Pessoa, os documentos das políticas curriculares nacionais, como Lei de Diretrizes e Bases nº 9394/96, Parâmetros Curriculares Nacionais 1ª a 4ª série, Diretrizes Curriculares Gerais para Educação Básica e Diretrizes Curriculares para o Ensino Fundamental de Nove Anos.

A partir desse conjunto de documentos, nos debruçamos sobre a classificação dos nossos corpora e resgatamos o que foi mostrado no tópico (2.1), onde elencamos alguns critérios a partir de Beber Sardinha (2004) sobre a tipologia de corpus, com o propósito de definição do conteúdo e do propósito dos corpora e chegamos à definição do nosso corpus investigativo como:

a. Modos: escritos;

b. Tempo: sincrônico e contemporâneo;

c. Seleção: de amostragem;

d. Conteúdo: especializado;

e. Autoria: de língua nativa;

f. Disposição interna: original em português;

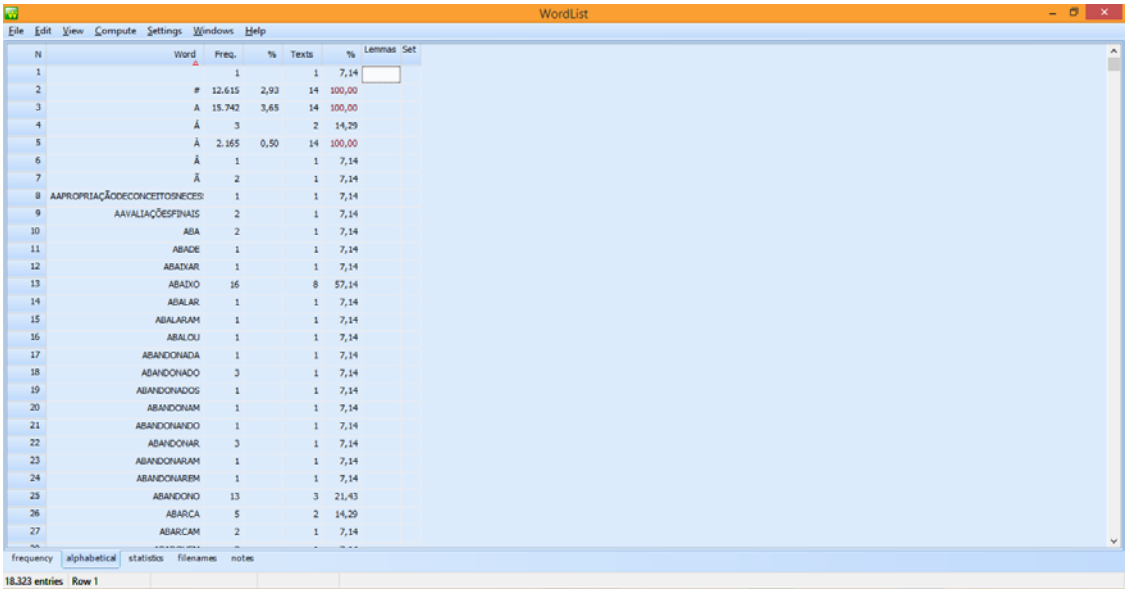
g. Finalidade: de estudo.

O *Wordlist* é responsável por criar uma lista de palavras individuais, essa lista vai conter as informações todas as palavras do arquivo ou arquivos selecionados, elencados juntamente com suas frequências absolutas e percentuais. Essa lista pode ser exibida de duas formas, em ordem alfabética ou pela frequência. Segundo nos aponta Beber Sardinha (1999) o *wordlist* é

> [...] propicia a feitura de listas de palavras. O programa é pré-definido para produzir, a cada vez, duas listas de palavras, uma ordenada alfabeticamente (identificada pela letra 'A' entre parênteses) e outra classificada por ordem de
>
> frequência das palavras (com a palavra mais frequente encabeçando a lista). Cada uma destas listas é apresentada em uma janela diferente, e juntamente com as duas janelas correspondentes à lista alfabética ('A') e por frequência ('F'), o programa oferece uma terceira janela ('S') na qual aparecem estatísticas relativas aos dados usados para produção das listas. Assim, para cada vez que o *WordList* é chamado para fazer uma lista de palavras, três janelas são produzidas: uma contendo uma lista de palavras ordenada por ordem alfabética, outra com uma lista classificada pela frequência das palavras, e uma terceira janela com estatísticas simples a respeito dos dados.

Essa ferramenta se consolida como fundamental, uma vez que é a partir dela que a lista de *keywords* será produzida como também vai possibilitar a exploração dos corpora de várias formas, em diferentes vertentes, depender do objeto de estudo da pesquisa.

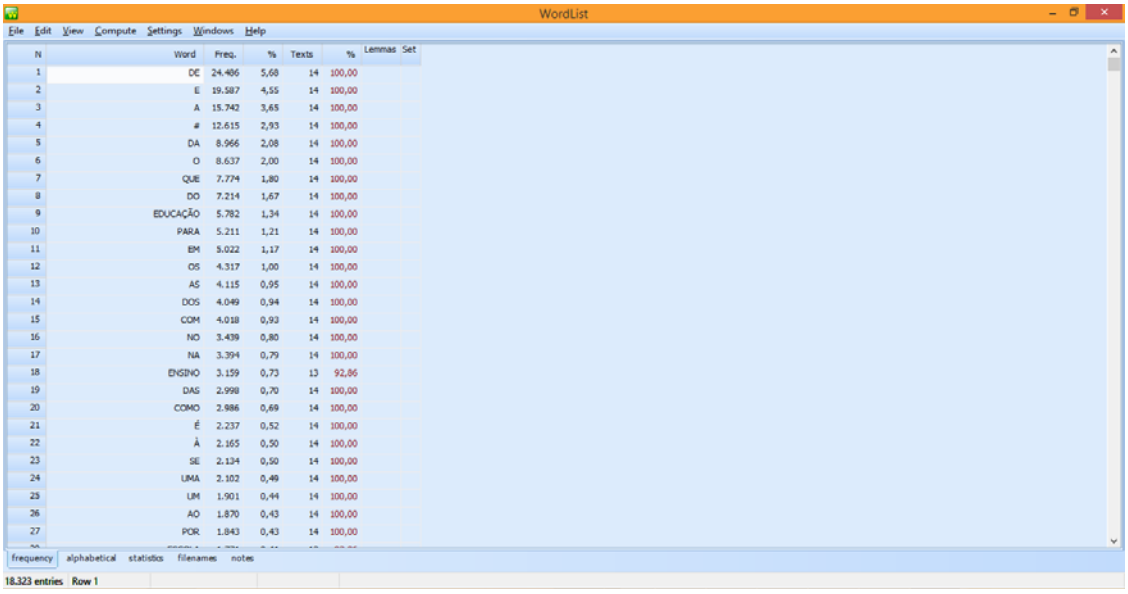
FIGURA 1 - *Wordlist* dos corpora em ordem alfabética



Fonte: Organizado pelos/as autores/as

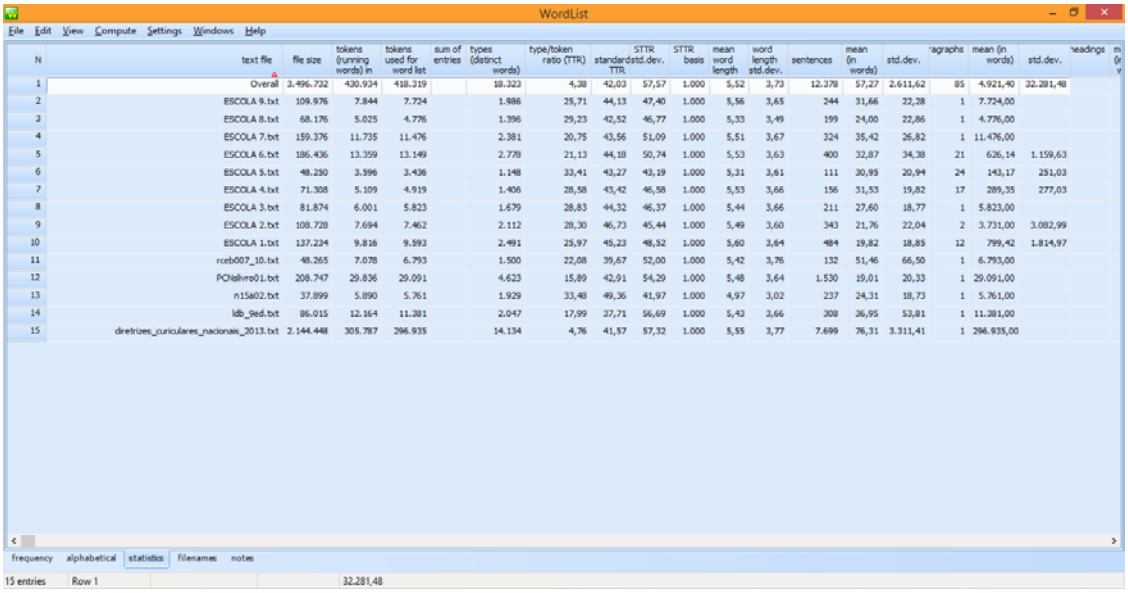
Não conseguimos, através das nossas pesquisas e dos treinamentos, determinar o que seria o espaço em branco na primeira linha da Fig. 1, mas observamos que ele está representando 7,14% dos nossos corpora e aparece em apenas um dos textos. O "#" por sua vez é a representação dos numerais e demais caracteres presentes nos corpora que o programa não conseguiu ler, logo abaixo se inicia a lista de palavras em ordem alfabética. A Fig. 2 ilustra a tema do *wordlist* por ordem de frequência, mostrando a quantidade de vezes que cada palavra apareceu nos nossos corpora de estudos.

FIGURA 2 - *Wordlist* dos corpora ordenada pelas frequências



Fonte: Organizado pelos/as autores/as

FIGURA 3 - *Wordlist* dos corpora, ferramenta *statistics*.



N	text file	file size	tokens (running words) in	tokens used for word list	sum of entries	types (distinct words)	type/token ratio (TTR)	STTR standard TTR	std.dev.	STTR basis	mean word length	word length std.dev.	sentences	mean (in words)	std.dev.	ragraphs	mean (in words)	std.dev.	eadings
1	Overall	3.496.732	430.934	418.319		18.323	4,38	42,03	57,57	1.000	5,52	3,73	12.378	57,27	2.611,62	85	4.921,40	32.281,48	
2	ESCOLA 9.txt	109.976	7.844	7.724		1.986	25,71	44,13	47,40	1.000	5,56	3,65	244	31,66	22,28	1	7.724,00		
3	ESCOLA 8.txt	68.176	5.025	4.776		1.396	29,23	42,52	46,77	1.000	5,33	3,49	199	24,00	22,86	1	4.776,00		
4	ESCOLA 7.txt	159.376	11.735	11.476		2.381	20,75	43,56	51,09	1.000	5,51	3,67	324	35,42	26,82	1	11.476,00		
5	ESCOLA 6.txt	186.436	13.359	13.149		2.778	21,13	44,18	50,74	1.000	5,53	3,63	400	32,87	34,38	21	626,14	1.159,63	
6	ESCOLA 5.txt	48.250	3.596	3.436		1.148	33,41	43,27	43,19	1.000	5,31	3,61	111	30,95	20,94	24	143,17	251,03	
7	ESCOLA 4.txt	71.308	5.109	4.919		1.406	28,58	43,42	46,58	1.000	5,53	3,66	156	31,53	19,82	17	289,35	277,03	
8	ESCOLA 3.txt	81.874	6.001	5.823		1.679	28,83	44,32	46,37	1.000	5,44	3,66	211	27,60	18,77	1	5.823,00		
9	ESCOLA 2.txt	108.728	7.694	7.462		2.112	28,30	46,73	45,44	1.000	5,49	3,60	343	21,76	22,04	2	3.731,00	3.082,99	
10	ESCOLA 1.txt	137.234	9.816	9.593		2.491	25,97	45,23	48,52	1.000	5,60	3,64	484	19,82	18,85	12	799,42	1.814,97	
11	rceb007_10.txt	48.265	7.078	6.793		1.500	22,08	39,67	52,00	1.000	5,42	3,76	132	51,46	66,50	1	6.793,00		
12	PCNalivro01.txt	208.747	29.836	29.091		4.623	15,89	42,91	54,29	1.000	5,48	3,64	1.530	19,01	20,33	1	29.091,00		
13	n15a02.txt	37.099	5.890	5.761		1.929	33,48	49,36	41,97	1.000	4,97	3,02	237	24,31	18,73	1	5.761,00		
14	ldb_9ed.txt	86.015	12.164	11.381		2.047	17,99	37,71	56,69	1.000	5,43	3,66	308	36,95	53,81	1	11.381,00		
15	diretrizes_curriculares_nacionais_2013.txt	2.144.448	305.787	296.935		14.134	4,76	41,57	57,32	1.000	5,55	3,77	7.699	76,31	3.311,41	1	296.935,00		

Fonte: Organizado pelos/as autores/as

A partir dos estudos de Beber Sardinha (1999), salientamos que a Fig. 3 é uma representação da lista de dimensões do corpus e densidade lexical, em outras palavras, é uma exposição das características estatística descritivas dos corpora. Apresenta muitas possibilidades de informação sobre os corpora, tais como: o tamanho e itens (tokens) e formas (types), a densidade lexical simples e em intervalos. O que nos apresentou um complemento numérico de 3.496.732 palavras e 18.323 palavras únicas ou types.

Após esse primeiro momento no *wordlist*, para apresentar as informações quantitativas do corpus com a intenção de mostrar a relevância em utilizar o programa *Wordsmith Tools*, passamos a explorar a ferramenta *Keywords*. Essa ferramenta nos possibilita encontrar as palavras chave dos corpora de estudos. Segundo Sardinha (2006), é "uma ferramenta das mais úteis na análise textual por computador" (p. 1), possibilita que o usuário compare uma lista de palavras dos corpora de estudos com o corpus de referência. A partir dos estudos de Beber Sardinha (2006), vimos que isso nos proporcionaria "uma seleção dos itens lexicais de seu corpus de estudo que são estatisticamente mais distintivos" (p. 2).

Todavia, para a utilização dessa ferramenta Sardinha (2006) nos mostra que é preciso dispor de dois corpora, um de estudos e outro de referencias sendo esses caracterizados como

> - um corpus de estudo, representado em uma lista de freqüência de palavras. O corpus de estudo é aquele que se pretende descrever. A ferramenta *KeyWords* aceita a análise simultânea de mais de um corpus de estudo.
> - um corpus de referência, também formatado como uma lista de freqüência de palavras. Também é conhecido como 'corpus de controle', e funciona como termo de comparação para a análise. A sua função é a de fornecer uma norma com a qual se fará a comparação das freqüências do corpus de estudo. A comparação é feita através de uma prova estatística selecionada pelo usuário (qui-quadrado ou log-likelihood). As palavras cujas freqüências no corpus de estudo forem significativamente maiores segundo o resultado da

> prova estatística são consideradas chave, e passam a compor uma listagem específica de palavras-chave. (p. 3)

No nosso caso, o corpus de referência tornou-se uma dificuldade, uma vez que os grupos e pesquisadores que já possuem esse instrumento estruturado não têm interesse de disponibilizar às demais pesquisas. Logo, fomos levados a uma tentativa de construir o nosso próprio corpus de referência com a colaboração de outros membros do grupo de pesquisa, pois o grupo conta com outras pesquisas que se utilizam do *wordsmith* como ferramenta para auxiliar na análise dos corpora.

Entretanto, chegamos à conclusão que nossos esforços não seriam suficientes uma vez que os estudos de Sardinha (2005) indicam

> [...] que um corpus de referência cujo tamanho seja cinco vezes maior do que o do corpus de estudo permite retirar uma quantidade de palavras-chave estatisticamente equivalente àquela de corpora maiores, o que sugere cinco vezes (maior do que o corpus de estudo) como sendo a ordem de magnitude mínima de um corpus de referência. (p. 184)

Levando em consideração que o nosso corpus de estudos tem um total de 3.496.732 palavras, o corpus de referência que se encaixasse nos pré-requisitos dos estudos de Beber Sardinha teria que ter 17.483.660 palavras. Juntando essa necessidade de baixar arquivos que teriam que ser convertidos para (.txt), além de retirar desses textos as estruturas textuais que seriam as referências, cabeçalhos entre outros, chegamos à conclusão que o período de que não seria possível concluirmos esse corpus de referência até o fim do projeto.

Mas, o trabalho se iniciou e deve ser concluídos nos projetos posteriores que venham a se utilizar da mesma ferramenta, isso caso não seja possível o estabelecimento de parcerias com outros grupos que possuam um corpus de referência estruturado e que possamos contribuir com a atualização desse durante os anos que pudermos utilizar.

Paralelo a nossas tentativas de construir esse corpus de referência, fomos nos enveredando pelos estudos e experimentação da utilização da ferramenta Concord que é um dos utilitários do *Wordsmith tools* 6.

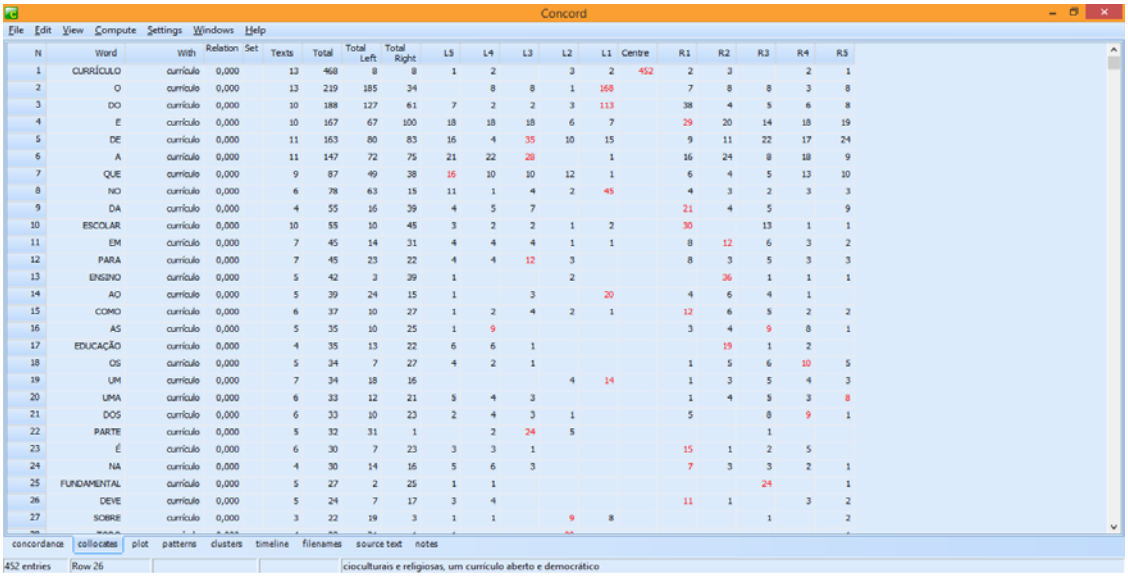
O *Concord* é um programa que produz concordâncias. Essas concordâncias por sua vez são listagens de ocorrências de um item específico indicado pelo pesquisador. Esse item de busca também pode ser conhecido “termo de busca ou nódulo” e pode ser formado por mais de uma palavra. (BEBER SARDINHA, 2009, p.87)

FIGURA 4 - Concordância dos corpora; palavra de busca: Currículo

Fonte: Organizado pelos/as autores/as

A Fig. 4 nos apresenta a ferramenta Concord, apresentando o nódulo “Currículo” que poderia ser chamado de palavra-chave caso tivéssemos conseguido utilizar a ferramenta *Keywords*. O nódulo está com 452 entradas ao centro marcado em azul, como outra alternativa de analisarmos quais palavras estão colocadas ao lado de currículo é através da aba collocates, observemos a Fig. 5.

FIGURA 5 - *Collocates* dos corpora no *Concord*



Fonte: Organizado pelos/as autores/as

Ao clicarmos em collocates, observamos que a palavra currículo está na primeira linha, com ocorrência de 452 entradas, a sua direita (R1) está a palavra “escolar” com um indicativo de 30 em destaque. Isso indica que a palavra “escolar” apareceu ao lado da palavra currículo, 30 vezes, conforme a Fig. 6 e ao clicarmos sobre essa palavra, iremos acessá-la no contexto.

O *wordsmith tools* permite ao usuário configurar predefinições de acordo com seus objetivos, possibilitando fazer ajustes para um melhor desempenho do programa no desenvolvimento da lista. Essa possibilidade de alterar configurações é uma prática optativa, pois o programa vem com suas definições padrões. Assim, esse trabalho se configura como meio de ter um primeiro contato com funcionalidade e as possibilidade de trabalhar com essa ferramenta na pesquisa em Educação.

A coleta de dados

Para iniciarmos o percurso que nos propomos nas exposições anteriores, foi necessário identificar como estavam organizadas as distribuições das escolas municipais de João Pessoa/PB. Após descobrirmos que as escolas eram divididas em nove pólos, tivemos que escolher uma forma de escolhê-las, então optou-se por sorteio, visando garantir uma maior confiabilidade aos dados. Em seguida, nos encaminhamos às escolas munidos da autorização da Secretaria Municipal de Educação, a qual nos dava autorização para solicitar das escolas a retirada de uma cópia do arquivo eletrônico do PPPs das instituições.

Destacamos que é preciso observar que as análises apresentadas sobre as discrepâncias e coincidências na apropriação do significado social e atribuição de sentido pessoal sobre currículo, mediante os dados levantados com base nessa pesquisa, são resultados dos processos formativos vivenciados por esses sujeitos. Dessa forma, as análises ultrapassam esses documentos, porque compreendemos o contexto de formação de opiniões dos indivíduos.

De posse dos Projetos Políticos Pedagógicos, fomos à busca das Políticas nacionais de currículo, as quais já listamos anteriormente. Esses documentos ficam disponíveis online e não apresentaram dificuldades em adquirirmos uma cópia. Entretanto, o arquivo dos corpora deve ser em .txt, ou seja, texto sem formatação e esses documentos não estavam, sendo fáceis de converter o que nos levou a pegar uma cópia direto do *site* do Ministério da Educação e Cultura, o que acarretou o trabalho de retirar do texto as partes que tinham sido atualizadas.

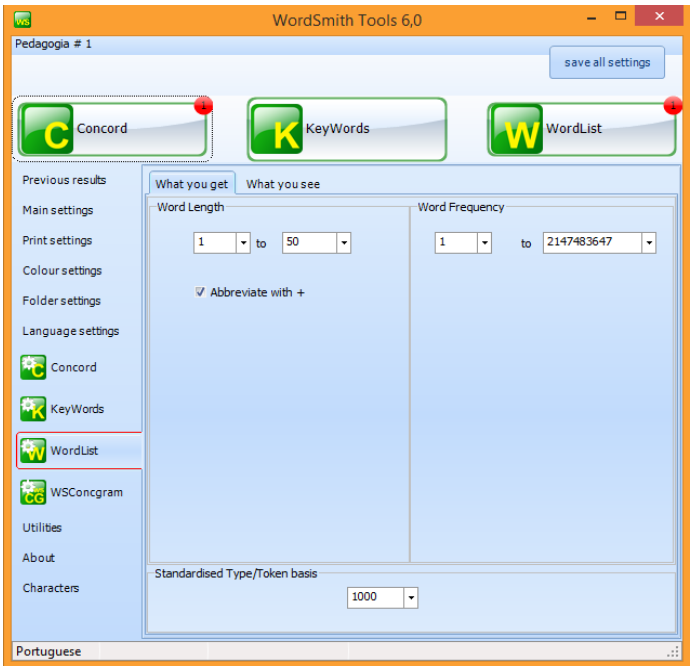
Na intenção de preservar a autenticidade do documento das políticas nacionais, as instituições no momento de as tornarem públicas bloqueiam esses arquivos, impossibilitando, assim, a execução de qualquer ação, como copiar, seja em partes ou na íntegra e fazer qualquer alteração.

O tratamento dos dados

Após a coleta dos dados, iniciamos a utilização do programa *Wordsimth tools 6,* gerando uma *wordlist* de cada documento que compõe os corpora. Em seguida, geramos uma *wordlist* com o conjunto de documentos das políticas curriculares nacionais e locais. Todas as *wordlist* foram salvas separadamente, lembramos que os textos que estamos manipulando no programa todos se encontram no formato de texto sem formatação (txt).

Fizemos algumas predefinições a partir da das configurações dos programas que aparecem na primeira tela do programa como mostra na Fig. 4.

FIGURA 6 - Tela inicial do *Wordsmith tools* 6, aba *What you get*.



Fonte: Organizado pelos/as autores/as

As alterações que fizemos significa que as listas de palavras produzidas frente às definições nos mostrariam as palavras utilizadas no corpus de estudo como uma frequência mínima de um, composta a partir de uma letra, igual aos artigos definidos que possuem apenas uma letra (a, o e etc) até a composta de várias letras. Isso é o que nos mostra a Fig. acima.

Assim, fomos até a *interface* do programa no botão do *wordlist* em *What you get* que indica as opções disponíveis para a composição da *wordlist* e configuramos de forma que selecionasse palavras constituídas a partir de uma só letra até 50, partindo da frequência 1 até o limite do programa.

FIGURA 7 - Tela inicial do *Wordsmith tools* 6, aba *What you see*.

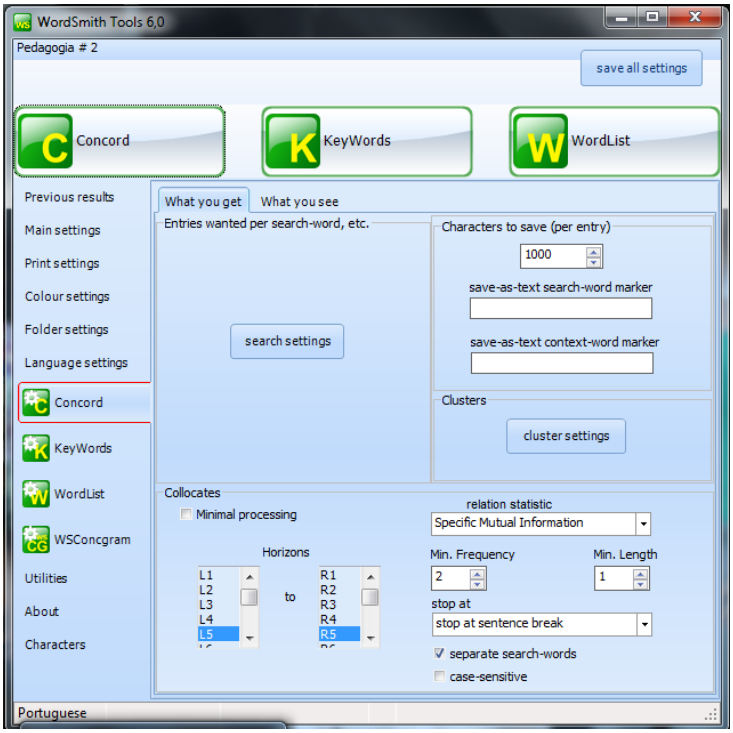
Fonte: Organizado pelos/as autores/as.

Na aba *What you see*, optamos por continuar com a configuração padrão do programa, ou seja, visualizar apenas palavras com colunas referentes à frequência, a porcentagem e seus respectivos textos de origens.

Na configuração da ferramenta Concord, optamos por definir a lista de *collocates* também conhecidos como colocados, em uma determinada posição com frequência mínima de uma vez. Pois essa configuração nos proporcionaria uma melhor visualização do posicionamento das diversas palavras que se apresentavam ao redor do nódulo, podendo perceber a elaboração dos clusters por parte dos documentos das políticas nacionais e locais que tratam do currículo e que compõem o nosso corpus de estudo. Veja isso na Fig. abaixo.

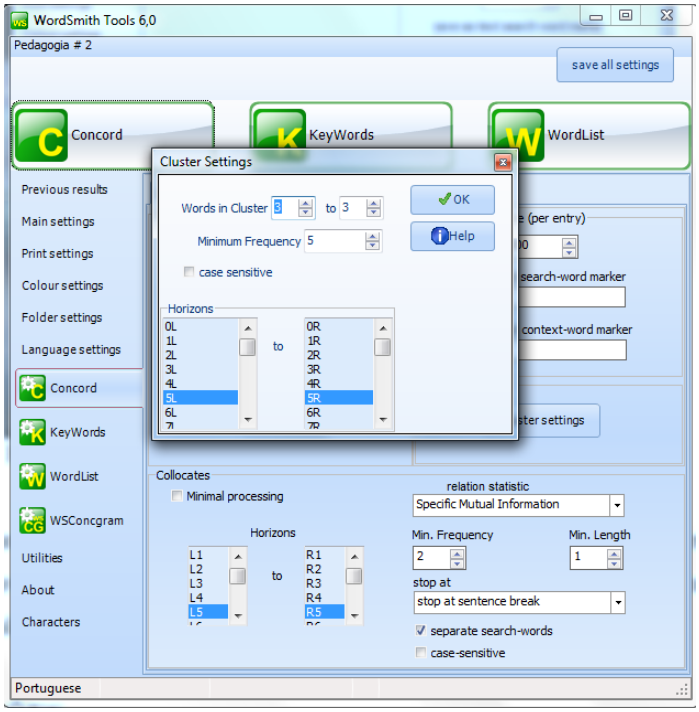
Ao acionarmos o botão *cluster settings*, é exposta outra tela onde poderemos escolher a quantidade de palavras que irá compor o *clusters*, como também a frequência mínima das palavras e abrangência das palavras ao redor do nódulo.

FIGURA 8 - Tela de configuração da ferramenta *Concord*



Fonte: Organizado pelos/as autores/as

FIGURA 9 - Tela de configuração da ferramenta Concord, aba do *cluster setting*



Fonte: Organizado pelos/as autores/as

Na continuidade dos nossos estudos e a partir da nossa *wordlist* extraída dos documentos nacionais e locais que norteiam a educação básica, localizamos a palavra currículo com uma frequência de 92,86% presente nos 13 textos que compõem os corpora. A partir dessa palavra, acionamos o *Concord* para a elaboração de nossas tabelas.

TABELA 2 - O nódulo currículo e as variações com os dados do *wordlist*

Palavras	Frequência	Frequência (%)	Nº de textos	Frequência textos (%)
Curricular	321	0,07	13	92,86
Curriculares	697	0,16	13	92,86
Currículo	452	0,10	13	92,86
Currículos	126	0,03	7	50,00

Fonte: Organizado pelos/as autores/as

Para compor nossas tabelas, foi necessário percorrer os *collocates* do nódulo currículo, buscando as palavras que identificam o currículo construindo a tabela com esses nomes e suas frequências no corpus de estudo.

TABELA 3 - Lista de concordância, com informações dos *collocates*

Palavras	R1	R2	R3	R4	R5
Aberto	3	1	1	-	-
Ação	-	3	1	3	1
Aprendizagem	-	-	1	1	2
Componentes	-	-	2	-	3
Conhecimentos	4	2	1	2	1
Conjunto	-	4	1	-	1
Conteúdos	-	1	1	1	1
Contextualização	-	-	4	2	-
Cultura	-	-	4	2	-
Democrático	-	-	2	-	1
Desigualdades	-	-	1	3	-
Escolar	30	-	13	1	1
Estudo	-	1	-	2	-
Experiências	-	1	1	1	-
Flexível	1	-	2	-	-
Formal	4	-	1	-	-
Integrado	5	-	-	-	-
Interdisciplinar	-	-	2	-	2
Obrigatório	1	-	-	-	1
Oculto	3	-	-	-	-
Orienta	1	-	1	-	1
Político-Pedagógico	-	-	-	2	1
Práticas	-	-	1	-	2
Processo	-	1	-	3	-
Propostas	-	-	-	2	1
Trabalho	1	4	1	1	-
Valores	-	-	-	2	2

Fonte: Organizado pelos/as autores/as

A tabela 3 nos mostra as palavras que aparecem no corpus de estudo que tem a função de caracterizar o nódulo currículo, ou seja, numa visão gramatical, são os adjetivos do substantivo "currículo". Entretanto, vale ressaltar que, por fatores que ainda não sabemos, algumas outras palavras não ficaram dentro dos limites estipulados por nós no lado direito do nódulo. O que queremos elencar é que as palavras arena, coração, criação, critérios, diálogo, dispositivo, diversificado, esforços, expressões, extrapolar, fronteiras, guia, histórico, homogeneizado, identificação, inclui, instrumento, integrador, interação, interlocução, metas, metodológico, modos, polissêmico, possibilidade, produção, proposta, redimensionado, relações, seleção, transgressor, transversal, visão e vivo não estão colocadas quantitativamente, mas estão sendo consideradas.

TABELA 4 - Amostragem das palavras e a quantidades de texto da ocorrência

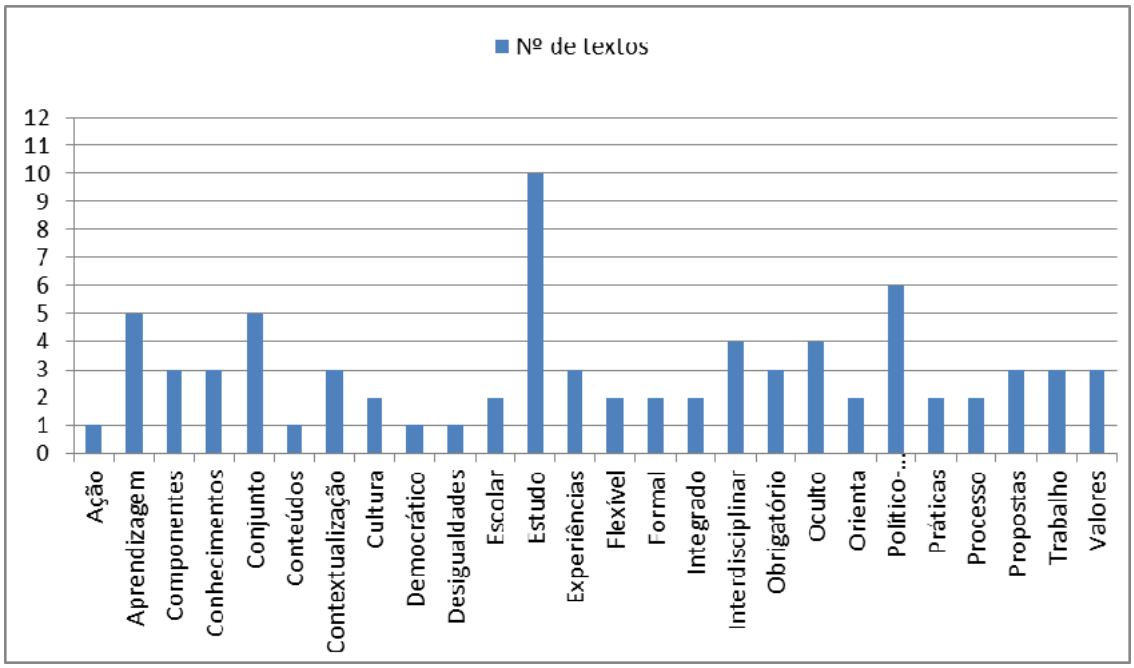
Palavras	Nº de Textos
Aberto	1
Ação	5
Aprendizagem	3
Componentes	3
Conhecimentos	5
Conjunto	1
Conteúdos	3
Contextualização	2
Cultura	1
Democrático	1
Desigualdades	2
Escolar	10
Estudo	3
Experiências	2
Flexível	2
Formal	2
Integrado	4
Interdisciplinar	3
Obrigatório	4
Oculto	2
Orienta	6

Político-Pedagógico	2
Práticas	2
Processo	3
Propostas	3
Trabalho	3
Valores	2

Fonte: Organziado pelos/as autores/as

Partindo da tabela acima, resolvemos expor um gráfico para melhor visualização dos termos com maior ocorrência entre os sujeitos pesquisados por nós.

GRÁFICO 1 - Amostragem das palavras e a quantidades de texto da ocorrência



Fonte: Organizado pelos/as autores/as

CONCLUSÕES

O trabalho realizado nos permitiu ter um conhecimento prático das pesquisas que utilizam a linguística como metodologia. Em nosso caso, se restringiu a linguística de corpus, como também explorar a potencialidade da ferramenta *Wordsmith tools* 6 no processo de exploração dos documentos que compunham os corpora, detalhando o passo a passo da nossa experimentação e experiência que fomos adquirindo ao longo da pesquisa, além da experiência sobre as etapas de uma pesquisa e dos procedimentos afins, respeitando os objetivos de forma a não alterá-los mesmo que sejam necessárias algumas modificações no processo metodológico.

No que ser refere à construção do corpus, destacamos que foram dados passos importantes quando nos referimos a utilizar a linguística de corpus, contemplando parte da nossa metodologia e o *Wordsmith tools* como ferramenta para auxiliar na análise dos

documentos. Os nossos contatos com a ferramenta aqui elencada nos mostrou o quanto ainda podemos aprender e colaborar com estudos que façam uso de uma metodologia próxima a que utilizamos.

O Grupo de Estudos e Pesquisas em Políticas Curriculares (GEPPC) possui pesquisas que fazem uso da ferramenta *wordsmith* e por consequência da Linguística de Corpus, por esse motivo, já foi oferecido um minicurso com o Profº Mr Felipe Aguiar, Professor Substituto do Instituto de Letras da Universidade Estadual do Rio de Janeiro (UERJ). Esse momento se caracterizou impar de aprendizado, reforçando o quão é importante para as pesquisas em educação considerar o discurso e os documentos que são construídos como representação dos processos discursivos dos sujeitos em sociedade.

Entretanto, para chegarmos a esse grau de confiabilidade, foi necessário um aprofundamento teórico que nos levou a conhecer mais profundamente a ferramenta e o que precisaríamos para poder utilizá-la de forma a potencializar os resultados da nossa pesquisa ampliando o leque de representatividade qualitativo e quantitativo dos resultados.

Todavia, o que nós não conseguimos alcançar e que era extremamente importante para uma melhor organização dos nossos dados foi à extração das palavras chave através do *Keywords*. A nossa dificuldade se deve a necessidade de um corpus de referência adequado ao nosso objeto de pesquisa ou a grande área a qual esse está inserido. Existem pesquisas na nossa área que possuem esse corpus de referência, mas não existe uma política de socialização para a uma construção coletiva do conhecimento e um aperfeiçoamento dos recursos. O corpus de referencia que conseguimos foi fornecido pelo professor que ministrou o minicurso, como falado anteriormente não contemplava a nossa pesquisa.

Frente a essas demandas que se caracterizaram como dificuldade e por sua vez nos fizerem ir a fundo buscando as possibilidades para sanar essas problemáticas, resolvemos junto, aos/as outros/as bolsistas PIBIC 2014/2015 e PROLICEN 2014/2015 que vão também se utilizar da mesma ferramenta, construir um corpus de referência que nos ajude nas nossas pesquisas que estão sendo desenvolvidas, permitindo assim utilizar o *Wordsimth* como um todo.

O GEPPC é um dos pioneiros na Universidade Federal da Paraíba (UFPB) a fazer esse vínculo da Linguística de Corpus e o *Wordsmith* com as pesquisas em Educação voltadas ao objeto Currículo Escolar. Essa experiência primeira vem acompanhada com desafios que vão se constituindo como momentos de aprendizado que precisam ser superados e marcados com resultados que mostrem alternativas de trabalho, acompanhando as novas demandas tecnológicas que surgem.

Podemos afirmar que as nossas descobertas com essa metodologia e instrumento devem ser compartilhadas por meios de artigos e oficinas. Assim, além de vir publicando os nossos resultados é uma realidade a construção de uma oficina teórica e prática de como utilizar o *Wordsmith*, a ser oferecida pelo grupo de estudos e ministrada pelos bolsistas deste projeto, para os interessados por aprenderem novas técnicas de coleta e tratamento dos dados de uma pesquisa em Educação.

Logo, vemos que a pesquisa foi além de suas propostas. Mais do que analisar os sentidos de currículo, deu margem a construção de outros processos de formação que podem fomentar investigações debruçadas em uma área considerada carente de estudos. Essa carência que foi elencada é vista na utilização da Linguística de Corpus e da ferramenta *Wordsimth Tools 6* nas pesquisas em Educação.

REFERÊNCIAS

BERBER SARDINHA, Tony**. Linguística de Corpus. Barueri,** São Paulo: Manole, 2004.

_____. **Como encontrar as palavras-chave mais importantes de umcorpus com Wordsmith Tools**. D.E.L.T.A. São Paulo, v. 21, n. 2, dez. 2005. Disponível em:<http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502005000200004&lng=en&nrm=iso>. Acesso em: 13 abr. 2011.

_____. **Linguística de Corpus: histórico e problemática**. **D.E.L.T.A.** São Paulo, v. 16, n. 2, p.323-367, 2000. Disponível em:<http://www.scielo.br/pdf/delta/v16n2

/a05v16n2.pdf>.Acesso em: 13 agosto. 2013.

_____. **O banco de palavras-chave como instrumento de identificação de Palavras-chaveexclusivas no programa Wordsmith Tools keyword**. TheSPecialist, São Paulo, v. 27, n. 1,p. 1-19, 2006. Disponível em:<http://revistas.pucsp.br/index

.php/esp/article/download/1626/1045>. Acesso em: 3setembro 2013.

_____. Pesquisa em linguística de corpus com WordSmith Tools. Campinas, SP: Mercad de letras, 2009.

_____. Usando Wordsmith Tools na investigação da linguagem. DIRECT Papers 40, SãoPaulo, 1999. Disponível em: <http://www2.lael.pucsp.br/direct/DirectPapers40.pdf>. Acesso em: 12 jul. 2011

RIBEIRO, G. C. B**. Tradução técnica, terminologia e lingüística de corpus: a ferramenta Wordsmith Tools**. Cadernos de Tradução, Vol. 2, No 14, p. 159-174, 2004.

SILVA, Tomaz Tadeu da**. Currículo e Identidade Social: territórios contestados.** In: Alienígenas na Sala de Aula: uma introdução aos estudos culturais em educação. Petrópolis: Vozes, 1995.

_____. Documentos de Identidade: uma introdução às teorias do currículo. 2ª ed., 11ª

reimpressão. Belo Horizonte: Autêntica, 2007.

PUBLICAÇÕES

HONORATO, R. F. S.; BARBOSA, S. W. X. Os Sentidos do Currículo nas Escolas da Rede Municipal de Ensino de João Pessoa/PB: wordsmith como ferramenta de exploração da corpora. In: VI Colóquio Internacional de Políticas e Práticas Curriculares, 2013, João Pessoa. CURRÍCULO: (Re)Contruindo os Sentidos de Educação e Ensino, 2013. p. 649-655.