



PennState
College of the
Liberal Arts

C-SODA
Center for Social Data Analytics

Day 8 - Multilingual Text as Data & Machine Translation

Advanced Text as Data: Natural Language Processing
Essex Summer School in Social Science Data Analysis

Burt L. Monroe (Instructor) & Sam Bestvater (TA)
Pennsylvania State University

August 4, 2021

Today

- In the first half, I will essentially deliver my talk (given as recently as three weeks ago) on the Goist & Monroe multilingual text analysis paper, but with extended comments about connections to related ideas and techniques we have discussed in class.
- In the second half, we will give machine translation a fairer shake than I do in the paper:
 - I will discuss contexts in which machine translation is useful and viable in social scientific text as data research.
 - Sam will go over a notebook on implementing machine translation with pretrained models / transfer learning (avoiding the costs associated with something like Google Translate).

Tomorrow

- Semantic change (text analysis across time), including “diachronic word embeddings” (two suggested readings - one “classic”, one SOTA)
- Fairness & bias in NLP (including Papakyriakopoulos, et al. 2020. on bias in word embeddings).
- The extent to which NLP models have or have not reached a level of “natural language understanding.” (including Bender and Koller 2020).
- (If I can, I’ll squeeze in something on the requested task of “custom NER.”)

“Scaling the Tower of Babel: Common-scale analysis of Political Text
in Multiple Languages.” (Goist & Monroe)

Just so I don't bury the lede ...

The meta-question I'll address today is:

Can text-as-data techniques be adapted or developed to enable credible quantitative cross-cultural social science research using untranslated primary textual sources that are written in multiple languages we don't necessarily read or speak?

Now, some more specific motivation by way of a controversial claim ...

Politicians talk a lot.

Politicians talk a lot

Consider the United States Congress.

Most days of the year, Representatives and Senators take to the floor of Congress and deliver speeches.

Stenographers write it all down and it is entered in the *Congressional Record*.

September 17, 1980

THE WAY OF OUR FATHERS?

HON. NEWT GINGRICH
OF GEORGIA
IN THE HOUSE OF REPRESENTATIVES

Wednesday, September 17, 1980

● Mr. GINGRICH. Mr. Speaker, the term "enterprise zones" is quickly becoming familiar across the United States, mainly due to a bill introduced by New York Congressmen JACK KEMP and ROBERT GARCIA. The bill is H.R. 7563, the Urban Jobs and Enterprise Zone Act.

Enterprise zones is not a new concept. The British and Irish Governments are experimenting with enterprise zones as a means to stop urban decay. And the idea has been introduced in Congress before, but was laid aside for the conventional methods of urban development.

Recently, an article on enterprise zones appeared in the Atlanta Constitution by Gene Tharpe, a noted economics writer. I found his points about "enterprise zones" being similar to the old-fashioned term "free enterprise" very interesting. I want to share his

Politicians talk a lot

Actually, they want to say so much, or be able to claim they said so much, that they don't even actually have to say it.

Gingrich's speech appears in a section called "Extensions of Remarks." It was "read into the Record."

That page has about 1800 words.

September 17, 1980

EXTENSIONS OF REMARKS

THE WAY OF OUR FATHERS?

HON. NEWT GINGRICH
OF GEORGIA
IN THE HOUSE OF REPRESENTATIVES
Wednesday, September 17, 1980

25922

Rita Sands, ABC News, Hauppauge, New York.
KOPPEL We'll have the latest on Iran in a moment.

* * * * *

KOPPEL It's day 313 of the crisis in Iran and Iranian President Bani Sadr says it is United States policy that has been delaying a solution to the hostage issue. Bani Sadr did not elaborate, but it's believed he was referring to Iranian demands that the U.S. apologize for its past actions in Iran.

That's our report for tonight. This is Ted Koppel of Washington. For all of us here at ABC News, good night.

ANNOUNCER This has been Nightline, a presentation of ABC News.

PATRIOTISM ALIVE AND WELL

HON. CARLOS J. MOORHEAD
OF CALIFORNIA
IN THE HOUSE OF REPRESENTATIVES
Wednesday, September 17, 1980

Mr. MOORHEAD of California. Mr. Speaker, Mr. Aarna Davidson of Sunland, Calif., was recently guest editorialist for the Record Ledger, one of the leading newspapers in California's 22d District. Mr. Davidson, a naturalized citizen of our country, is a well-known writer and has been a recipient of the Freedoms Foundation Award on four different occasions. His love for his adopted country is often and openly expressed in his writings. I believe my colleagues will appreciate Mr. Davidson's view that patriotism is not dead in America.

The editorial follows:

PATRIOTISM ALIVE AND WELL
(By Aarno Davidson)

Contrary to some people's conception of patriotism, like God, is not dead. Not in the U.S.A. True, some vociferous soap-box orators of pessimism are dishing out the gospel of Armageddon through the three D's: Damn, Doom, Discord. In this respect they resemble the front page headlines which accentuate the worst traumatic occurrences. Thus, the adage, "no news is good news" has been changed to "good news is no news."

To counteract this fashionable trend of thought, numerous organizations are providing to us and to the outside world that it is fallacious to judge this great country by front page headlines by, for example, bad movies in movie houses, radio and T.V. ("Tantrum Violence"). Contrastingly, our guardians of positivism, peace and freedom continue to herald the wholesome side of American life. Freedoms Foundation at Valley Forge, Pa. leads the way. During the past three decades it has made enormous strides for the enhancement of freedom in general and the preservation of the American heritage in particular. To elucidate:

Every year it inaugurates a National Awards Program, stimulating people from all walks of life to participate in a hundred different contests on vital subjects: education, religion, science, journalism, literature, radio, television, film, corporations, non-profit organizations, etc. These contests are judged by a panel of Supreme Court Judges, top officers of national and patriotic organizations. These judges select almost a thousand winners whose names are inscribed permanently in the Foundation's Annual National Awards Book. Sixty pages of information about and pictures of the truly great, the near-great and hundreds of everyday thinking people devoted to freedom's cause.

Then, too, almost 5,000 high school juniors have participated in seminars on the U.S.A. and their role in the responsibilities for its future. Similarly, about 5,000 high school teachers from every state have taken in graduate credit seminars preparing themselves to better teach American values, including the American private enterprise system.

The national headquarters of this non-profit, non-sectarian organization is Valley Forge, Pennsylvania. No other historical landmark offers more surety symbols of American Heritage than this town where self-sacrifice, discipline and devotion keep uninterrupted the torch of freedom and patriotism alive. Here one may awe at "Faith of Our Father's Chapel," an imposing edifice representing the major religious faiths in the United States. "Independence demands," a collection of stones and bricks from the birthplaces of the signers of the Declaration of Independence; "Spirit of '76 Library Building" with innumerable original early American manuscripts. Many more kindred memorabilia of America's glorious past are here enshrined on public display.

It is stated that I hail from the cosmopolitan metropolis of Amsterdam, Holland, where chauvinism is practically non-existent. Furthermore, I've traveled extensively through Europe, South America, certain parts of the Orient, Canada and Mexico. With deep conviction I feel justified to emulate those who say that in spite of all its shortcomings this is still the sunniest spot on our shadowy planet. Freedoms Foundation is doing its utmost to keep it so.

Author Ben J. Wattenberg, "Only a nation that knows what it has accomplished in the past can know what it is able to do in the future. We are ill-served by the 'nothing works' boys. Americans came through a turbulent value-busting era and they coped well. The real heroes of the past dozen years have been American people. Wiser, tougher, calmer, shrewder and more forgiving than the critics to Right and to the Left." In his fascinating book of poems, Western Star, Stephen Vincent Benét wrote pensively:

Americans are always on the move
And when the whistle blows they go away.
Sometimes there never was a whistle blown
But they don't care, they blow their own.
A dozen tunes but only one refrain:
We know where we are going and are on our way.

I was privileged to acquire that whistle the day an amiable Federal Judge handed me my American Naturalization Certificate. This certificate I consider, together with my four Freedoms Foundation Awards, the most treasured possession of all my life. Since acquiring this whistle I've often put it to good use. Sometimes it sounds a bit out of tune but this does not stop me from whistling. . . .

So long as there are American organizations like Freedoms Foundation I shall keep on whistling the haunting tunes of "America The Beautiful" or Irving Berlin's popular song, "God Bless America, my home sweet home." A home I adopted decades ago and have been cherishing it ever since.

Just as the United States, the Enterprise Zones idea began catching on last year, and some politicians soon were touting the new-found old-time religion.

This summer in Congress, a conservative Republican and liberal Democrat teamed up to introduce the Urban Jobs and Enterprise Zone Act.

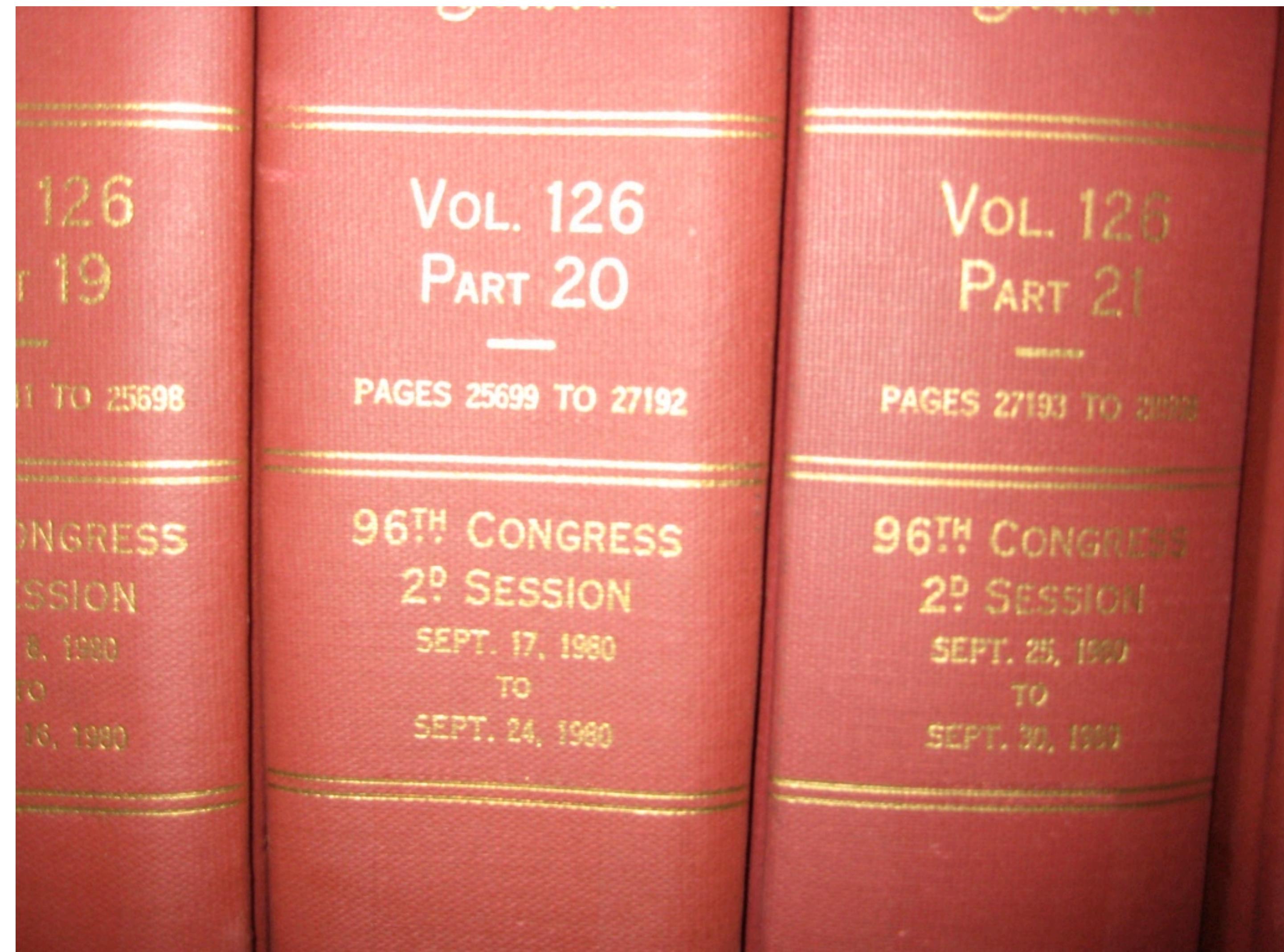
Politicians talk a lot

The volume that page is in
has just under 1500 pages.

It covers a week.

A year of the *Record*,
covering House and Senate,
has about 50,000,000 words.

You would have to read three
War-and-Peace-lengths a
day, every day, to keep pace.



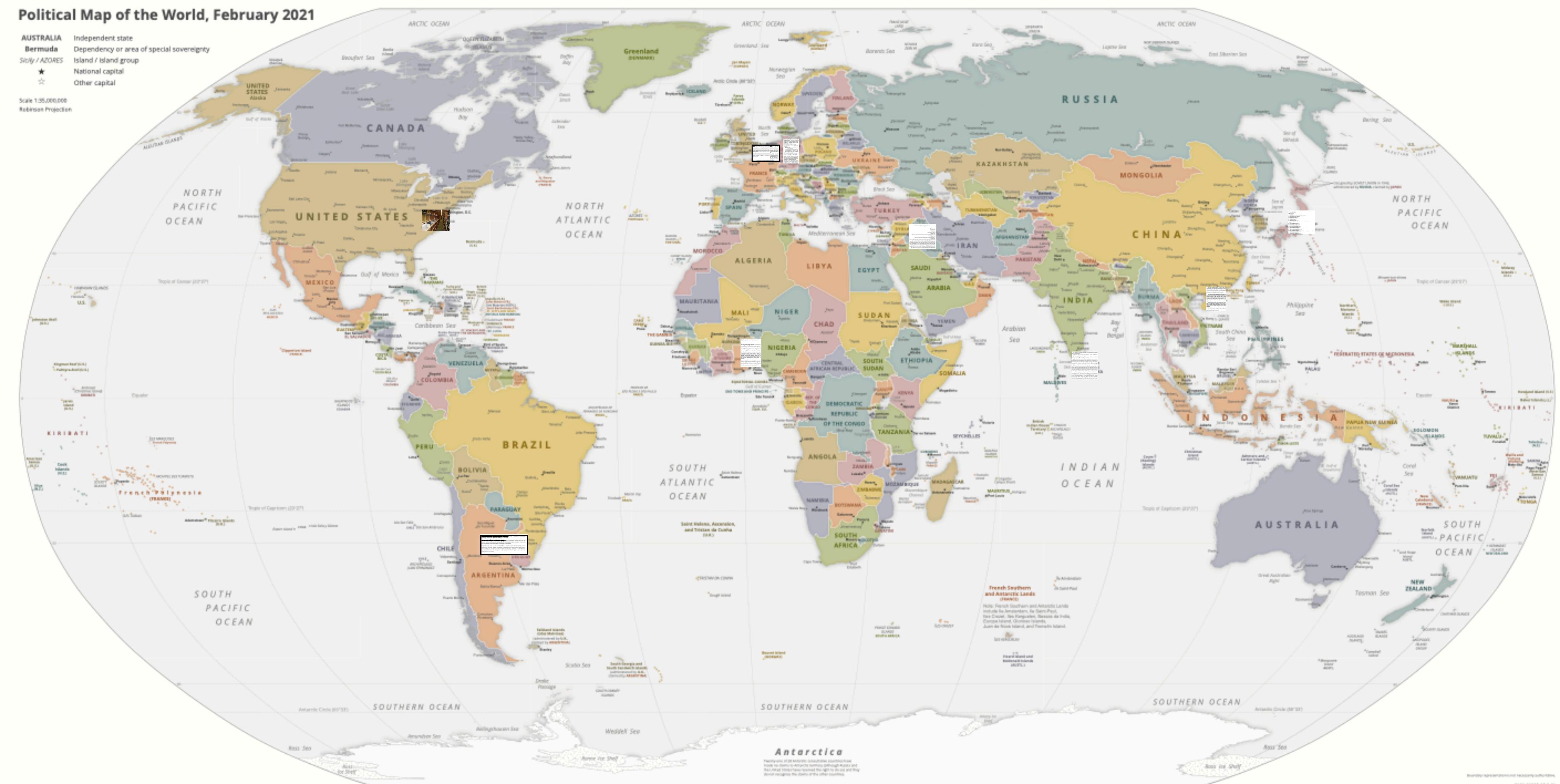
Politicians talk a lot

You'd have to go faster than that if you wanted to make a dent in the 200+ years — roughly 3.5 million pages — of *Congressional Record* and precursors that already exist.

And that's just the United States Congress.



Politicians talk a lot - everywhere



Politicians talk a lot - everywhere

Uruguay

Political Map of the World, February 2021

AUSTRALIA
Bermuda
Sicily / AZORES

Independent state
Dependency or area of special sovereignty
Island / Island group



El señor PRESIDENTE (Marcial Ayaipoma Alvarado).— Tiene la palabra el congresista Diez Canseco Cisneros, por tres minutos.

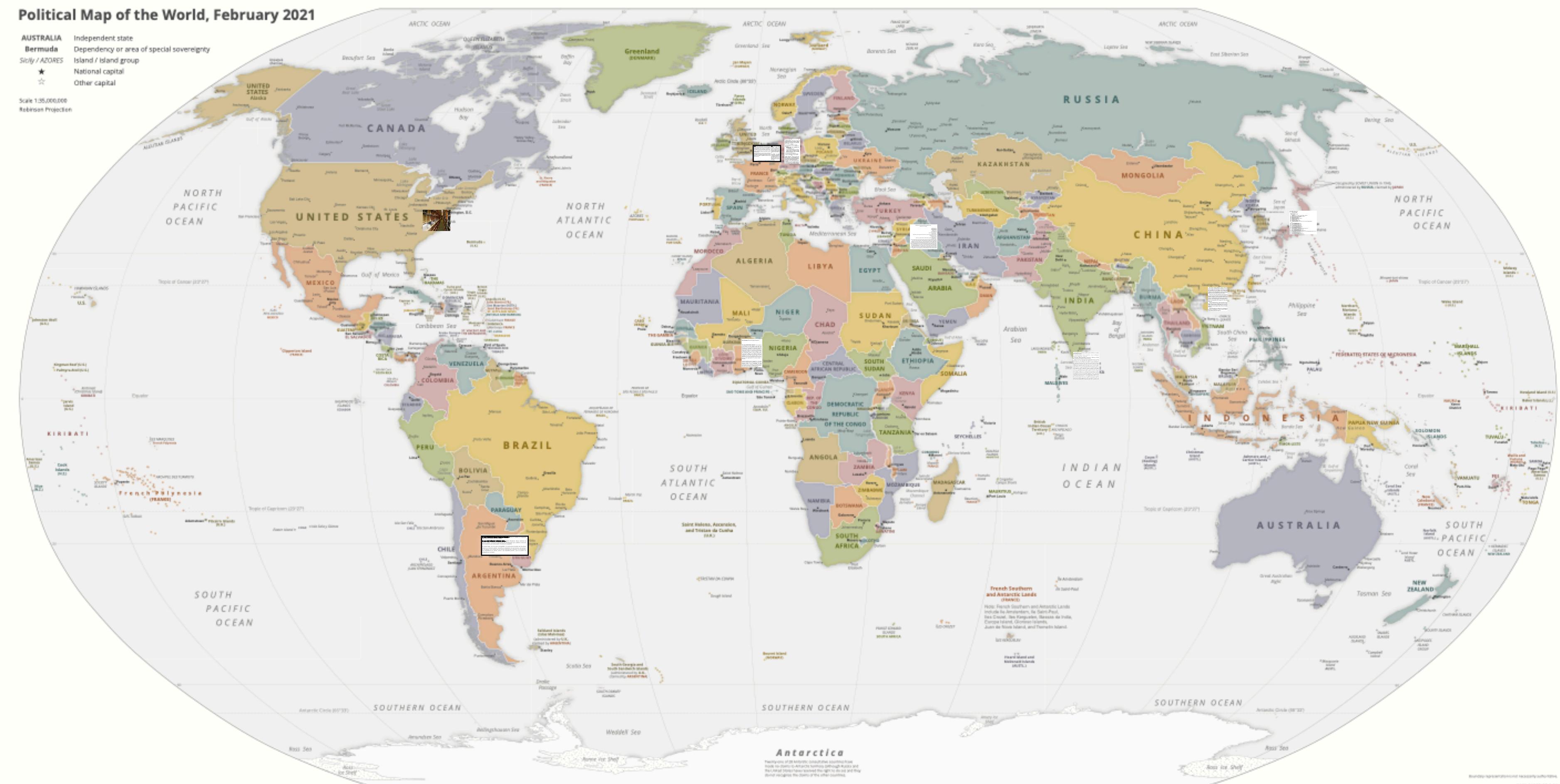
El señor DIEZ CANSECO CISNEROS (UPD).— Señor Presidente: Quiero reafirmar de manera enfática mi profunda discrepancia del planteamiento para modificar el tema de las concesiones y que estas puedan convertirse en propiedad.

En primer lugar, creo que eso es inaceptable, y no solo porque aquellos que hoy tienen terrenos en concesión podrían convertirse en propietarios de esos terrenos, sino porque, en mi opinión, es un crimen que los recursos naturales que, conforme a la Constitución, son propiedad de la Nación, pasen a ser propiedad privada y que no puedan ser entregados en concesión.



2020-08-28-28-21

Politicians talk a lot - everywhere



Politicians talk a lot - everywhere

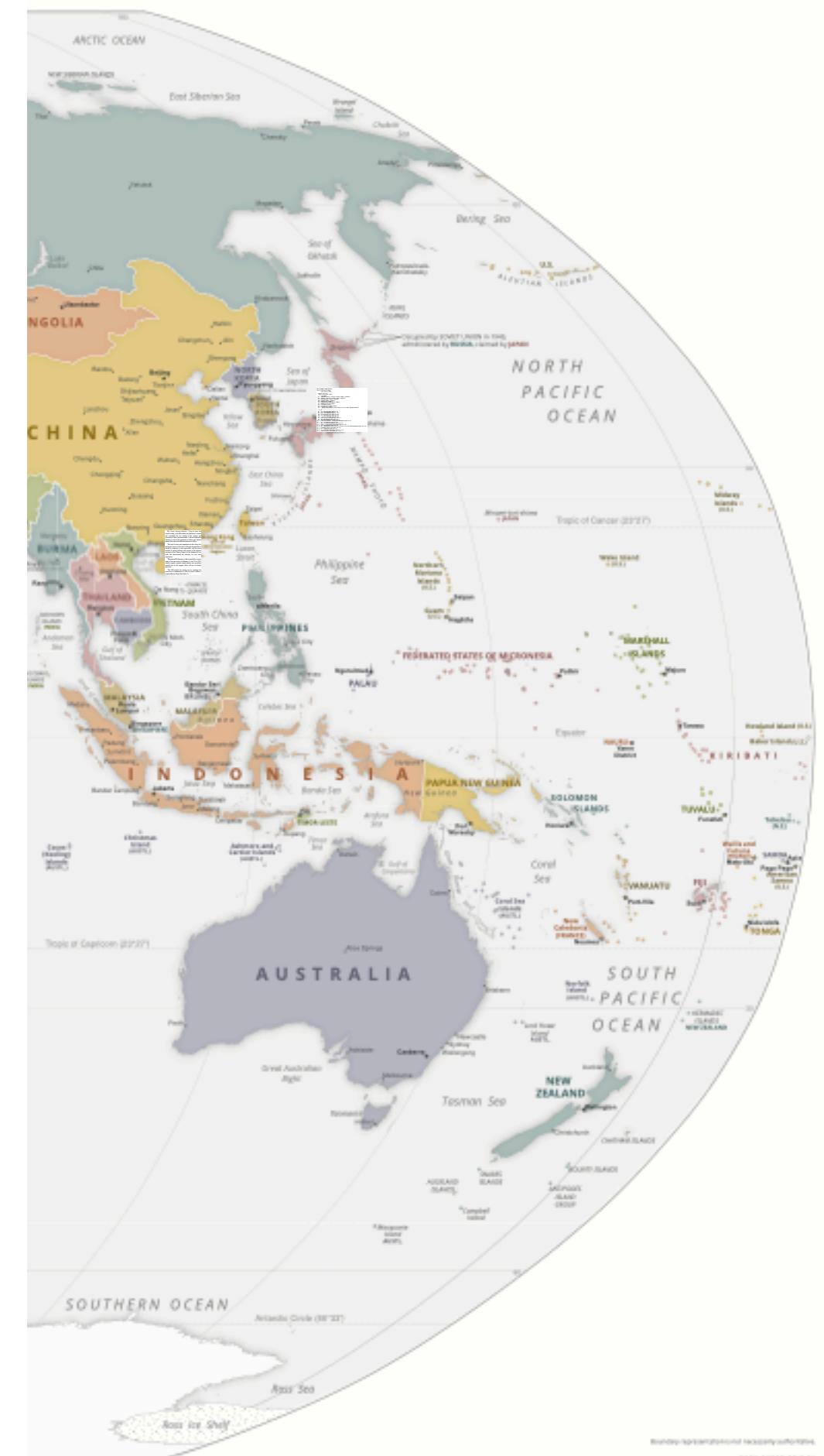
Benin



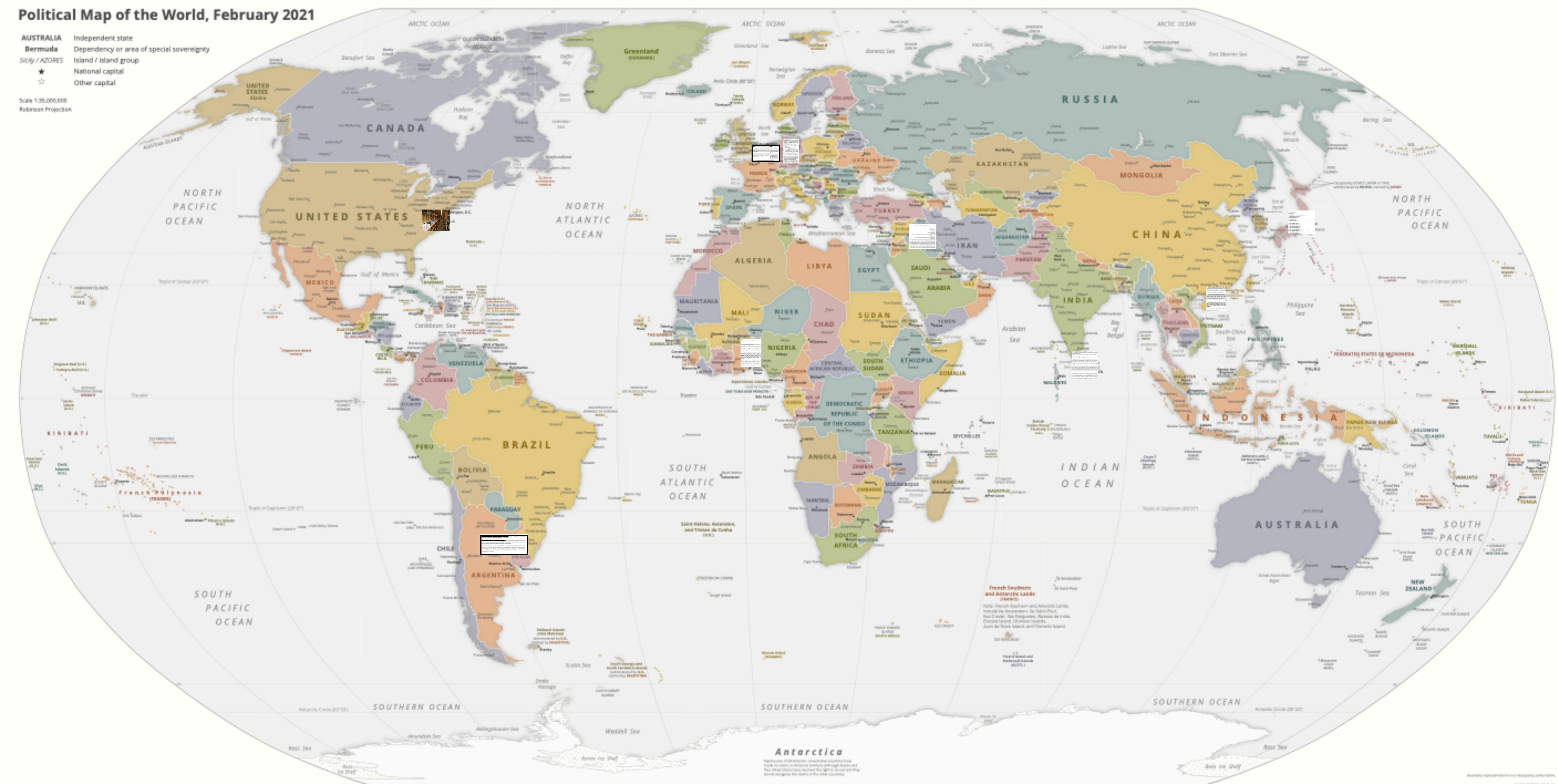
M. le Président. Honorable député Okounlola !

M. André OKOUNLOLA-BIAOU. Je me réjouis et je me félicite de telle initiative pour que cette proposition de loi soit aujourd'hui en étude. Parce que le passage de l'analogique au numérique dépasse le cadre du Bénin seul. Tout le monde sait que dans le monde entier, en 2015 on veut passer de l'analogique au numérique. Même si vous êtes à Gamian, si vous êtes à Bembèrèkè, même si vous êtes à Bopa, même si vous êtes à Diho, on est obligé de passer de l'analogique au numérique. S'il doit avoir un arsenal juridique qui doit nous permettre d'aller vers cet objectif, il n'y a pas de raison que nous ne nous jetons pas sur cette opportunité-là de pouvoir régler ce problème.

L'arsenal juridique qui est proposé ici, c'est pour régler un certain nombre de problèmes qui vont nous permettre d'aller vers le numérique. Et nous connaissons bien les avantages du numérique. Quand bien même que nous ne sommes pas du domaine, mais nous savons l'importance du numérique.



Politicians talk a lot - everywhere



Politicians talk a lot - everywhere

Belgium (in two languages)

Political Map of the World, February 2021

29.01 Pierrette Cahay-André (MR): Monsieur le président, vous savez que c'est ma dernière séance. Je ne ferai pas le même discours que celui de M. Henry, que j'ai par ailleurs apprécié. Je voudrais simplement parler au nom de la commission des Naturalisations, dont je fus la vice-présidente et remercier M. Hove, notre président, qui a très bien mené les débats. (*Applaudissements*)

De plus, je souhaiterais remercier mes deux collègues de la "petite Chambre", dont j'étais la présidente: André Frédéric et Stijn Bex. (*Applaudissements*)

Monsieur le président, je tiens à vous dire que nous avons travaillé en bonne intelligence. Nous avons octroyé la nationalité à bon nombre de candidats. Il me faut également rappeler que, pour bon nombre de nos concitoyens, cet accès à la nationalité constitue un accès à l'emploi, à la stabilité, à l'intégration et au droit de vote. (*Applaudissements*)

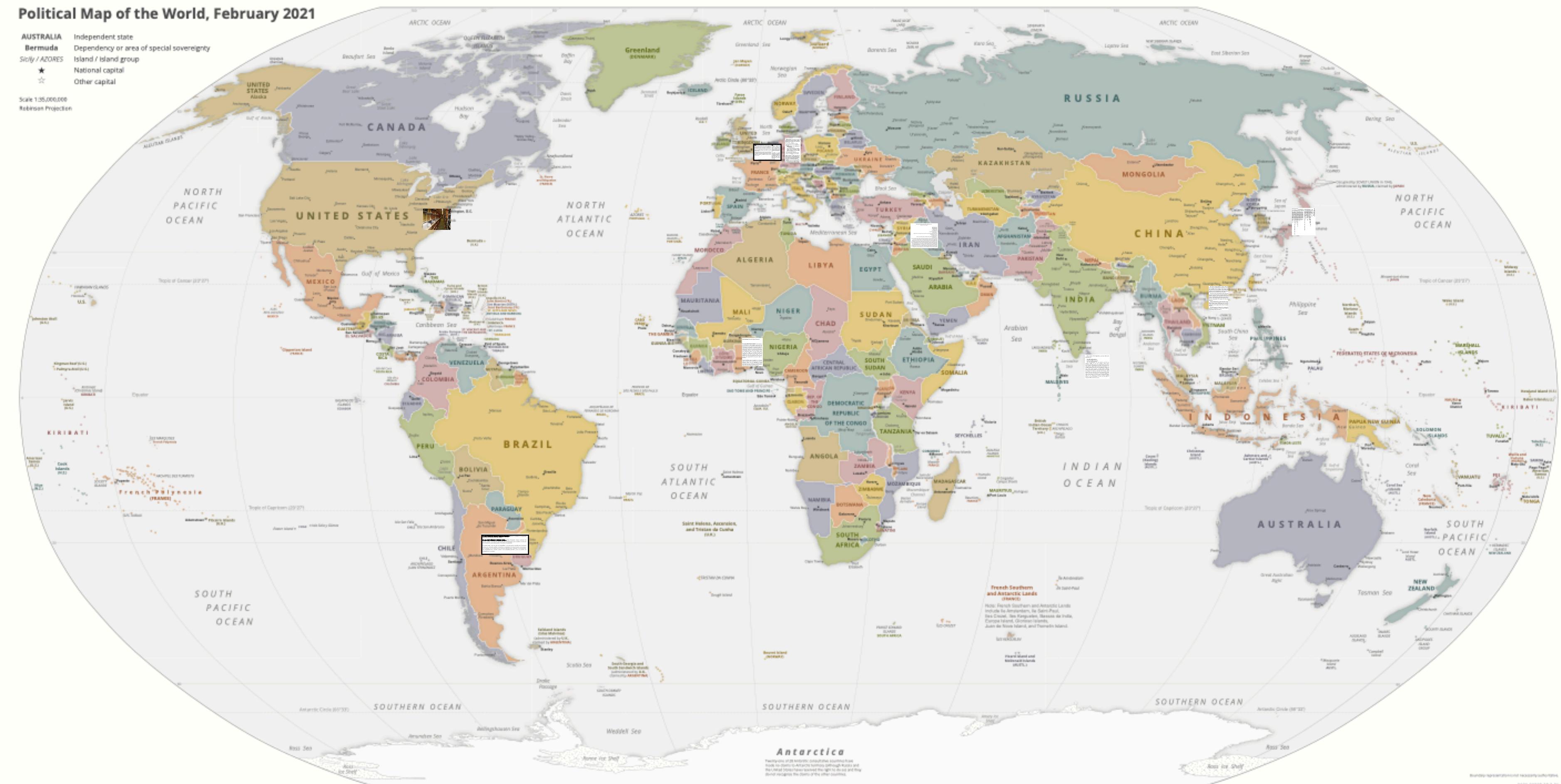
29.01 Pierrette Cahay-André (MR): Dit is mijn laatste vergadering. Namens de commissie voor de Naturalisaties, waarvan ik ondervoorzitter was, dank ik de voorzitter, de heer Hove, die de debatten zeer goed heeft geleid. Ik dank ook mijn twee collega's van de 'kleine kamer' die ik voorzag, de heren Frédéric en Bex.

Wij hebben in goede verstandhouding samengewerkt en hebben aan tal van gegadigden de nationaliteit toegekend. Voor velen betekent het verwerven van de nationaliteit ook dat zij een baan zullen vinden, dat zij stabiliteit zullen vinden, zich zullen kunnen integreren en stemrecht zullen krijgen.

Le président: Ce n'est pas terminé, chers collègues! L'atmosphère n'est pas mauvaise!

Facilities of the Antarctic Treaty member states have been located on the map. The map does not show the exact location of all facilities. The United States have claimed the right to do so and they do not recognize the claims of the other countries.

Politicians talk a lot - everywhere



Politicians talk a lot - everywhere

Sri Lanka



Hon. Leader of the Opposition, you may make your statement now.

ගරු සංචීත් ප්‍රේමදාස මහතා
(මාණ්පුමික සංඝිත් පිරෝමතාස)
(The Hon. Sajith Premadasa)

ගරු කරානායකතුමති, මට මේ අවස්ථාව ලබා දීම ගැන ඔබතුමාට ස්තුතියි. මා මේ ගැන ඔබතුමාට දැනුම් දුන්නා. දියවත්තා ඔයේ පරිසර දූෂණය ගැන මාර්තු 24වැනි දා මම ප්‍රකාශ කළා. ඔබතුමාට පාර්ලිමේන්තුවේ බලධාරීන් උපදෙස් දුන්නා, එය පොකුණක් සුද්ධ කිරීම හේතුවෙන් සිදු වුණු දෙයක් කියලා. අද ද්‍රව්‍යෙන් එක අසත්‍යයක් බවට පත් වෙලා තිබෙනවා. ඔබතුමා මේ ගැන හොඳට සොයා බලන්න. මා හිතන විධියට, drainage systems, sewerage systems ගණනාවක් දියවත්තා ඔයට යොමු කර තිබෙනවා. එම නිසා අද දියවත්තා ඔයේ බරපතල පරිසර භානියක් සිදු වෙලා තිබෙනවා. මා මේ ගැන එදා සඳහන් කරන කොට අද ඔය කැ ගහන ඇත්තේ මට දොස් කිවිවා, ලකුණු දමා ගන්න හදනවා කියලා. ඒ වාගේම, අසහා වචනත් පාවිච්ච කළා. නමුත් අද මම ඔබතුමාගෙන් බොහෝම කාරුණික ඉල්ලීමක් කරනවා, ගරු කරානායකතුමති. එදා සිට අද ද්‍රව්‍ය දක්වා ඒ වෙනුවෙන් කිසීම දෙයක් කරලා තැහැ.



It is easy to look like a fool using Google Translate

Hon. Leader of the Opposition, you may make your statement now.

රු සංචීත් ප්‍රේමදාස මහතා
(මාණ්‍යුමිකු සංඝිත් පිරෝමතාලා)
(The Hon. Sajith Premadasa)

රු කථානායකතුමනි, මට මේ අවස්ථාව ලබා දීම ගැන ඔබතුමාට ස්තුතියි. මා මේ ගැන ඔබතුමාට දැනුම් දුන්නා. දියවත්තා ඔයේ පරිසර දූෂණය ගැන මාර්තු 24වැනි දා මම ප්‍රකාශ කළා. ඔබතුමාට පාර්ලිමේන්තුවේ බලධාරීන් උපදෙස් දුන්නා, එය පොකුණක් සුද්ද කිරීම හේතුවෙන් සිදු වුණු දෙයක් කියලා. අද ද්‍රව්‍යේන් ඒක අසත්‍යයක් බවට පත් වෙලා තිබෙනවා. ඔබතුමා මේ ගැන නොදුට සොයා බලන්න. මා හිතන විධියට, drainage systems, sewerage systems ගණනාවක් දියවත්තා ඔයට යොමු කර තිබෙනවා. එම නිසා අද දියවත්තා ඔයේ බරපතල පරිසර භාණියක් සිදු වෙලා තිබෙනවා. මා මේ ගැන එදා සඳහන් කරන කොට අද ඔය කැ ගහන ඇත්තේ මට දොස් කිවවා, ලකුණු දමා ගන්න හදනවා කියලා. ඒ වාගේම, අසහා වවතන් පාවිච්චි කළා. නමුත් අද මම ඔබතුමාගෙන් බොහෝම කාරුණික ඉල්ලීමක් කරනවා, රු කථානායකතුමනි. එදා සිට අද ද්‍රව්‍ය දක්වා ඒ වෙනුවෙන් කිසීම දෙයක් කරලා තැහැ.

Hon. Leader of the Opposition, you may make your statement now.

Mr. Guth Prame
(Sajith Pilarithosa to Inbuiku)
(The Hon. Sajith Premadasa)

Your Excellency, I am grateful to you for giving me the peace of mind. I wish you the best of luck. The sender will be notified on March 4 of the environmental pollution. You are advised by the authorities in Parliament that it was a sound that caused the cleaning of the shack. Even today there are corpses that have become untrue. Please find out more about Sham Geena. In my opinion, the sender of a drainage system, a sewerage system, will be directed to you. Therefore, the sender today has suffered serious environmental damage. I'm not sure what I'm talking about then, but I'm trying to score points. In the same way, obscenities are not used. But today I would like to make a very kind request to you, Hon. Thanaya. From that day to this day, there is no sound like that.

It is easy to look like a fool using Google Translate

Hon. Leader of the Opposition, you may make your statement now.

Honourable Mr Sajith Premadasa

(This is in Tamil and is presumably his name)
(The Hon. Sajith Premadasa)

Honourable Speaker, thank you for giving me this opportunity. I notified you ([honorific form](#)) of this. I made a statement on the pollution in Diyawanna Lake on March 24th. The parliamentary officials advised you that it was due to a pond being cleaned. Today that too has become a falsehood. Please investigate this closely. I think a number of *drainage systems, sewerage systems* have been diverted to Diyawanna Lake. Because of this, significant environmental harm has been caused to Diyawanna Lake. When I spoke about it previously, those who are shouting at me today said ([literal trans. 'scolding'](#)) I was trying to score points. Also used obscene words. But today, I make a very kind request from you, Honourable Speaker. Nothing has been done about this from that day to now.

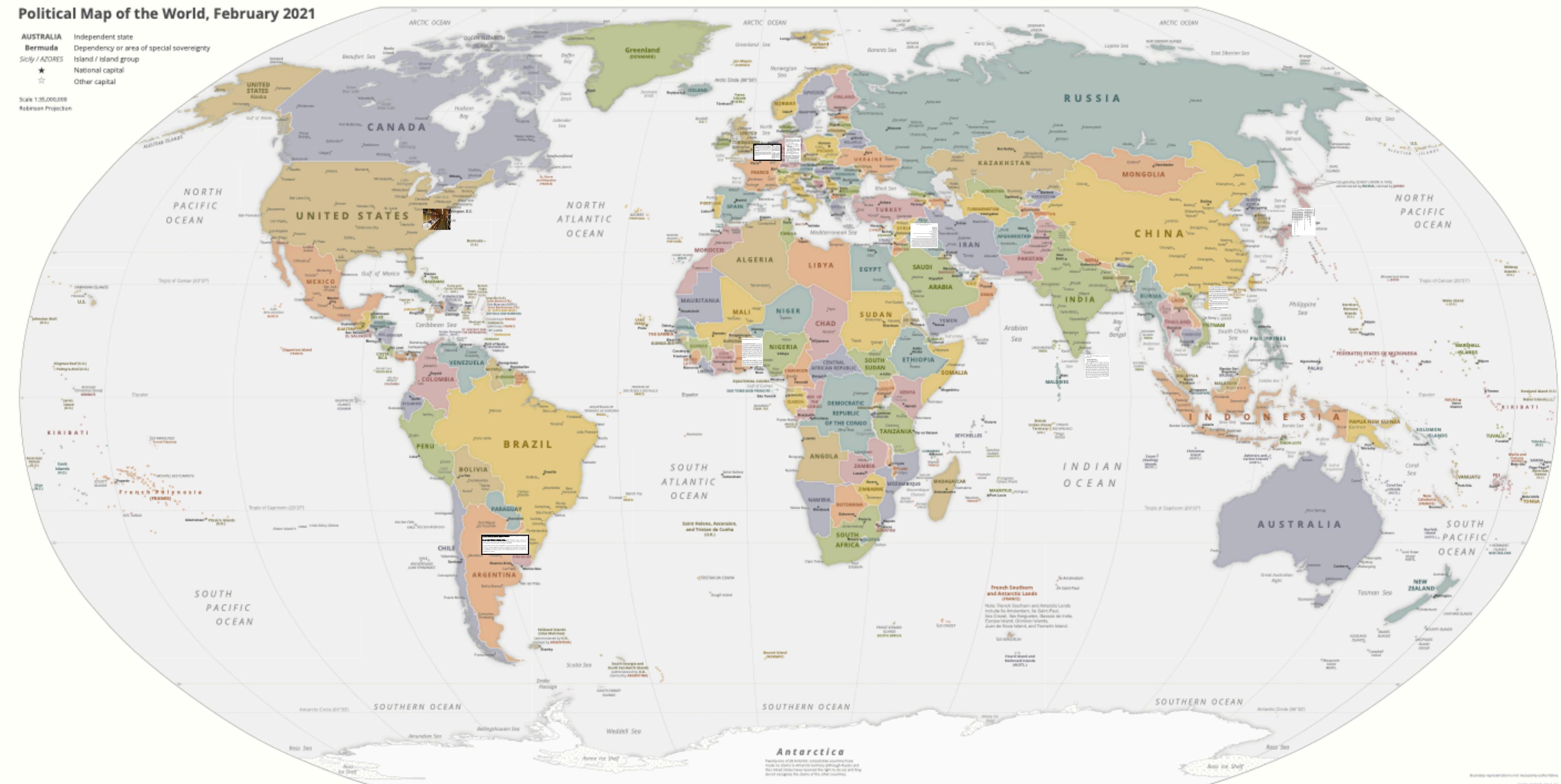
Hon. Leader of the Opposition, you may make your statement now.

Mr. Guth Prame

(**Sajith Pilarithosa to Inbuiku**)
(The Hon. Sajith Premadasa)

Your Excellency, I am grateful to you for giving me the peace of mind. I wish you the best of luck. The sender will be notified on March 4 of the environmental pollution. You are advised by the authorities in Parliament that it was a sound that caused the cleaning of the shack. Even today there are corpses that have become untrue. Please find out more about Sham Geena. In my opinion, the sender of a drainage system, a sewerage system, will be directed to you. Therefore, the sender today has suffered serious environmental damage. I'm not sure what I'm talking about then, but I'm trying to score points. In the same way, obscenities are not used. But today I would like to make a very kind request to you, Hon. Thanaya. From that day to this day, there is no sound like that.

Politicians talk a lot - everywhere



Politicians talk a lot - everywhere

Iraq



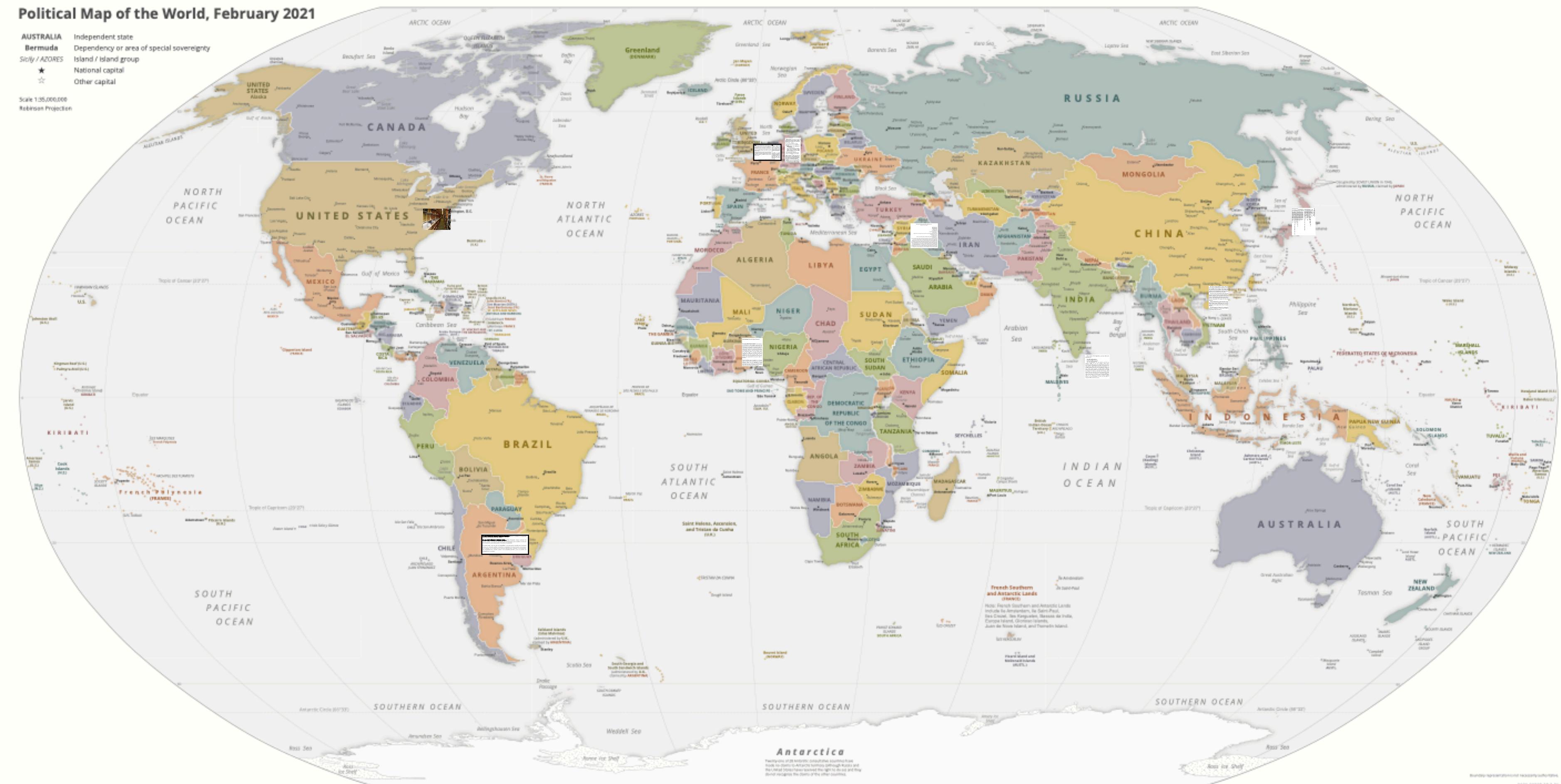
محضر الجلسة (40) من الفصل التشريعي الاول السنة التشريعية الثالثة
التاريخ: الأربعاء 10 أيلول 2008

الدورة الانتخابية الأولى
السنة التشريعية الثالثة
الفصل التشريعي الأول
الجلسة رقم (40)
الإثنين (12/7/2008) م
م / محضر الجلسة
بدأت الجلسة الساعة (1:40) ظهراً.
- الشيخ خالد العطية:-
نيابة عن الشعب نفتتح أعمال الجلسة الأربعين من أعمال مجلس النواب العراقي بقراءة آيات من القرآن الكريم.
- السيد محمد ناجي السامرائي:-
يقرأ آيات من القرآن الكريم.
- الشيخ خالد العطية:-
الى ان تجهز الصيغة النهائية لقانون انتخابات المحافظات نبدأ بالفقرات التالية من جدول الأعمال.
- السيد مثل الألوسي:-

يؤسفني أن أبلغ هيئة الرئاسة والأخوة والأخوات الزملاء أعضاء مجلس النواب الموقر إن صباح هذا اليوم قامت مجموعة إرهابية بتفجير منزل والدي وعائلتي في حي الجامعة تفجيراً بالكامل. ويؤسفني أيضاً أن أقول أن هذا الأمر وهذه التهديدات بلغتني منذ أيام بعد أن قمت بإسمكم جميعاً وأعلنت أنا واحد من أعضاء مجلس النواب أمثل جميع الأحزاب والطوائف والكتل وإذا كنت هنا في الشارع أقوم بواجبي فأخوتي وزملائي يقومون بواجب آخر وطني وواحدنا يكمل الآخر. وقلت أنا هنا ليس باسم حزب الأمة العراقي بل باسم المجلس الإسلامي الأعلى والتيار الصدري والفضيلة والدعوة والتوافق والوفاق والعراقية وكل القوائم والأعضاء ذكرتها بالترتيب لأنني أشعر أنا واحد منكم وكلنا لبعضنا من أجل هذا الوطن، لكن التهديدات مستمرة لأننا طالبنا بإعادة المهجرين إلى منازلهم. ويجب أن أخبر هيئة الرئاسة الموقرة إن كلا الطرفين المهجران الذي لا زال يسكن في حي الجامعة وبمفردات أخرى لا أحب أن أذكرها لكن للتوضيح أهاليينا من السنة الساكنين في حي الجامعة وهي العدل يقولون لي وبرسائل موثقة يقولون عودوا إلينا بأهلاًنا الشيعة والشيعة المهجرين يقولون دعونا نذهب إلى أهلاًنا السنة ونعود إلى الحياة الطبيعية. وحصلت على دعم رئيس الوزراء الأخ نوري المالكي في هذا الشأن لكن الدنيا قامت ولم تقدر وأصابع الإتهام واضحة لمن لا يريد الخير للعراق ولمن يشعر أن هذا المشروع يهدد



Politicians talk a lot - everywhere



Politicians talk a lot - and always have

Hong Kong
(an 1895 debate
on gun control)



The ACTING ATTORNEY-GENERAL—I beg to move the second reading of the Bill entitled An Ordinance to amend and consolidate the law relating to the carriage and possession of arms and ammunition. I will ask honourable members to turn to the statement of objects and reasons printed at the end of the Bill and which are as follow:—

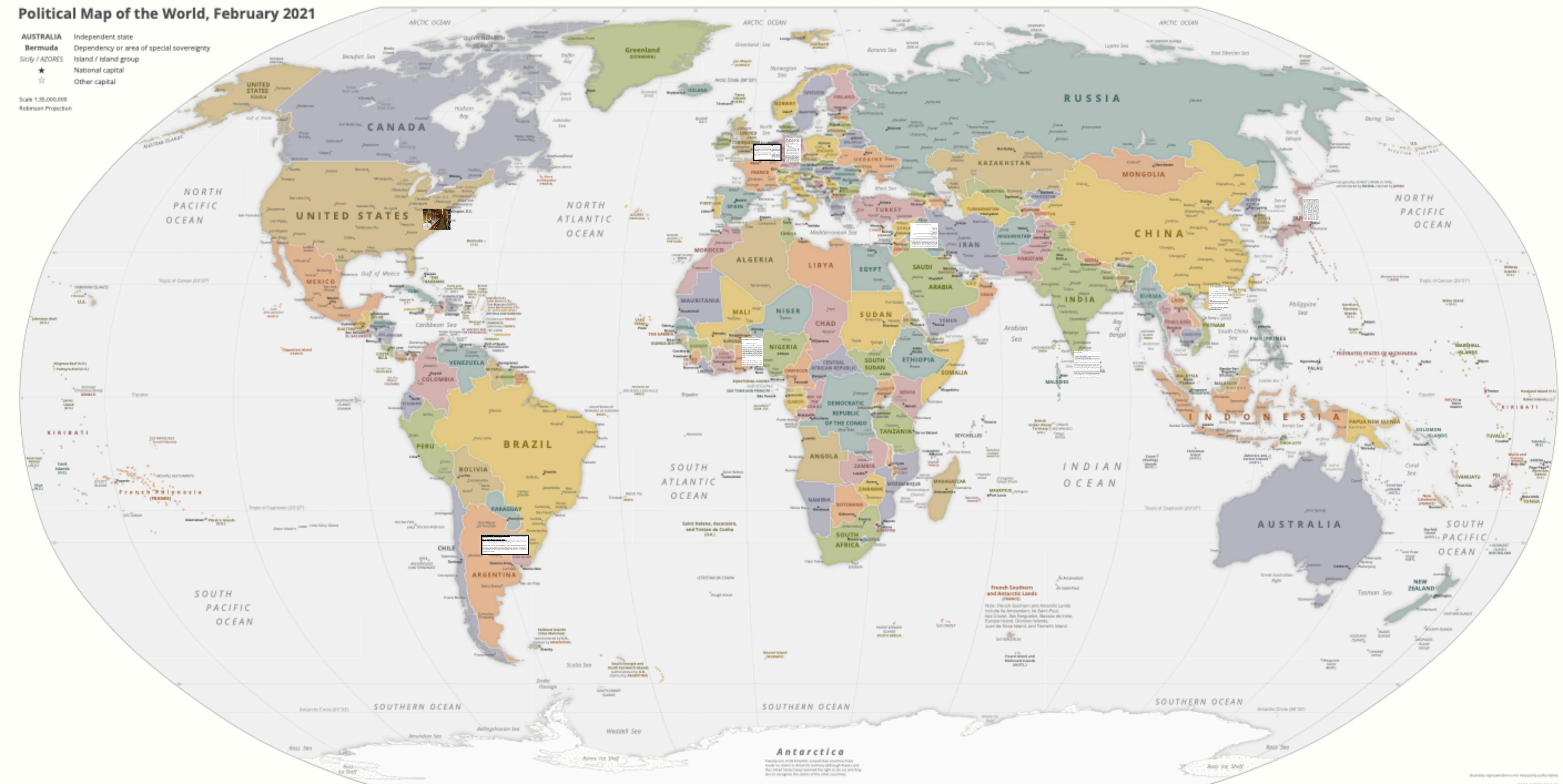
"The trade in arms and ammunition in this colony has increased so much of late that it is thought desirable that it should be subjected to strict supervision, and the recent increase of armed robberies and piracies in the adjacent provinces of the Chinese Empire, to the detriment of our trade, has demonstrated the necessity for such strict supervision.

"The present Bill, therefore, while it practically re-enacts many of the provisions of Ordinances 8 and 14 of 1895, which it repeals, includes certain entirely new provisions (marked new in the margin) which will now be briefly discussed.

"This Bill amends the existing law by rendering the possession of ammunition, as well as of arms, without a proper authority, illegal. (See clause 5.)



Politicians talk a lot - and always have



Politicians talk a lot - and always have

Japan's Meiji Imperial Diet, 1893



衆議院議事速記録第二十四號 明治二十六年一月十一日 豫算案申號歲出臨時部 文部省所管

五九六

於キマシテモ決シテ之ヲヤラヌト云フ御話デハナリ、他ニ方法ガアルト云フ御話デアリマシタガ、若シ左様ナ譯デアリマスレバ、尙更此ノ如キ事柄ハ矢張帝國大學内ニ於テ致スコトガ甚ダ至便デ且ツ至當デアラウト考ヘル、之ヲ要スルニ一ハ工業ヲ盛ニシ國家ヲ富マス所ノ基本ヲ作ル事柄デアリ、一ハ吾人ノ最モ重ンズベキ生命ヲ安全ニ保續スル所ノ方法ノ要項デアリマシテ、孰モ国家ノ富強ニ關係ヲスルモノデアリマスカラ、宜シク諸君ノ御熟考アッテ要求ノ通り決議アランコトヲ希望致シマス、尙ホ終ニ一言申シテ諸君ノ御参考ニ供シテ置キタイコトガアリマス、ソレハ他デハアリマセヌガ帝國大學ノコトデアリマス、帝國大學ハ未ダ完全ナル建物ヲ持ッテ居ラヌト云フコトデアル、即チ不整頓極タル建物デアルト云フコトヲ申スノデ、ソレ故ニ今年ノミナラズ今後モ年々多少ノ建築ヲ矢張請求致サナケレバナリマセヌ、抑帝國大學ハ去ヌル十九年ニ東京大學ニ工部大學ト云フモノヲ合併致シタ、又二十三年ニ東京農林學校ヲ合併致シタモノデアリマス、而シテ其位置ハ本鄉元富士町ニアリマスガ、其處ハ元來醫學部ノミヲ建テタ所デアリマス、然ルニ此法學部理學部文學部此三學部ヲ此處ニ移スコトニナリマシテ、明治十七年及十八年ニ一つ橋外ノ處カラ此三學部ヲ此處ニ移シマシタ、然ルニ各學部ノ校舎ト云フモノハ何モナイ、唯法科ト文科ト此二ツノタメニ設ケテアル一棟ガアリマス、ソレハ即チ今書籍館ニ假用シテ居リマスガ、其他ニハ何モ建テナカツタノデアリマス、一時從來在ル所ノ醫學部其醫學部ノ事務室ノ一部分、ソレカラ先刻長谷川君ノ御話ノアツタ非常ニ古イ御殿、ソレカラ病院ノ一部分、爾來理科大學ノ一部分ヲ即チ化學ノ教室ヲ新築ヲ致シマシタ、是ハモウ既ニ落成ヲ致シタ、此建物ニ化學、物理學、數學ナドノ教室ガ移シテアルマデマシテ工科大學ハ明治十九年工部大學校合併ノ時ニ之ヲ新築致シマシタ、又二十二年以來圖書館、博物教室、是等モ亦完全デハアリマセヌガ大半新築ヲ致シテ落成ヲスル様ニナツテ居リマス、要スルニ帝國大學ニ要スル建物ハ丁度總計務ヲ大學ニ屬シタリ、ソレ等ノ事務ヲ執ル所ヲ要スルト云フヤウナ話デ、而シテ工科大學ハ明治十九年工部大學校合併ノ時ニ之ヲ新築致シマシタ、又二十二年以來圖書館、博物教室、是等モ亦完全デハアリマス、或ハ教授上或ハ實驗上其用ニ堪ヘヌ所少カヌ譯デアリマス、然カノミナラズ近來學術ノ進歩學生ノ增加等ニ依ッテ段此建築ヲ要スルモノガアリマシテ、當分ノ中ハ矢張年々此新營費ヲ要スル譯デアリマスガ、是ハ實ニ萬止ムヲ得ナイデアリマシテ、今迄大學ノタメニデ一萬坪アリマスガ、其内二千坪許リ漸ク煉瓦造リテ出來テ居ル、後トマダ八千三百坪許リテ、木造テ古イ建物ハ追々腐朽ヲシテモ又修繕スルコトモ出來止ムヲ得ヌモノデアリマス、此他追々新營ヲ要スルモノハ年々請求ヲ致シマス、ソレカラ又圖書館モマダ完全ニナツテ居ラヌ、是モ未ダ建増ヲシナケレバナリマセヌ、ソレカラ

○教室モ建増ヲシナケレバナリマセヌ、ソレカラ植物園ニ溫室ガ未ダゴザリマセヌ、ソレカラ農科大學ガ御承知ノ如ク誠ニ古イ酷イ家デ、是モ年々改築ヲシナケレバナリマセヌ、ソレカラ大學本部ノ事務室ガゴザリマセヌ、ソレカラ大ナル講堂ト云フモノガナイ、其他吸水ノ如キ排水ノ如キ瓦斯杯ノ附屬物ガ未ダ整頓シテ居リマセヌ、此ノ如ク大學ハ尙ホ漸次ニ經費ヲ要スル様ナ狀況デアリマスル、故ニ諸君ニ於テモ御含ミ置カレンコトヲ希望致シマス

○議長(星吉君)質問デスカ

○鈴木萬次郎君(六十一番)質問デス

○鈴木萬次郎君(六十一番)宣シイ

○鈴木萬次郎君(六十一番)質問デス

○鈴木萬次郎君(六十一番)質問デス

○鈴木萬次郎君(六十一番)質問デス

○政府委員(久保田讓君)是ハ先刻主査ノ長谷川君カラモ御報告ニナツテ、政府委員ハ勿論茲ニ書イテアル通リト申サナケレバナリマセヌガ、豫算委員ノ三君デスカ四君デスカ今年實地ヲ御覽ニナツテ、是ハ止ムヲ得ナイモノデアルト云フコトニ御認メニナツタ譯デアリマスカラ、左様御承知ヲ願ヒマス

○鈴木萬次郎君(六十一番)若シ政府委員ノ自ラ醫科大學ノ病室ト、ソレカラ一般私立出來テ居ル病院ノ、所謂入院料ヲ取テ病人ヲ入レテ居ル一般ノ病院ノ病室トヲ御較ベニナリマシタコトナラバ、今ノ醫科大學ノ此假病室ハ立派デアルカ、又堅牢デアルト御覽デアリマセウカ、私立病院ノ病室ヨリハ餘程下ヅテ居ルト御覽デアリマスカ

○政府委員(久保田讓君)私ハ私立病院ハ委シク知リマセヌガ、當年或ル病院ノ病室トヲ御較ベニナリマシタコトナラバ、今ノ醫科大學ノ此假病室著シク往ケマセヌ、酷クナツテ居リマス、外ハ餘リ委シク知リマセヌ

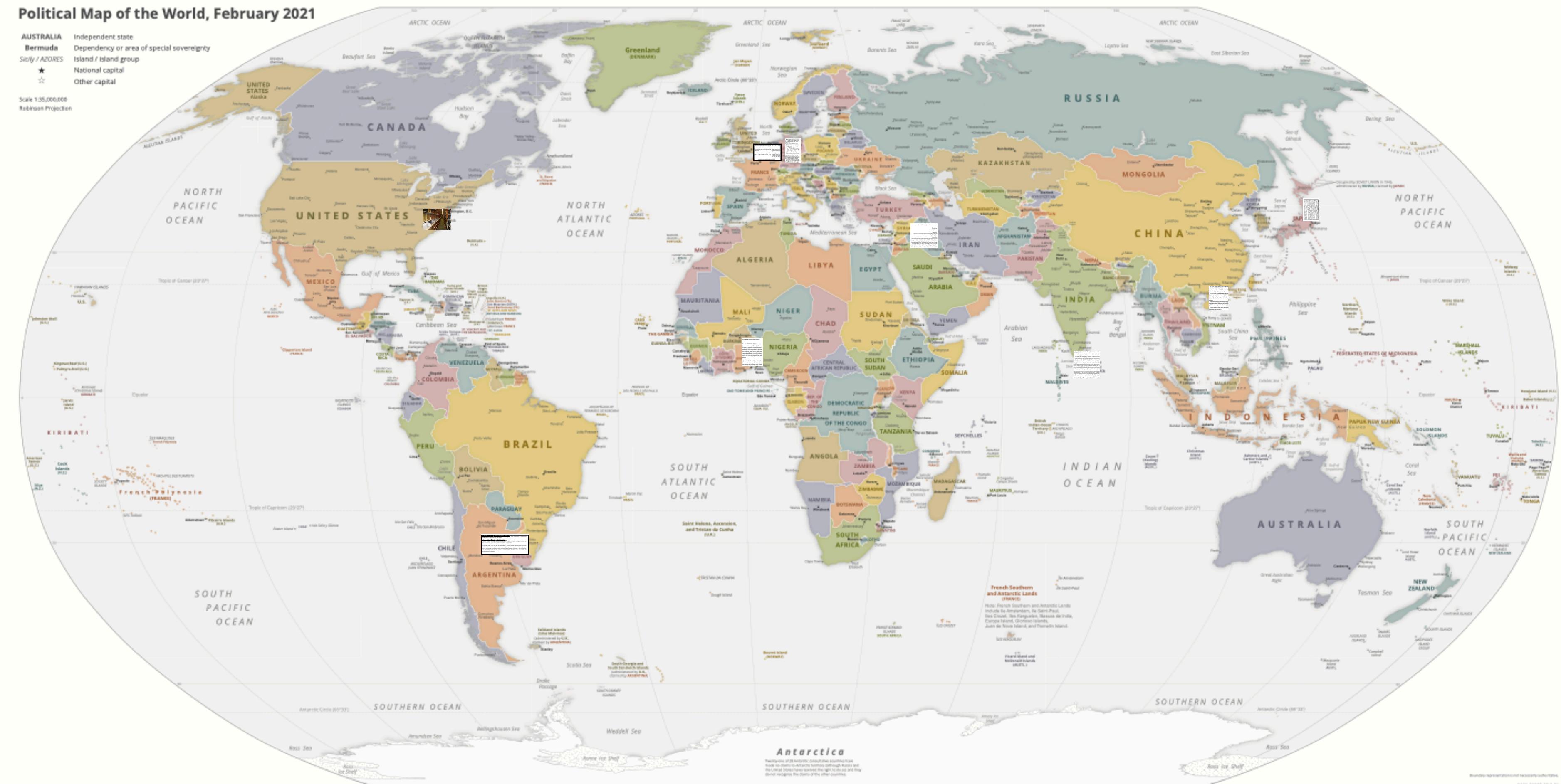
○鈴木萬次郎君(六十一番)大學ノ病院ノ方ガ惡ルイト云フコトデアリマス

○政府委員(久保田讓君)サウデス

○鈴木萬次郎君(六十一番)私ノ見タ所デハ——、尙ホ質問デアリマスガ、一向變ラヌ寧ロ宜イト云フ方ニ見受ケマシタガ、唯茲ニ質問致シテ置カケレバナラヌノハ、大學病院ハ此說明書ニアルガ如ク病人ヲ實地研究ヲ致ス、モウ一ツ進シテ言ヘバ之ヲ或ハ試驗的ニ用フルモノデアル、故ニ政府委員モ知ラル、如ク各國ニ於キマシテハ入院料ヲ取ラズ、之ヲ徒療治ラシテ居ルノデアリマス、所謂お客トスルノデハナリ、然ルニ若シ他ノ一般ノ私立病院ヨリ大變壯大ナル立派ナル病院ヲ擇ヘ、サウシテ今日迄ノ如ク實地研究ニ之ヲ用ヒ、其上ニ入院料ヲ矢張取ルト云フ方針デハドウ云フモノデヤラウカ、矢張政府ニ於テハ實地研究ニ之ヲ用ヒナガラ矢張入院料モ取シテ、私立病院



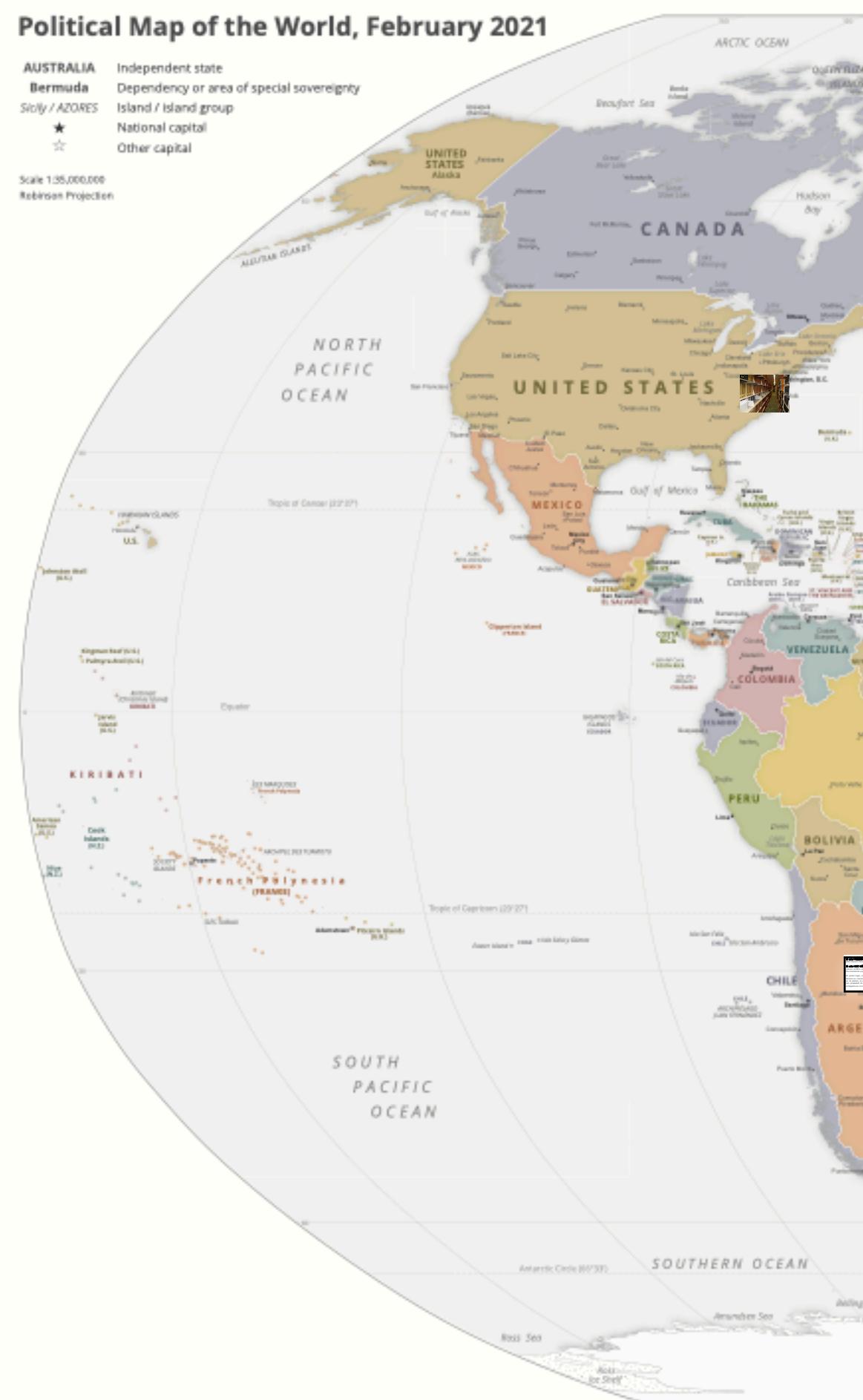
Politicians talk a lot - and always have



Politicians talk a lot - and always have

Weimar Republic

Göring introducing Hitler on March 23, 1933, to begin debate on the Enabling Act.



innerstag den 23. März 1933.

25

Präsident Göring: Meine Damen und Herren! (C)

Ich lasse jetzt über den Antrag abstimmen. Ich bitte diejenigen Damen und Herren, die für die Ablehnung des Antrags auf Haftentlassung sind, sich von den Sitzen zu erheben.

(Geschieht.)

Das ist die Mehrheit; der Antrag ist abgelehnt.

Wir kommen jetzt zum zweiten Gegenstand der Tagesordnung:

Entgegennahme einer Erklärung der Reichsregierung

in Verbindung mit der

ersten und zweiten Beratung des von den Abgeordneten Dr. Frick, Dr. Oberfohran u. Gen. eingebrachten Entwurfs eines Gesetzes zur Begebung der Not von Volk und Reich (Nr. 6 der Drucksachen).

Das Wort hat der Herr Reichskanzler.

(Die nationalsozialistische Fraktion erhebt sich und begrüßt den Reichskanzler mit einem dreifachen Heil.)

Hitler, Reichskanzler: Männer und Frauen des Deutschen Reichstags! Im Einvernehmen mit der Reichsregierung haben die Nationalsozialistische Partei und die Deutschnationale Volkspartei Ihnen durch einen Initiativantrag ein Gesetz zur Begebung der Not von Volk und Reich zur Beschlussfassung unterbreitet.

Die Gründe für diesen außerordentlichen Vorgang sind folgende.

Im November 1918 rissen marxistische Organisationen durch eine Revolution die vollziehende Gewalt an sich. Die Monarchen wurden entthront, die Reichs-



So ... parliamentary records are a rich source of political, social, and historical information, offering the potential for fine-grained and dynamic understanding of representation, ideology, issue ownership, realignment, polarization, fragmentation, agenda setting, oppositional strategy, and on and on.

But they're "BIG."

Machine learning / NLP / "Text-as-Data" can help with that volume and velocity.

But what about that variety?

The fundamental existential problem of comparative social science



The Curse of Babel

בָּלְל (bālal) -
to jumble, confuse

Comparative social science is generally undertaken by ...

- Individual scholars with 2-5 languages (“area specialists”)
- Individual scholars who compare many countries using “pre-quantified” data from secondary sources, e.g., election results, FDI (“quants”)
- Polyglot groups of scholars superficially coordinated on a theme (“edited volumes”)
- Polyglot groups of scholars tightly coordinated in elaborate and expensive multinational infrastructure, e.g.,



Conventional text-as-data techniques are little help

- Individual documents are represented by features of the text.
- In a common “bag-of-words” or “one-hot” representation, a document is represented by presence/absence or counts, possibly weighted in some way, of the words, or similarly defined tokens, it contains.
- Some modern approaches — like that used in language model “BERT” — define features based on *pieces* of words, or even individual characters.
- Whatever the features in the representation, every task we might want to do — search, information extraction, topic modeling, etc. — leverages patterns of similarity / difference in those features. There must be mutual nonzeros.

Similar content, different representation

Bag-of-words, or
“one-hot encoding”

Nell'area dell'euro turbolenze e tensioni sui debiti sovrani e sui mercati borsistici sono riemerse dal mese di aprile scorso e si sono amplificate nei mesi estivi.

In the euro area, turbulence and tensions on sovereign debts and stock markets have re-emerged since last April and have amplified during the summer months.



People will be concerned about the turmoil in the world's financial markets and what it means for economies here and across the globe. I want to update the House on what we are doing to protect Britain from the storm and to help lead a more effective international response to the fundamental causes of this instability.



	<i>turbolenze</i>	<i>mercati</i>	...	<i>turmoil</i>	<i>markets</i>
	0	0	...	1	1
	1	1	...	0	0

Is translation a solution?

- A few parliamentary records are manually translated to provide parallel bi- or multi-lingual records (Canada, Belgium, Switzerland, Hong Kong, European Parliament, ...).
 - These, in fact, have been a key source of training data for machine translation.
- Bespoke manual translation is possible with documents the scale of party manifestos, but not at the scale of parliamentary records.
- Google Translate can make you look foolish, but tasks that don't depend on sentence-level syntax or semantics, like search or topic modeling, tend to do OK with machine translated documents, at least in well-studied languages. But ... machine translation at scale with, e.g., Google/Microsoft, is cost-prohibitive for most, or requires scholars to violate Terms of Service if not laws.

The objectives of the method I'll demonstrate today

- Using primary data on parliamentary speech in different languages, without manual or machine translation ...
 - Estimate a **multilingual topic model**, including topics that correspond roughly to “things governments have ministries of” and major cross-cutting issues like immigration or terrorism, which can be used to capture relative attention of MPs and aggregates like parties, over time, in different parliaments on shared metrics.
 - Estimate a **multilingual embedding space** that captures linear political semantics and complete cross-national party analogies (e.g., UK Conservative is to UK Labour as German Christian Democrat is to ____).
 - Estimate a **multilingual scaling model**, describing shared ideological relationships among political issues, and placing parties from different parliaments in a common space.
 - Spoiler alert — we demonstrate that these three constructs are, or can be, the same thing.
 - Use these to **answer substantive political science questions** like ... is there coherent cross-national ideological content within conventionally described “party families” (e.g., nationalist, green, liberal, left, etc.) ... is there commonality in the strategies deployed by political oppositions?

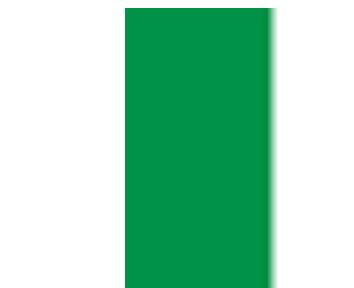
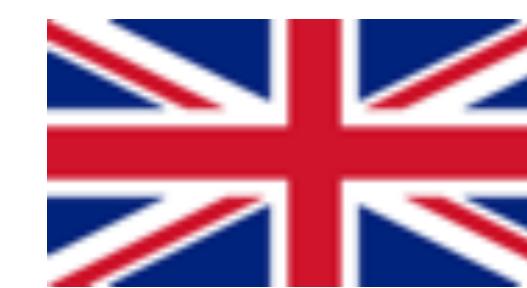
Data - Floor speech from eight lower houses

ParlSpeech (Rauh, et. al. 2017)



Chamber	<i>Riksdag</i>	<i>Eduskunta</i>	<i>Parlament</i>	<i>Tweede Kamer</i>	<i>Bundestag</i>	<i>Congreso</i>
Years	1990-2015	1991-2015	1993-2016	1994-2015	1991-2013	1989-2015
Speeches	317,132	245,852	329,135	900,796	299,844	290,680
Vocabulary	638,825	1,422,492	337,127	409,323	750,866	533,097

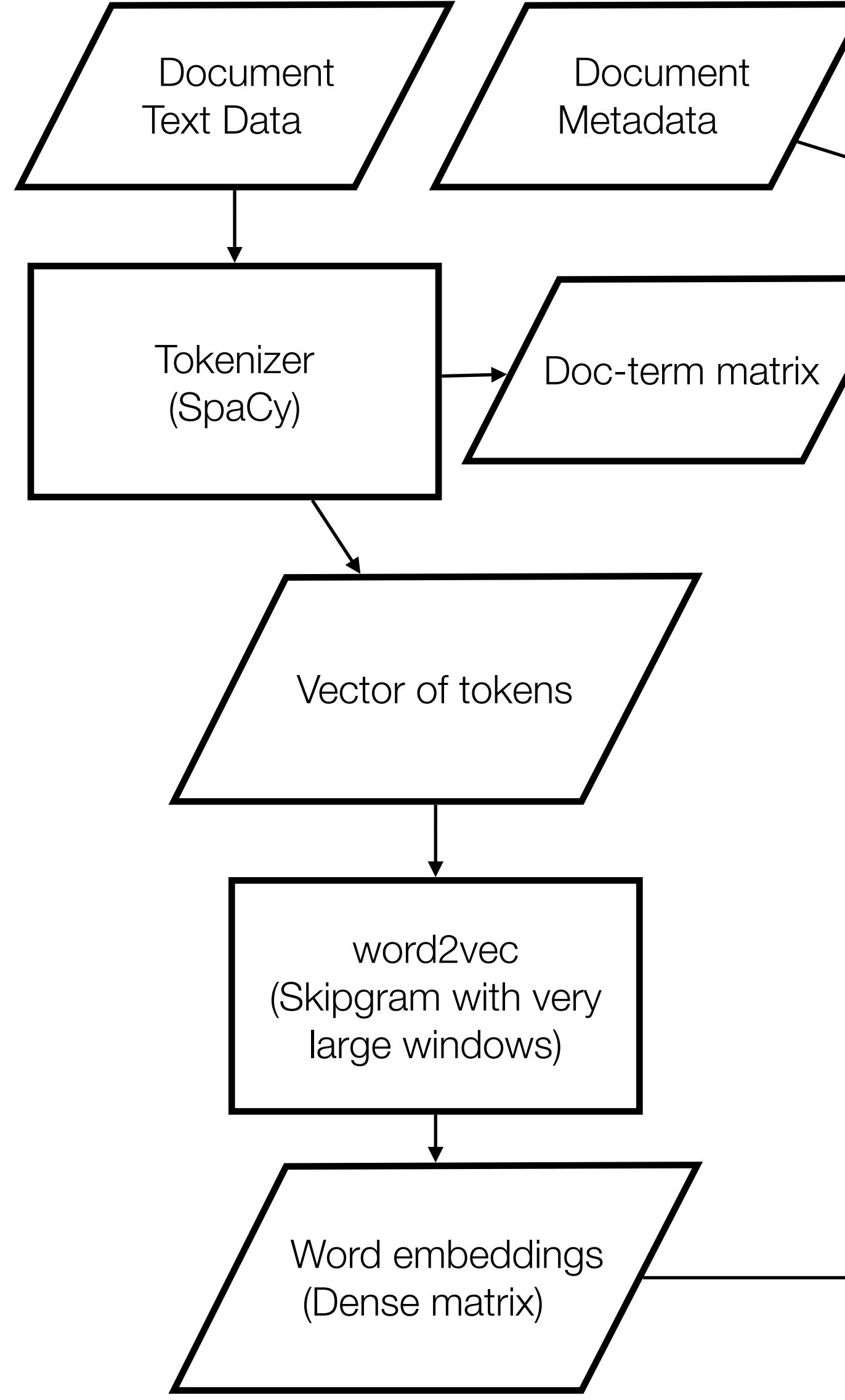
Custom



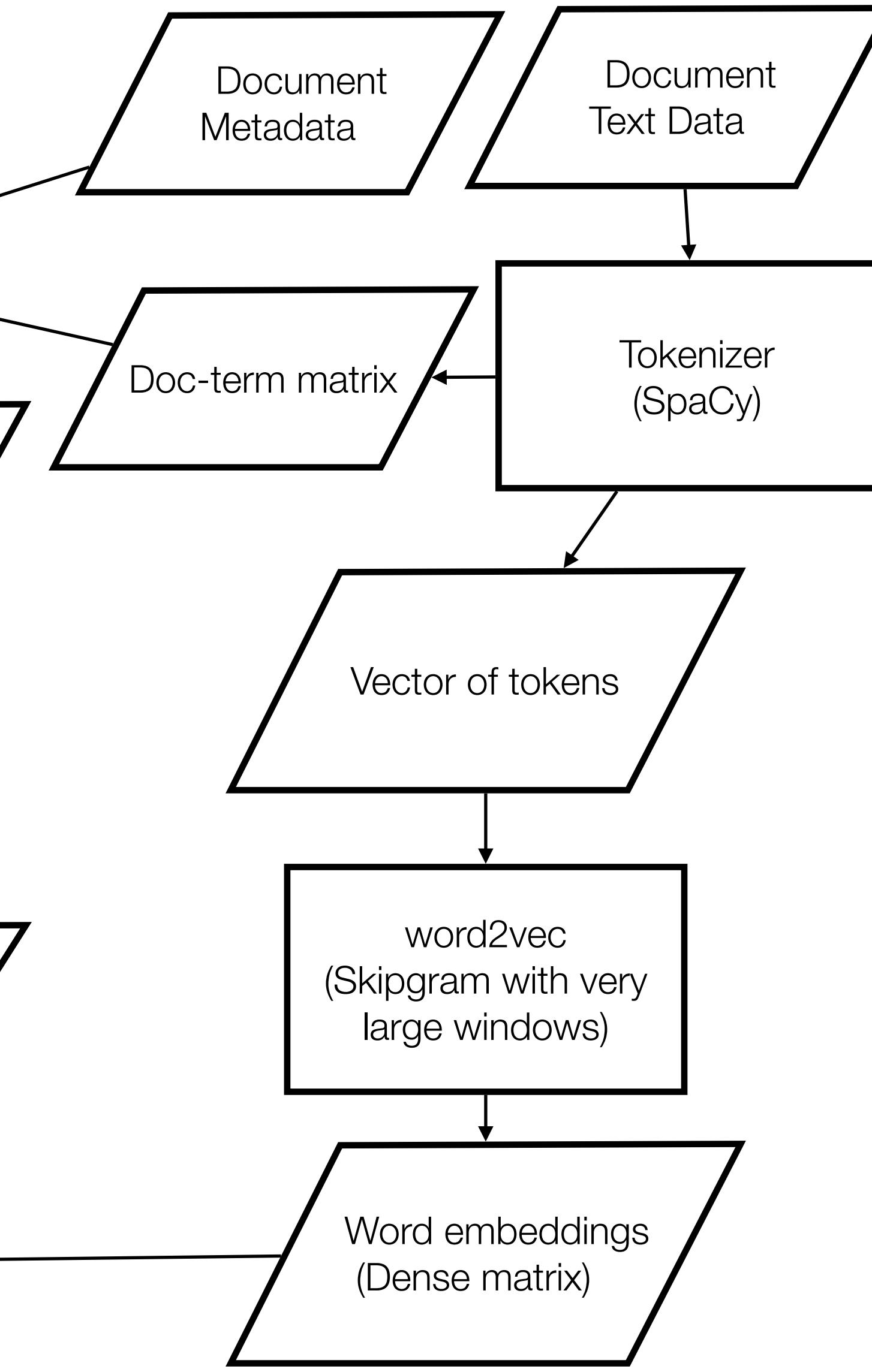
Chamber	<i>House of Commons</i>	<i>Camera dei Deputati</i>
Years	2006-2017	2006-2017
Speeches	602,763	215,022
Vocabulary	140,717	284,959

The pipeline ...

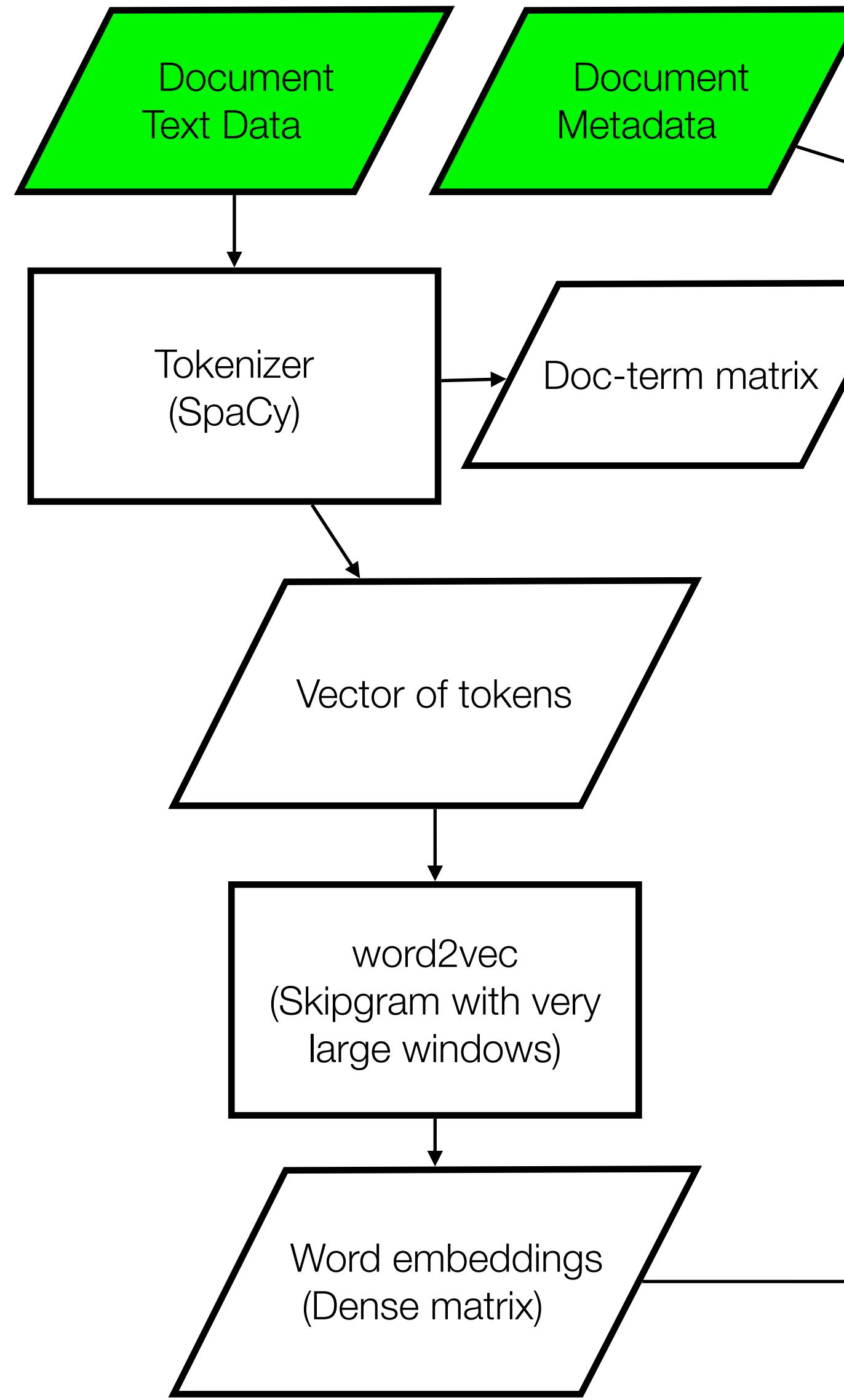
Language 1



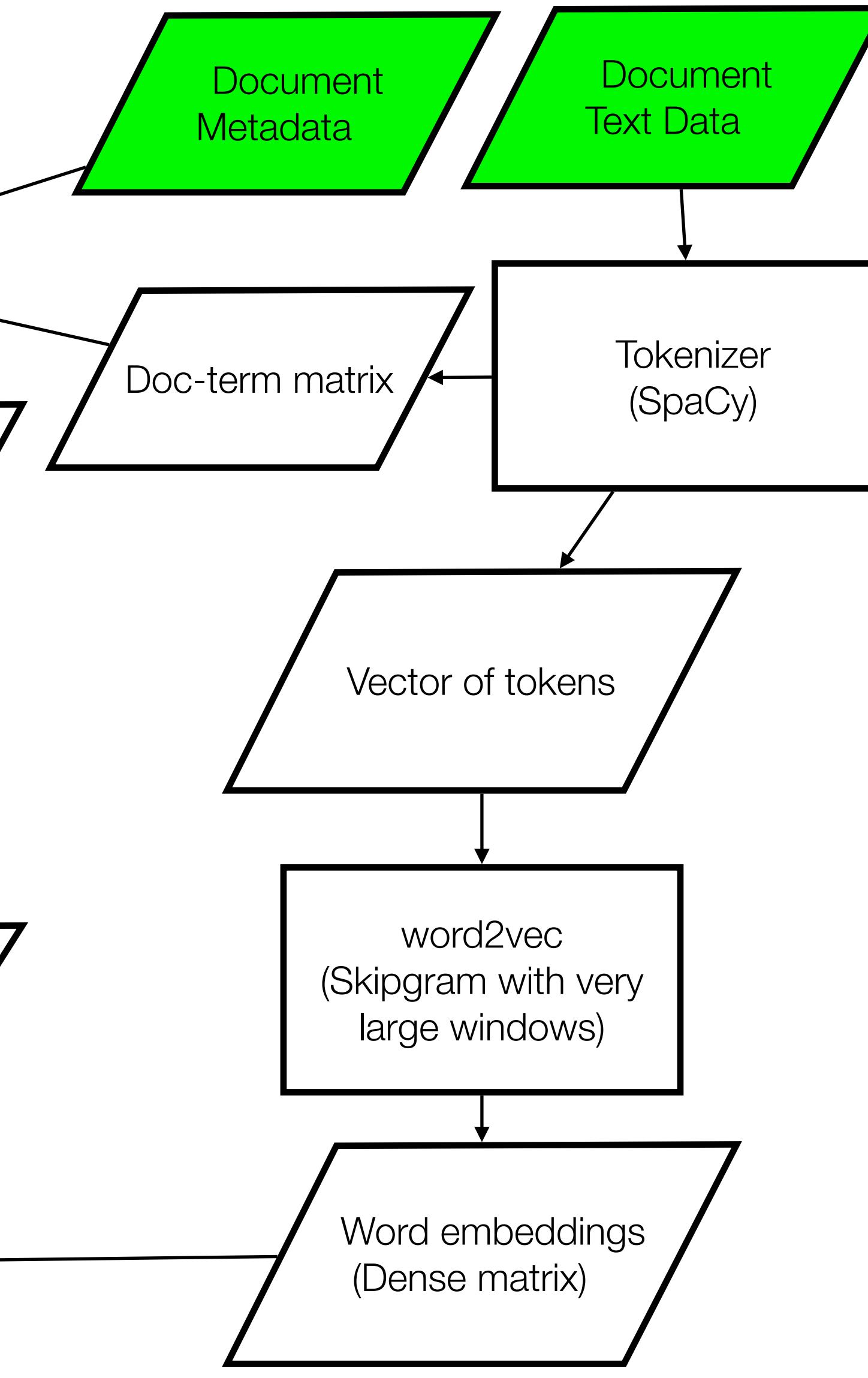
Language 2



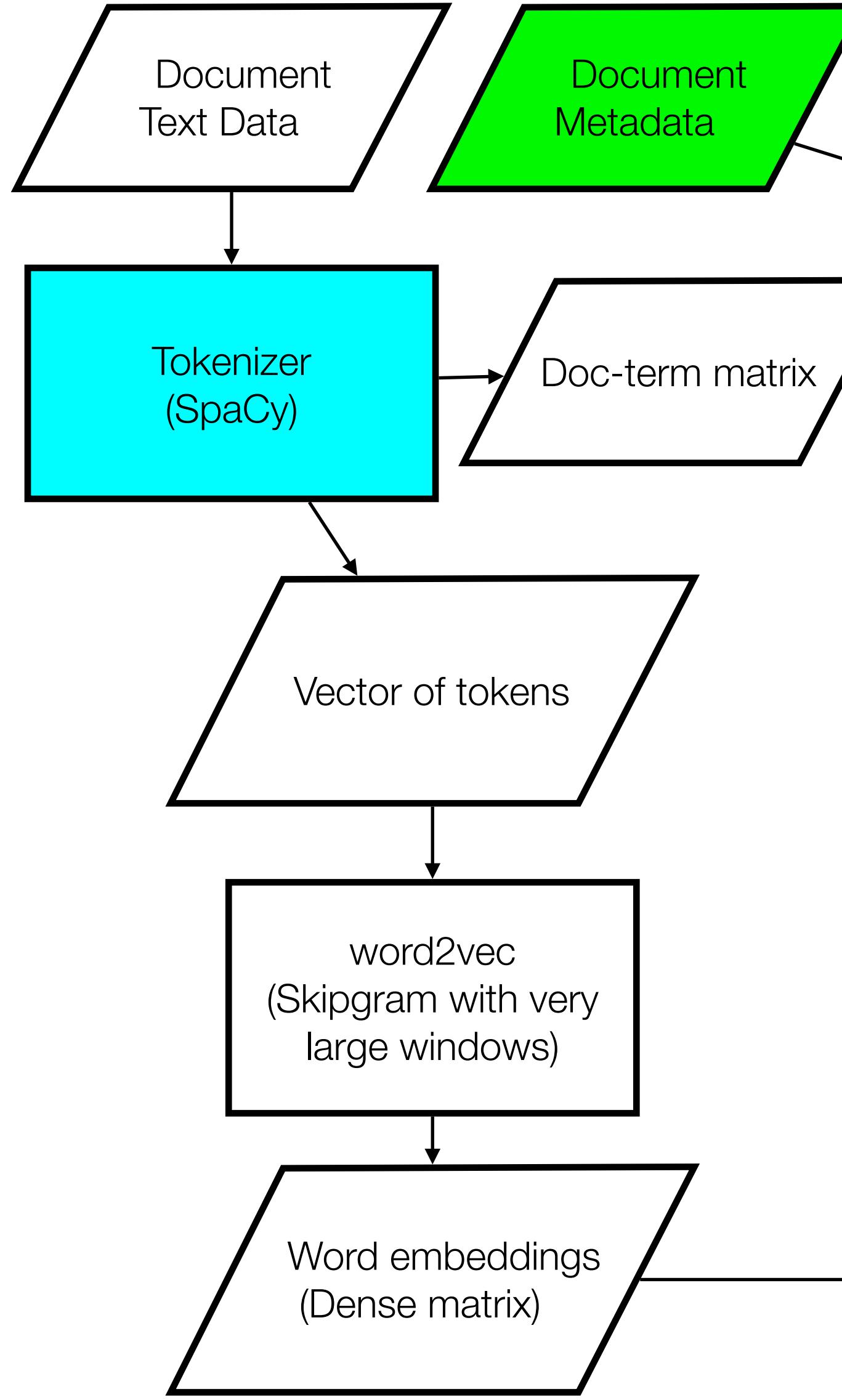
Language 1



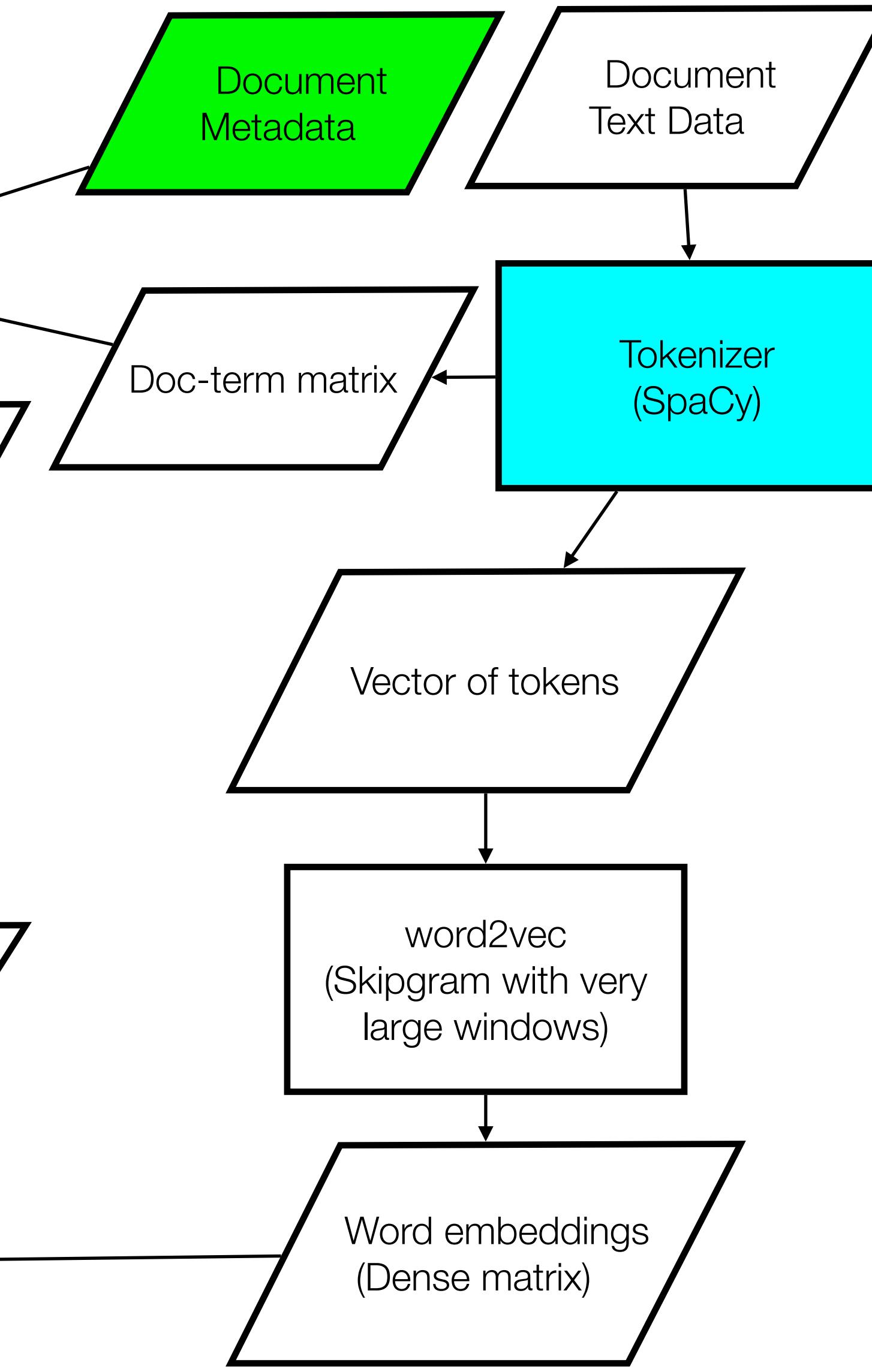
Language 2



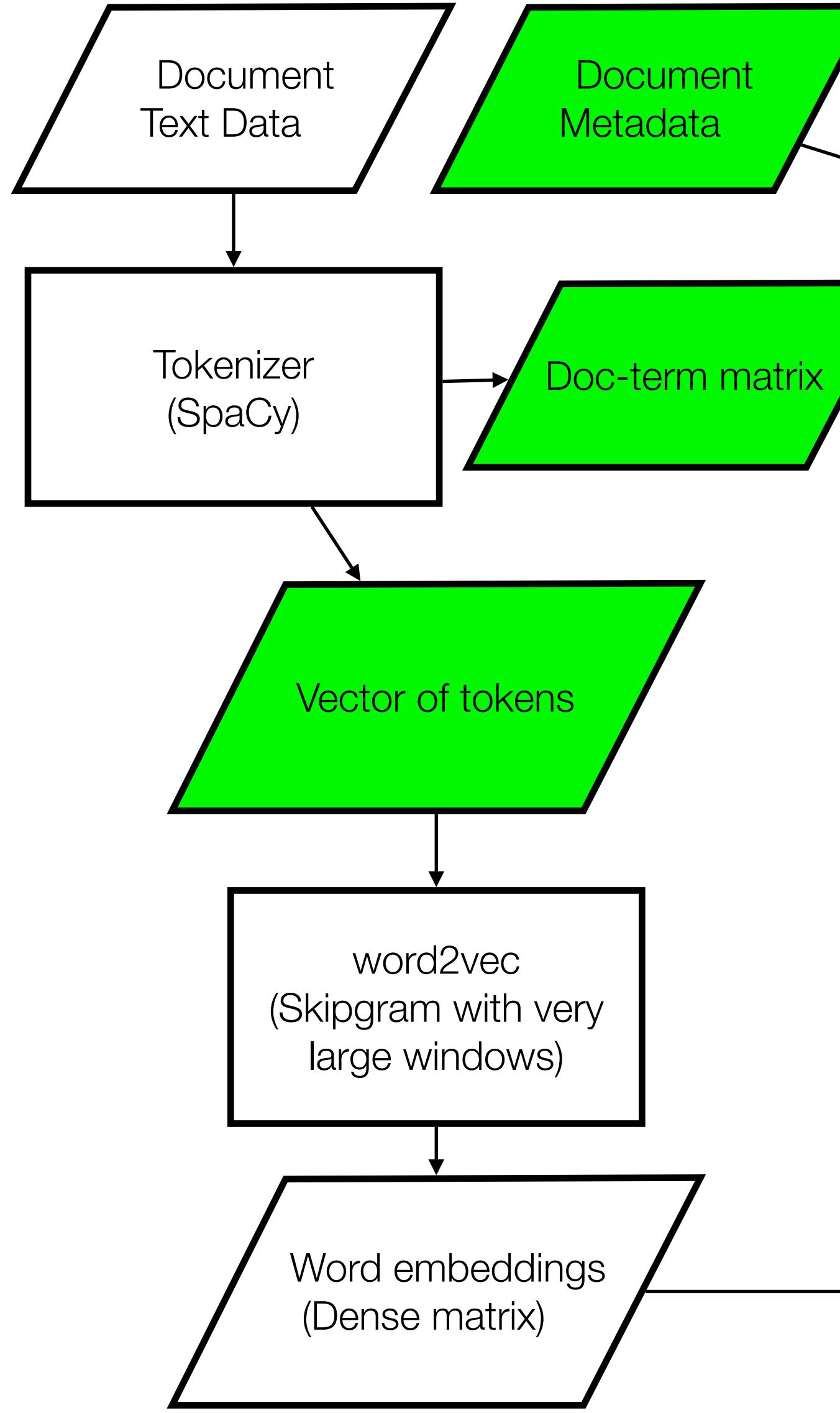
Language 1



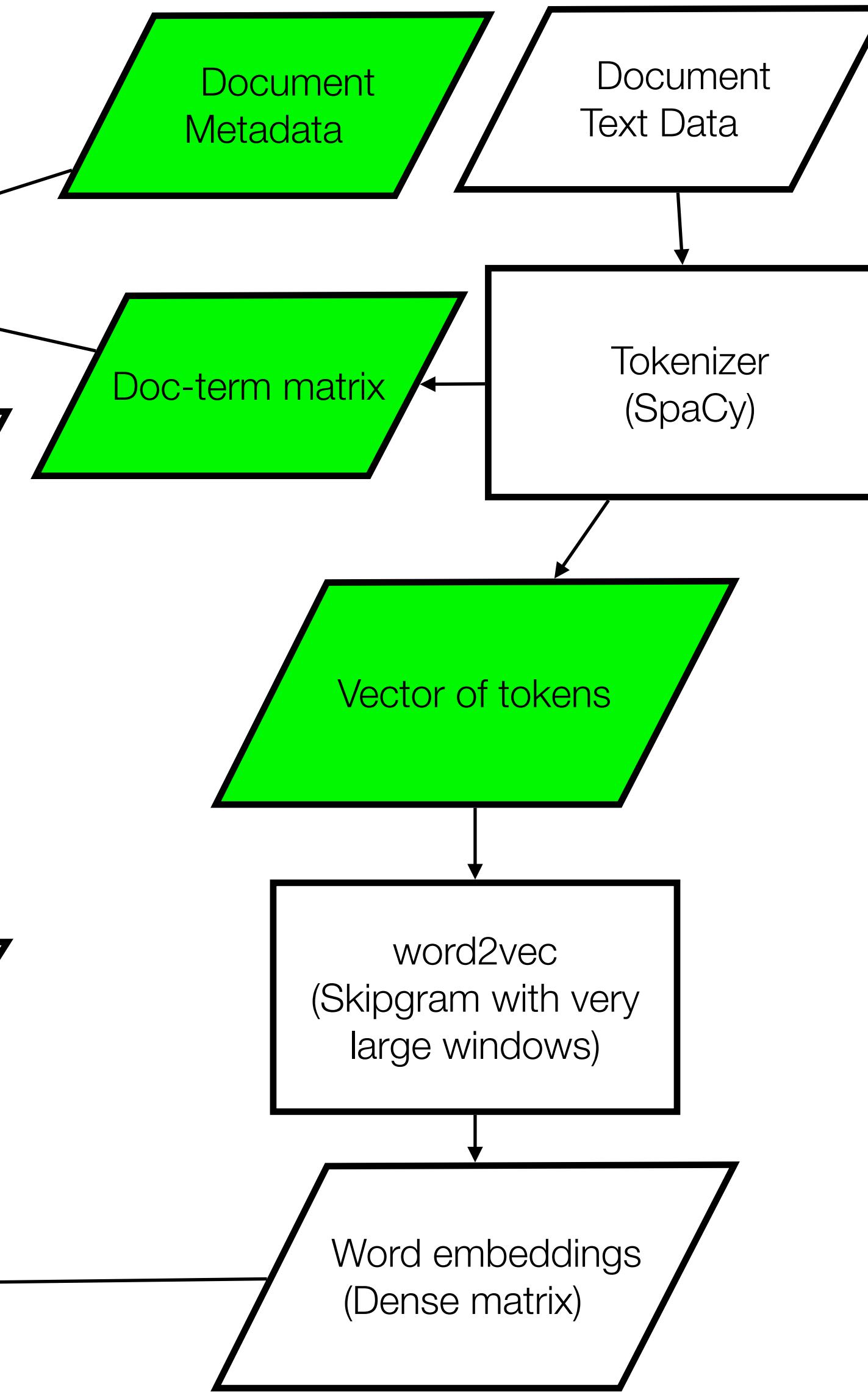
Language 2



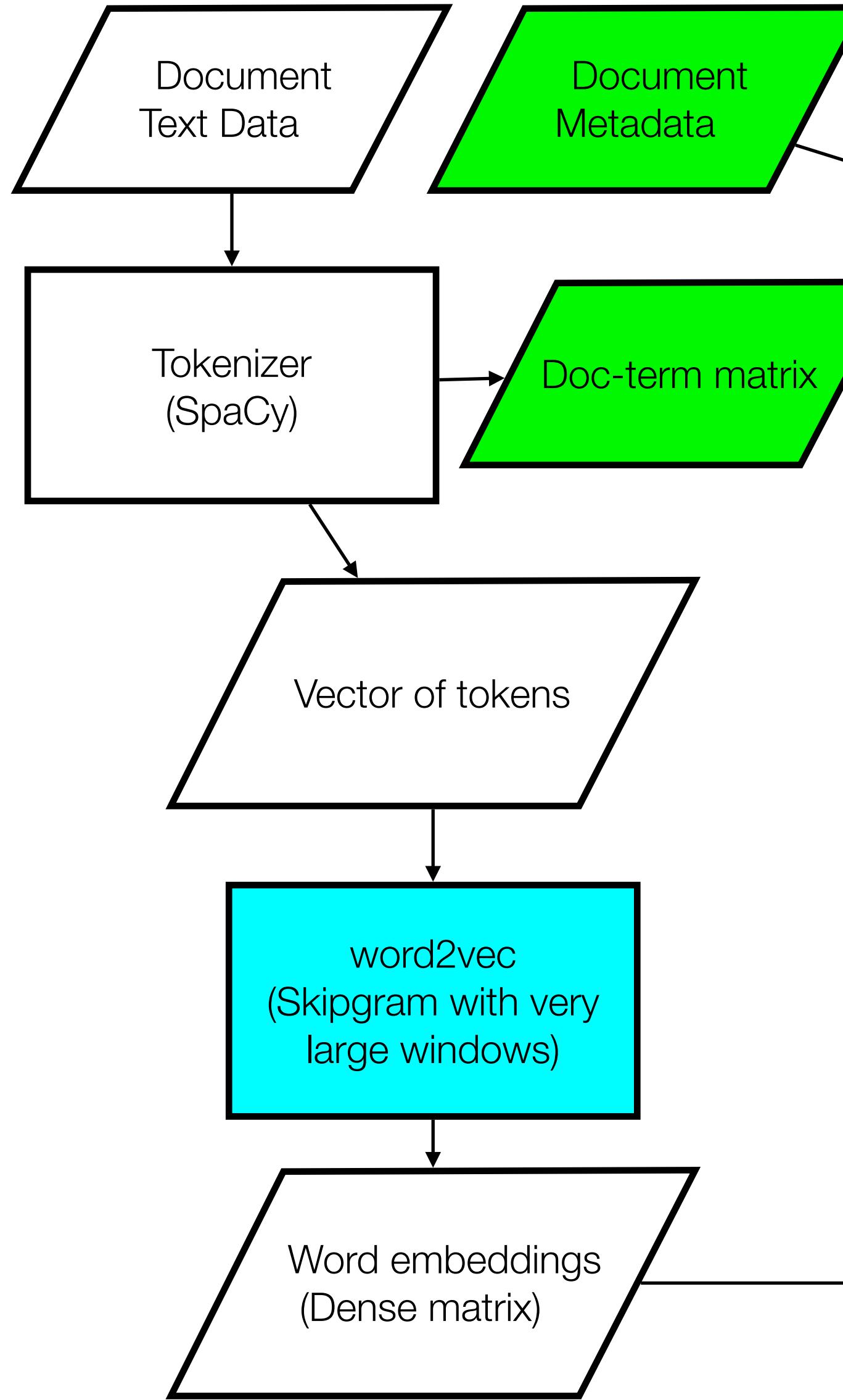
Language 1



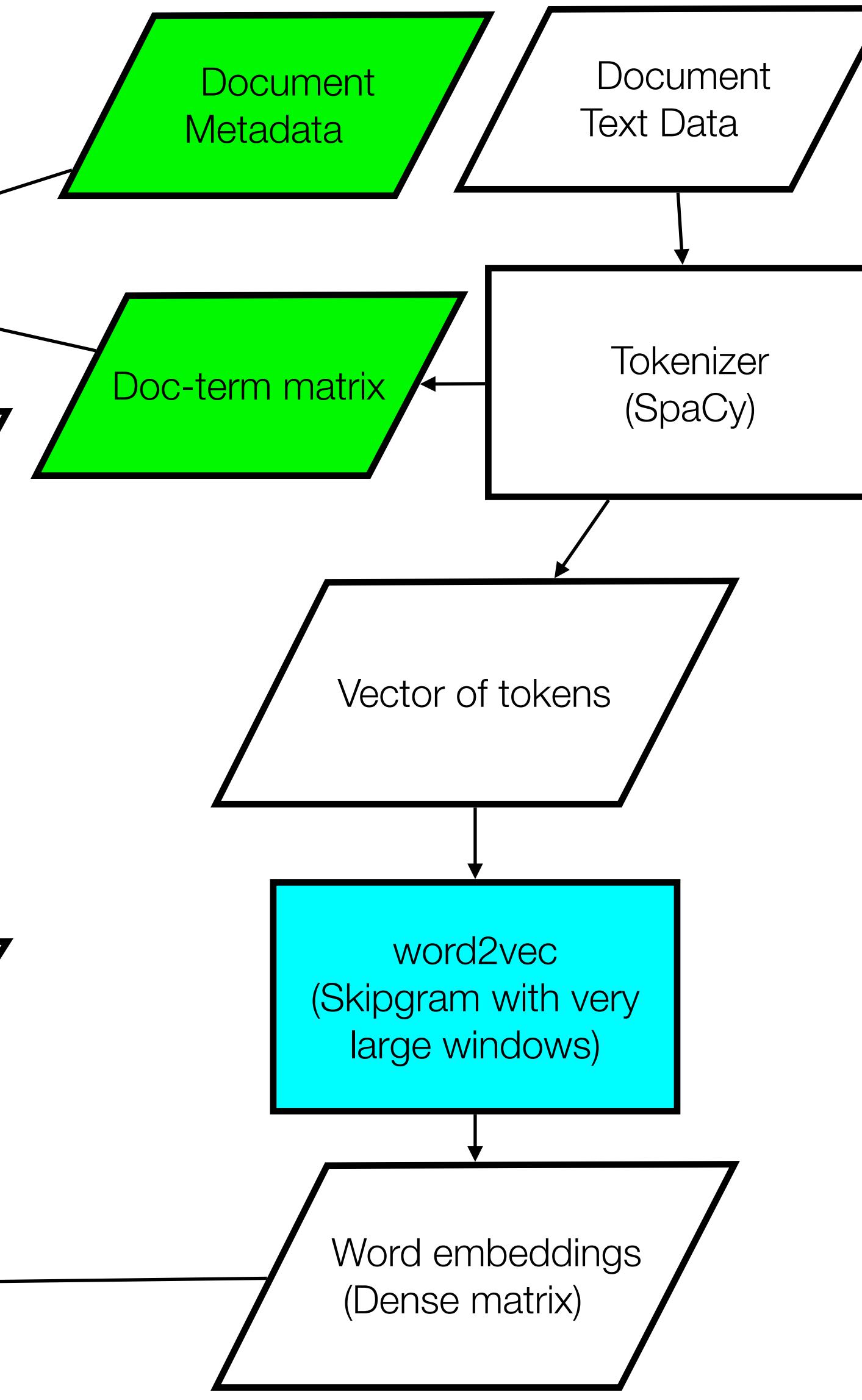
Language 2



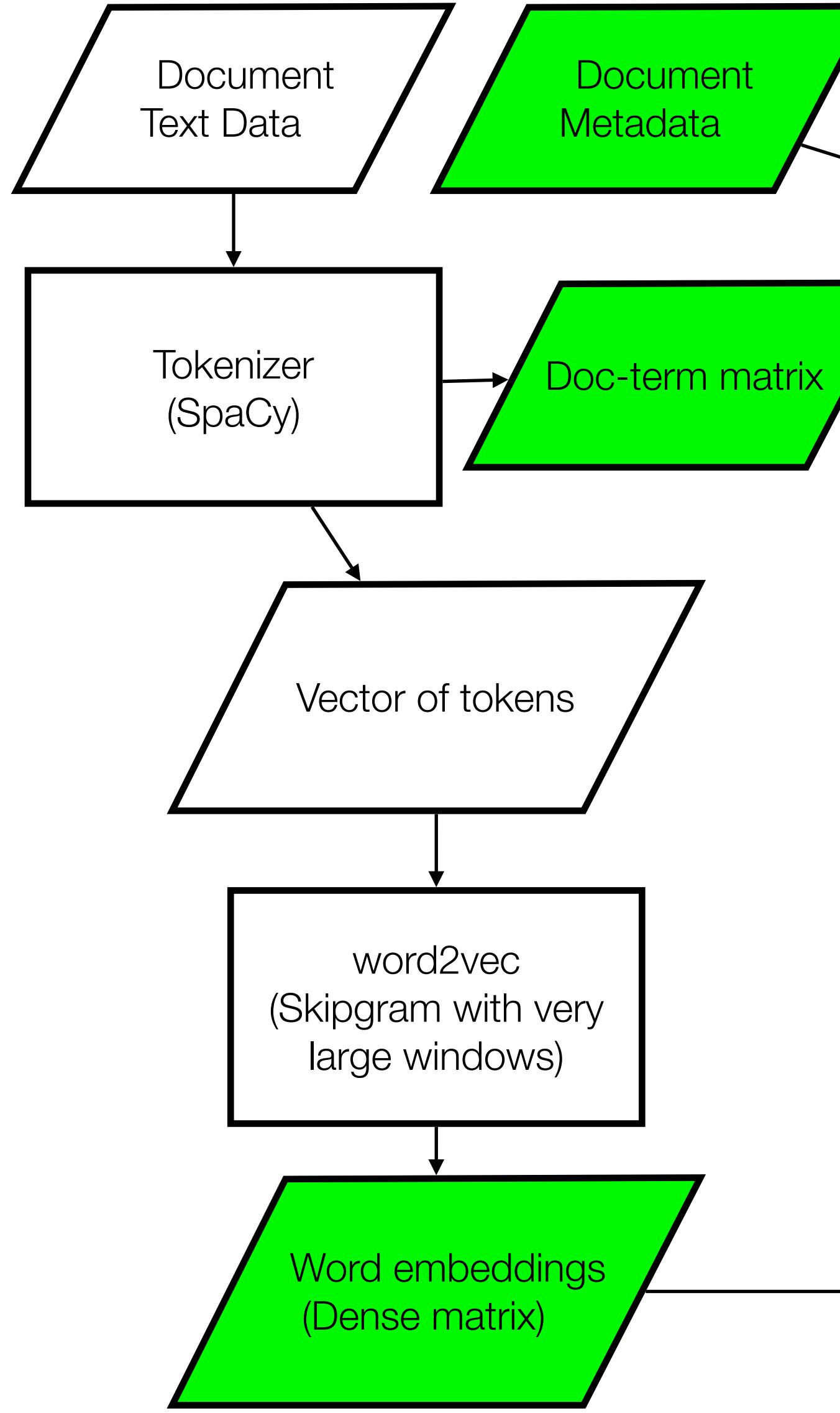
Language 1



Language 2



Language 1



Science ...

Scores
(Docs, Aggregates)

Transformations &
matrix algebra

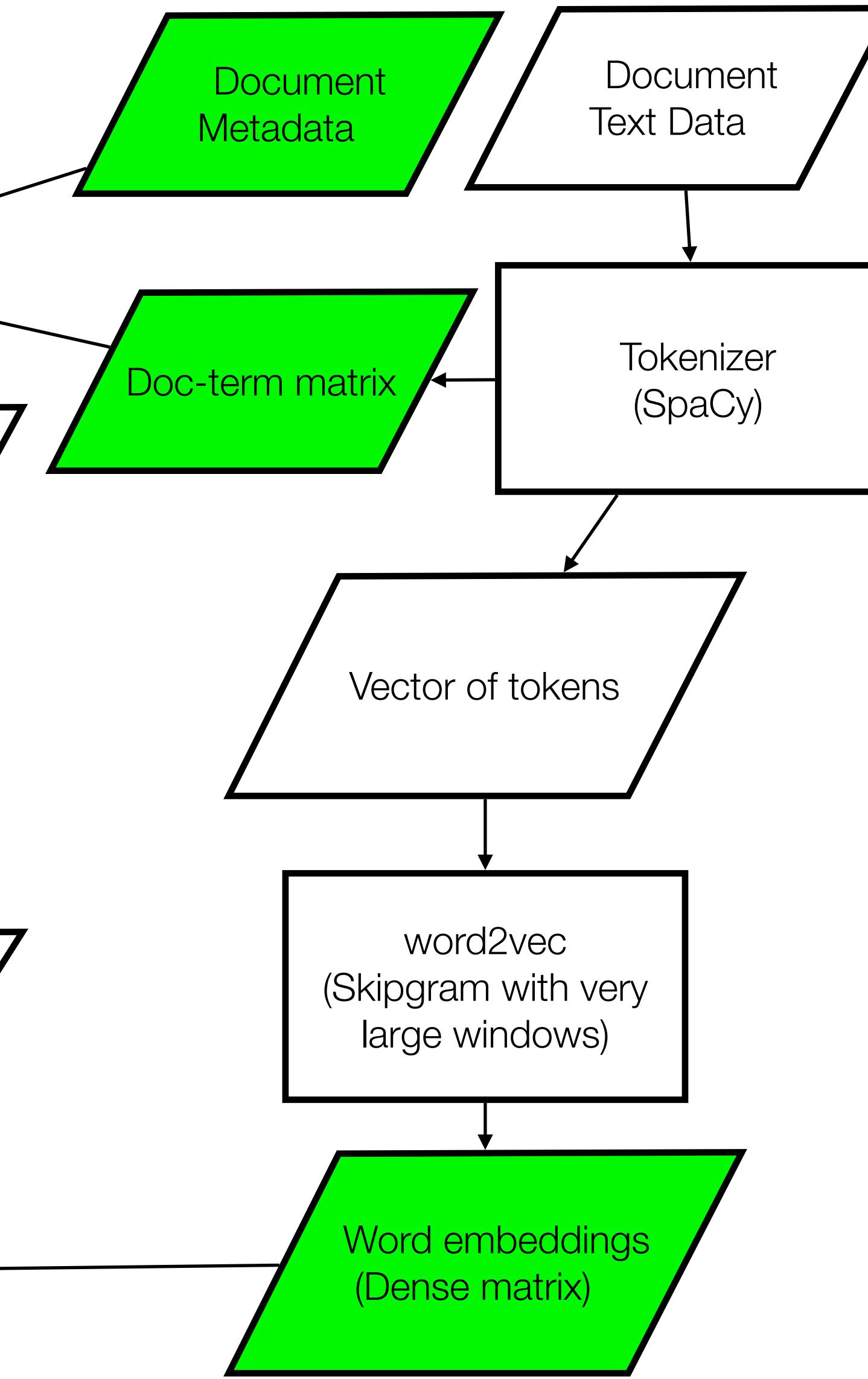
Interpretable
multilingual word
embeddings

Rotated SVD

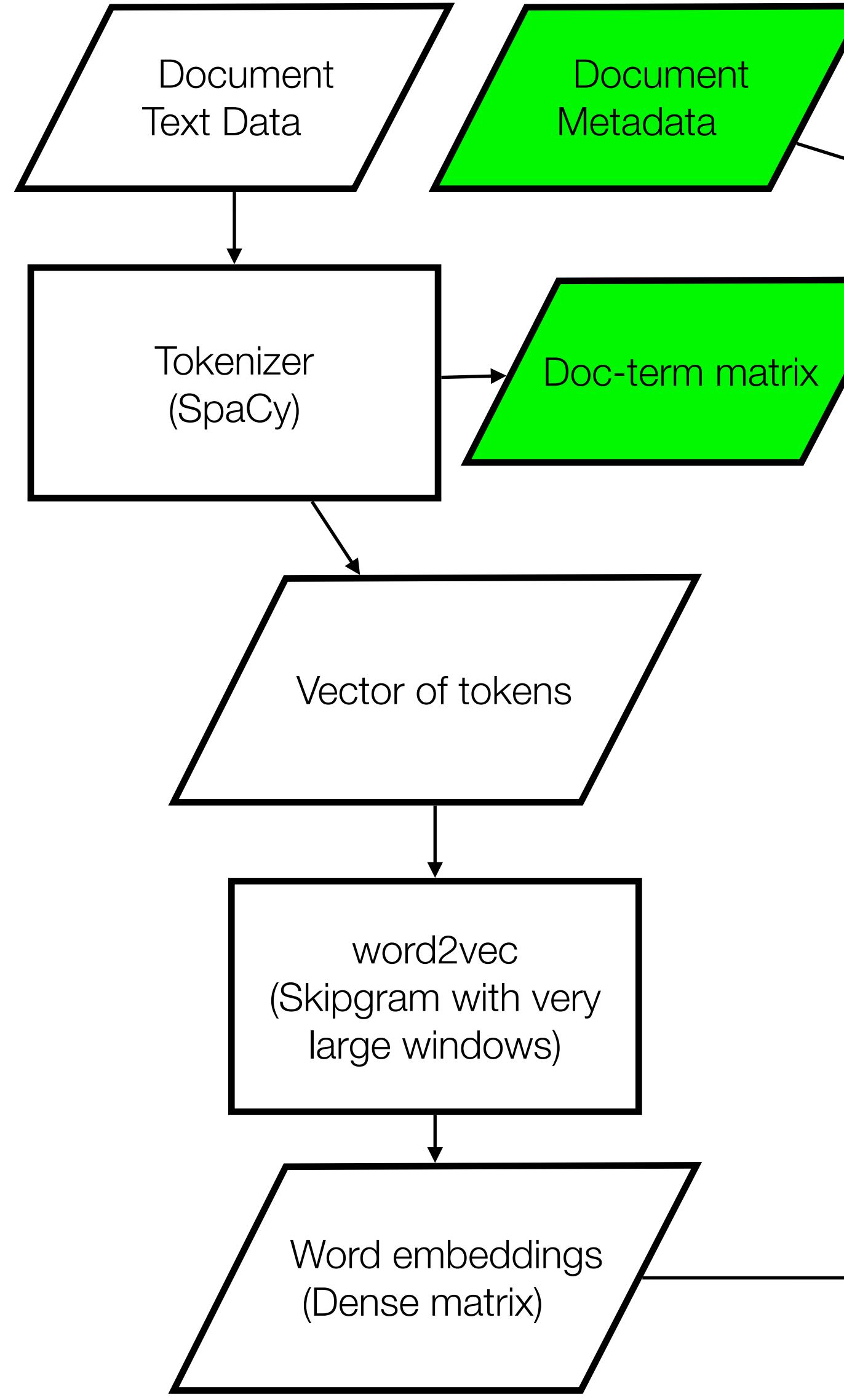
Multilingual word
embeddings

Procrustes alignment
using polysemic
bilingual dictionaries

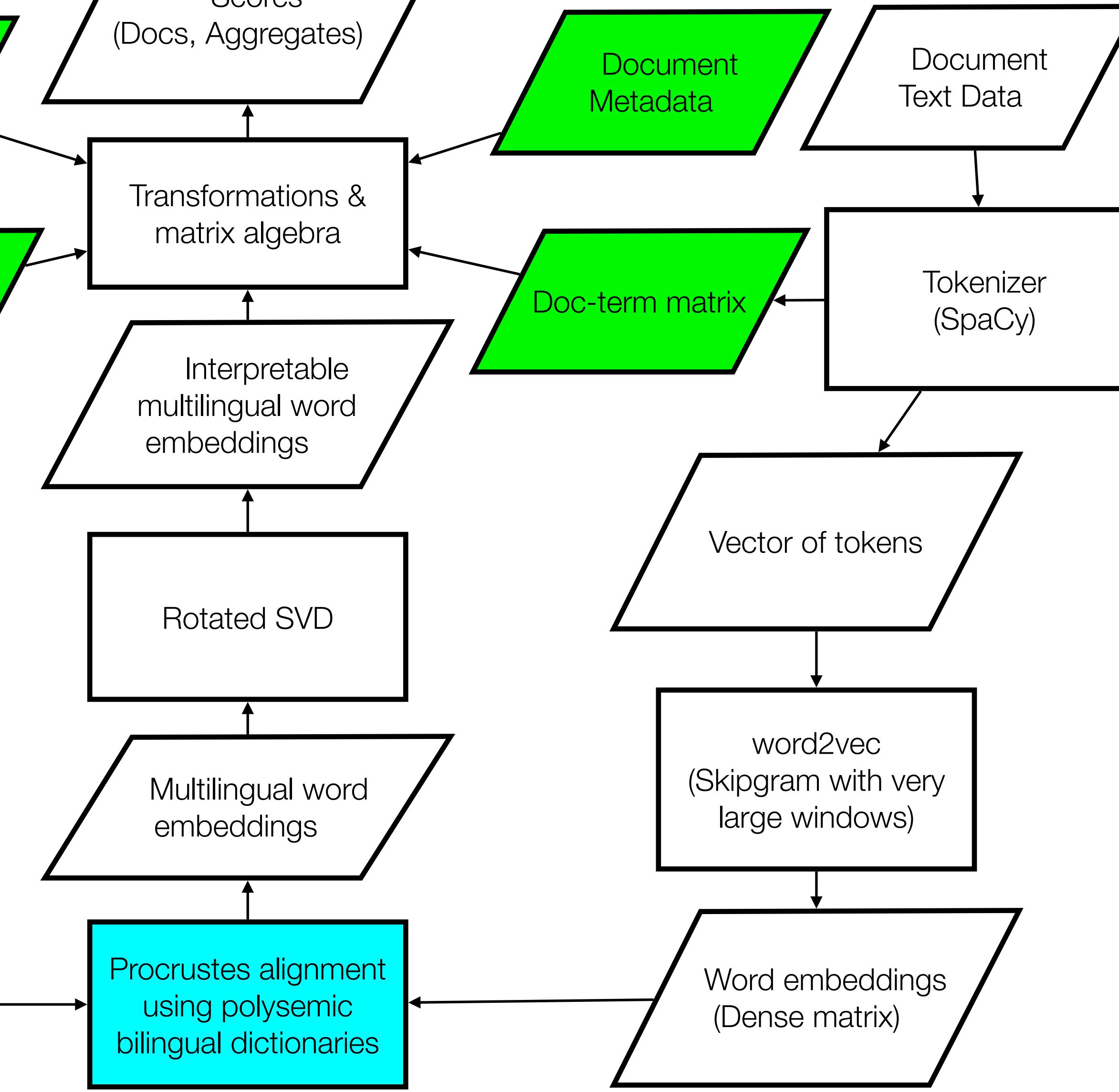
Language 2



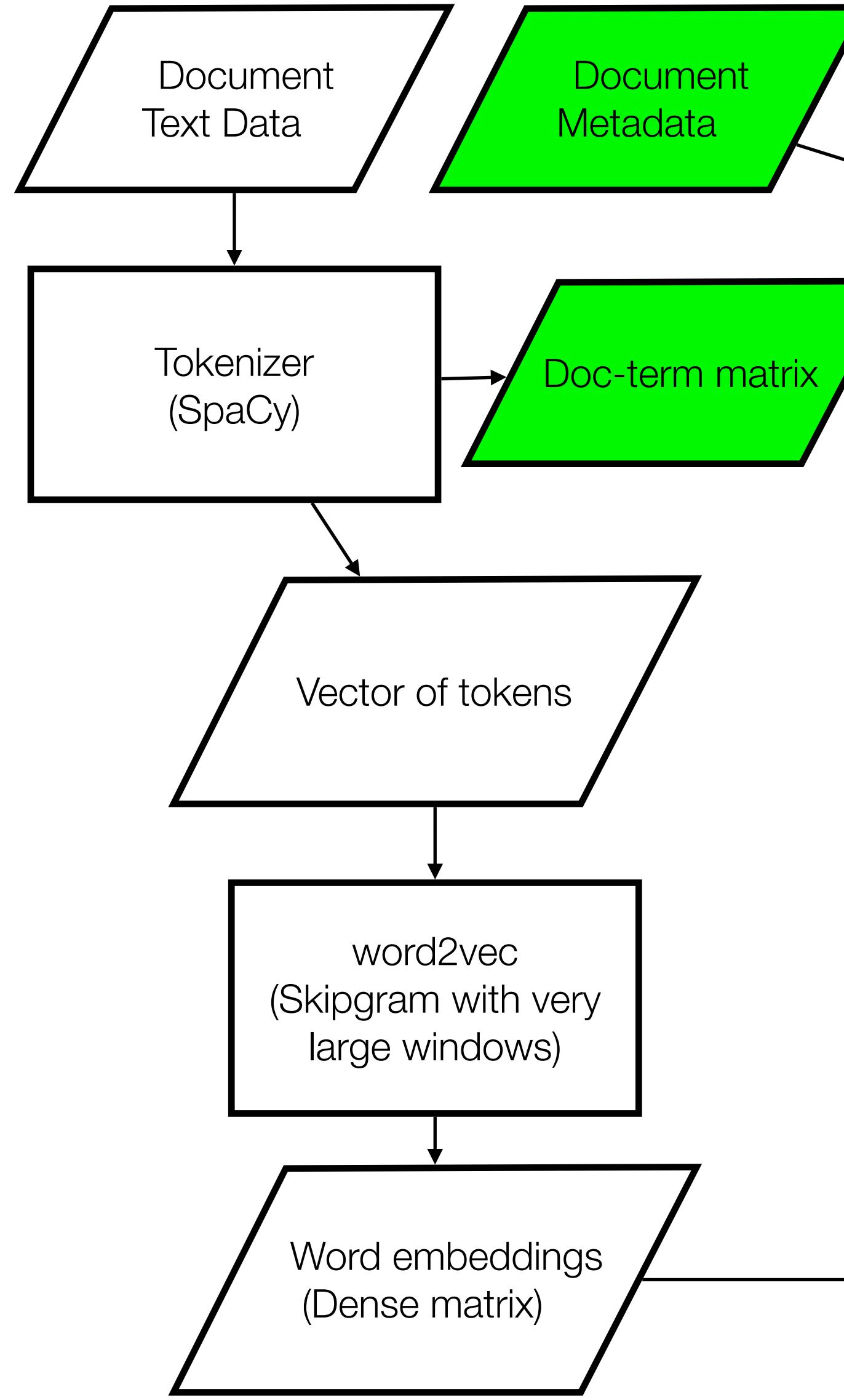
Language 1



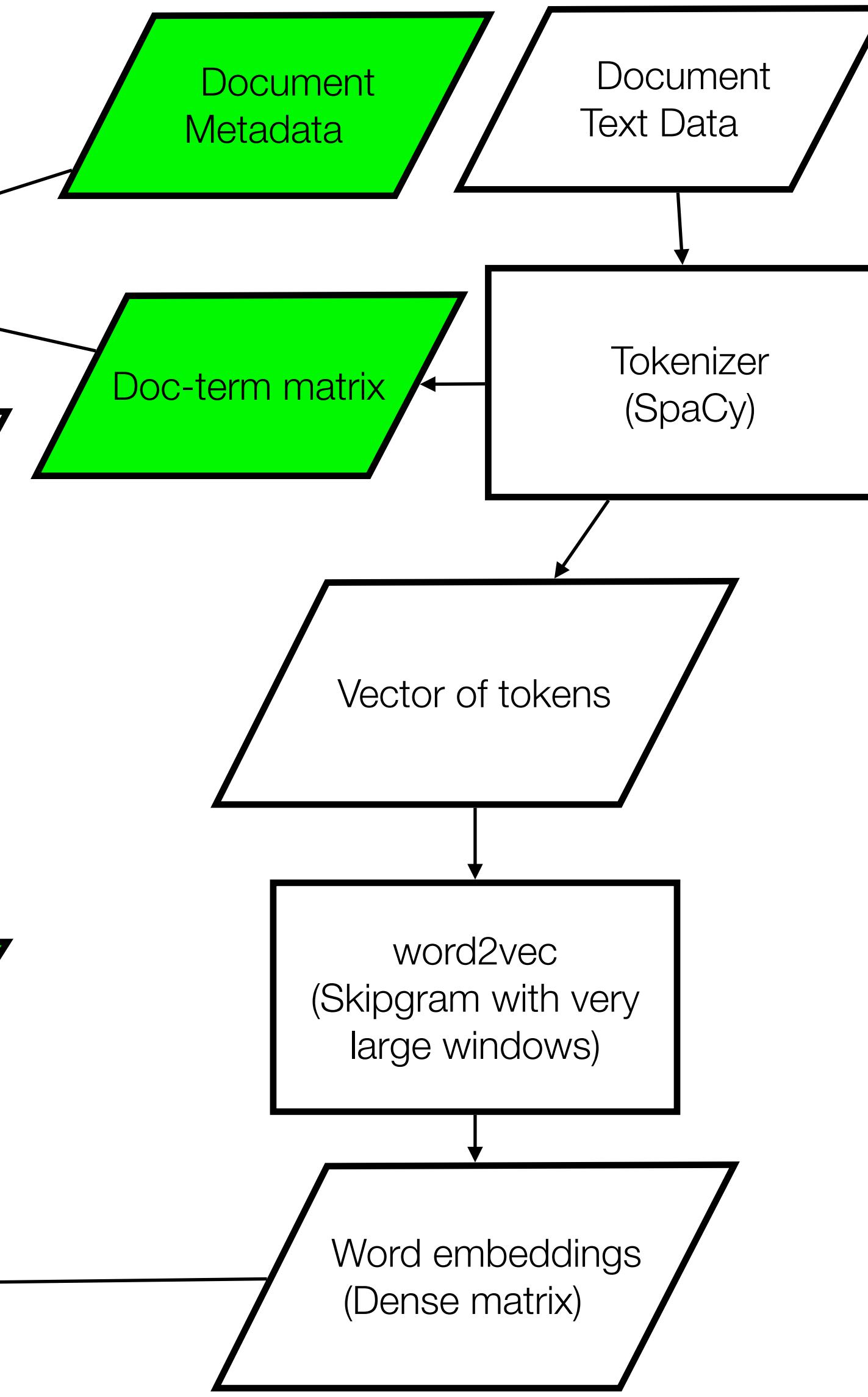
Language 2



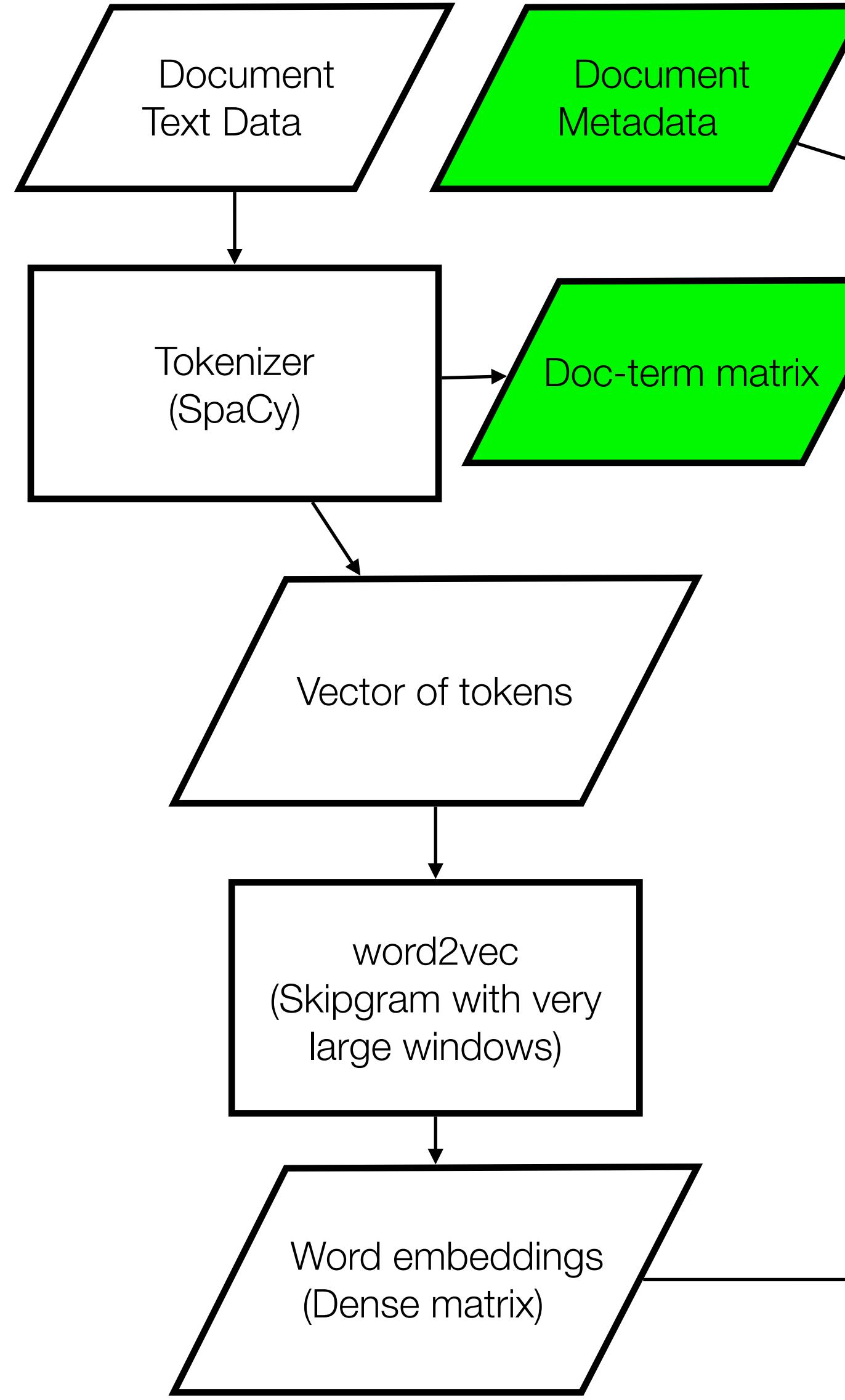
Language 1



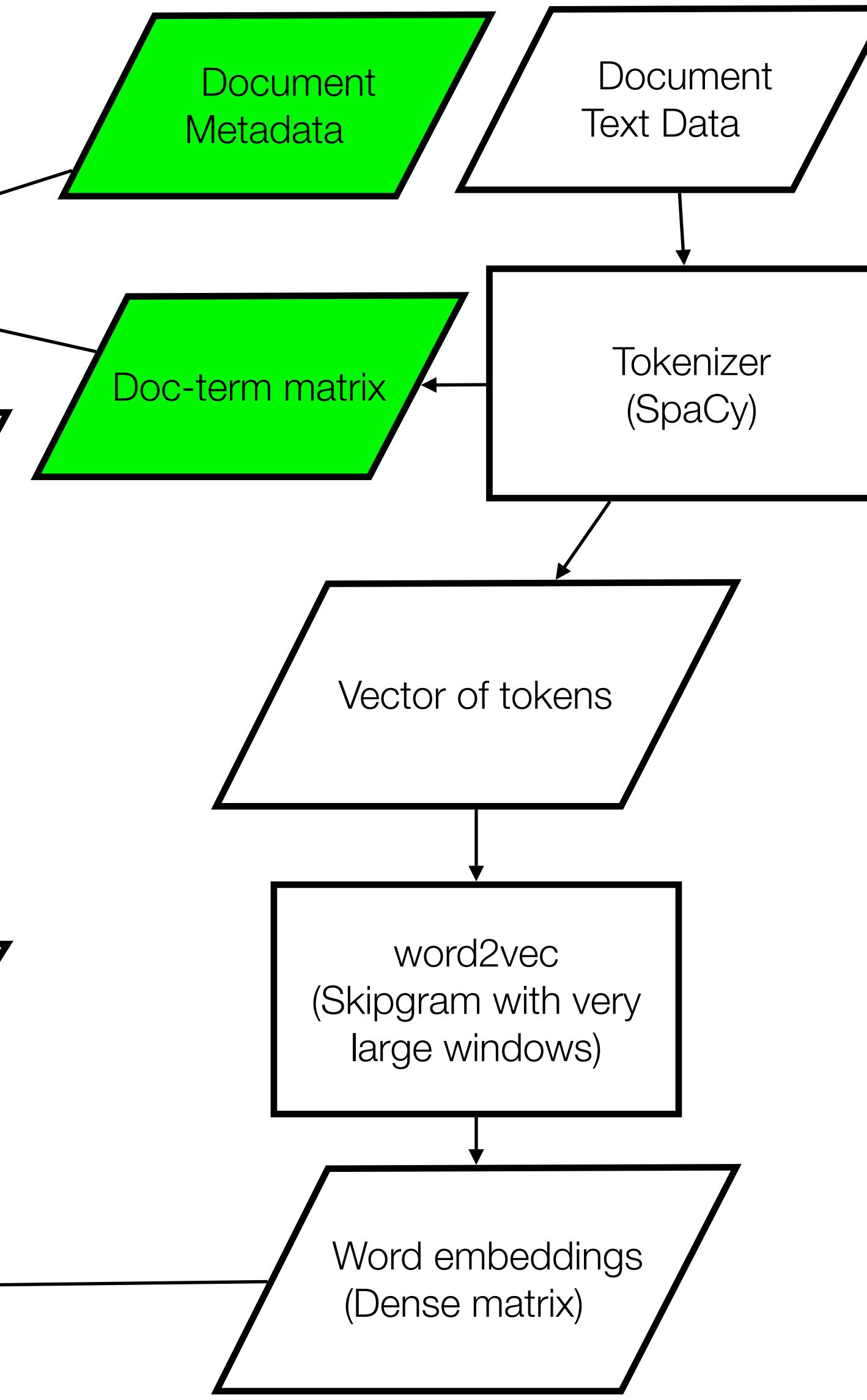
Language 2



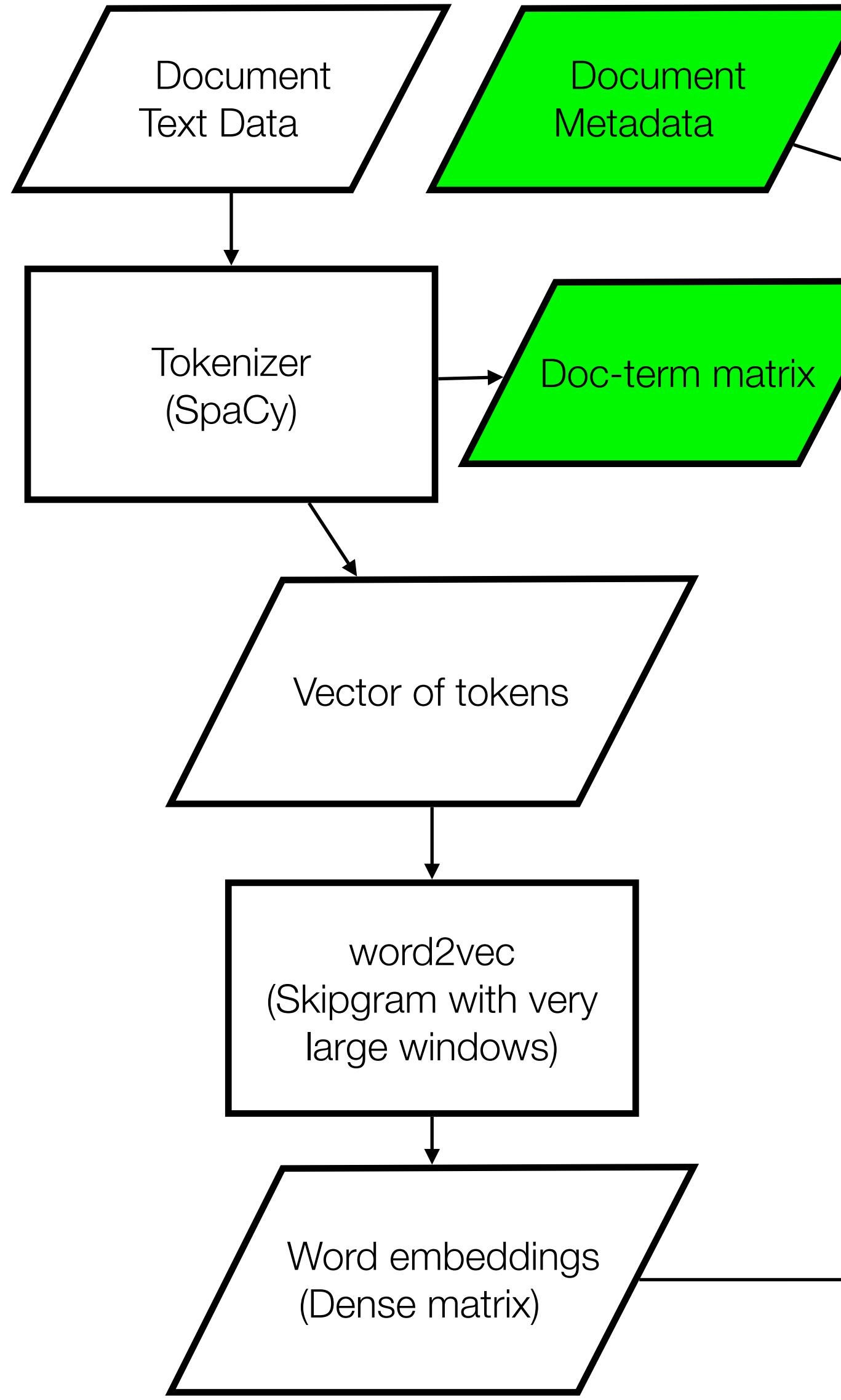
Language 1



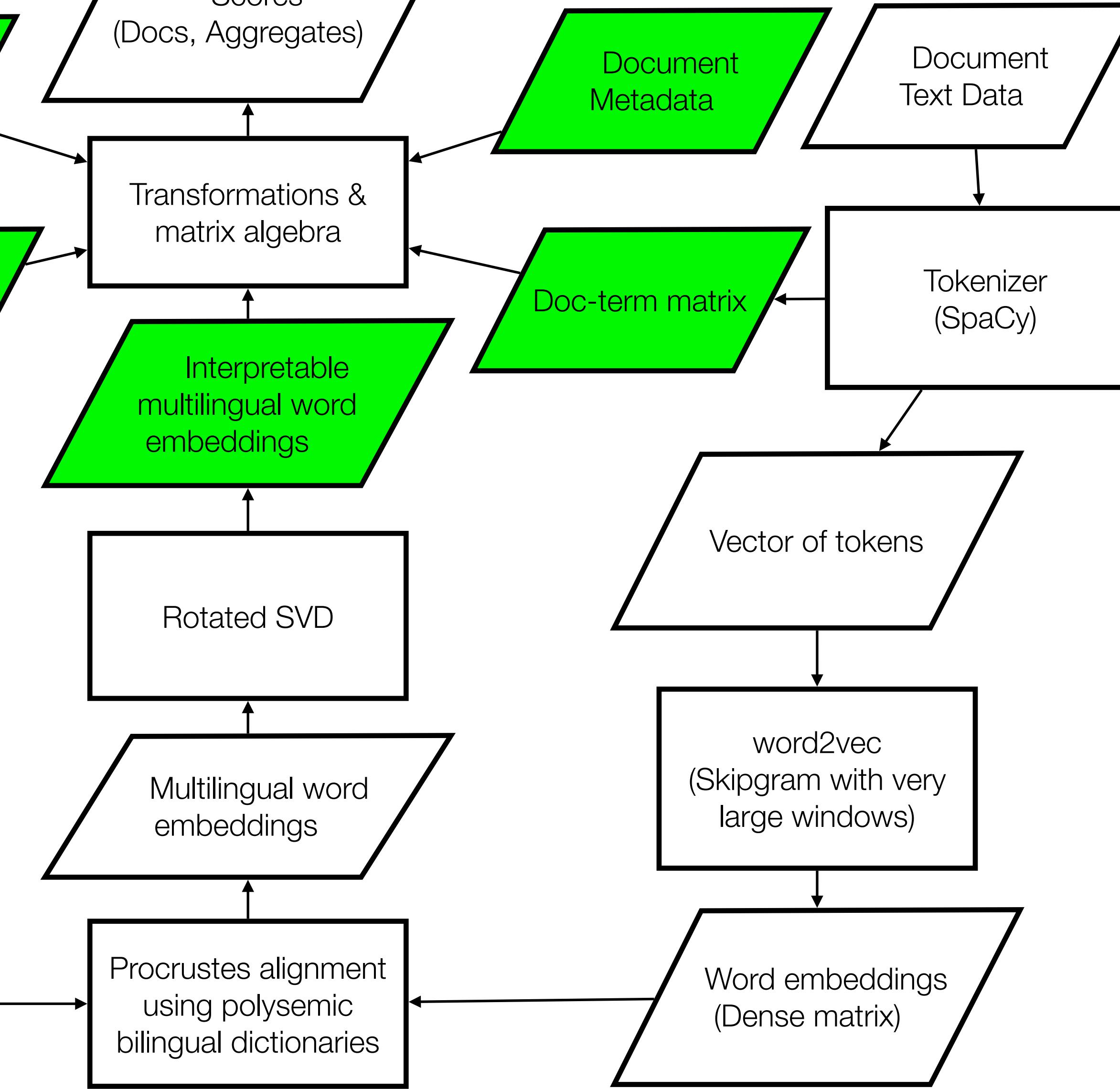
Language 2



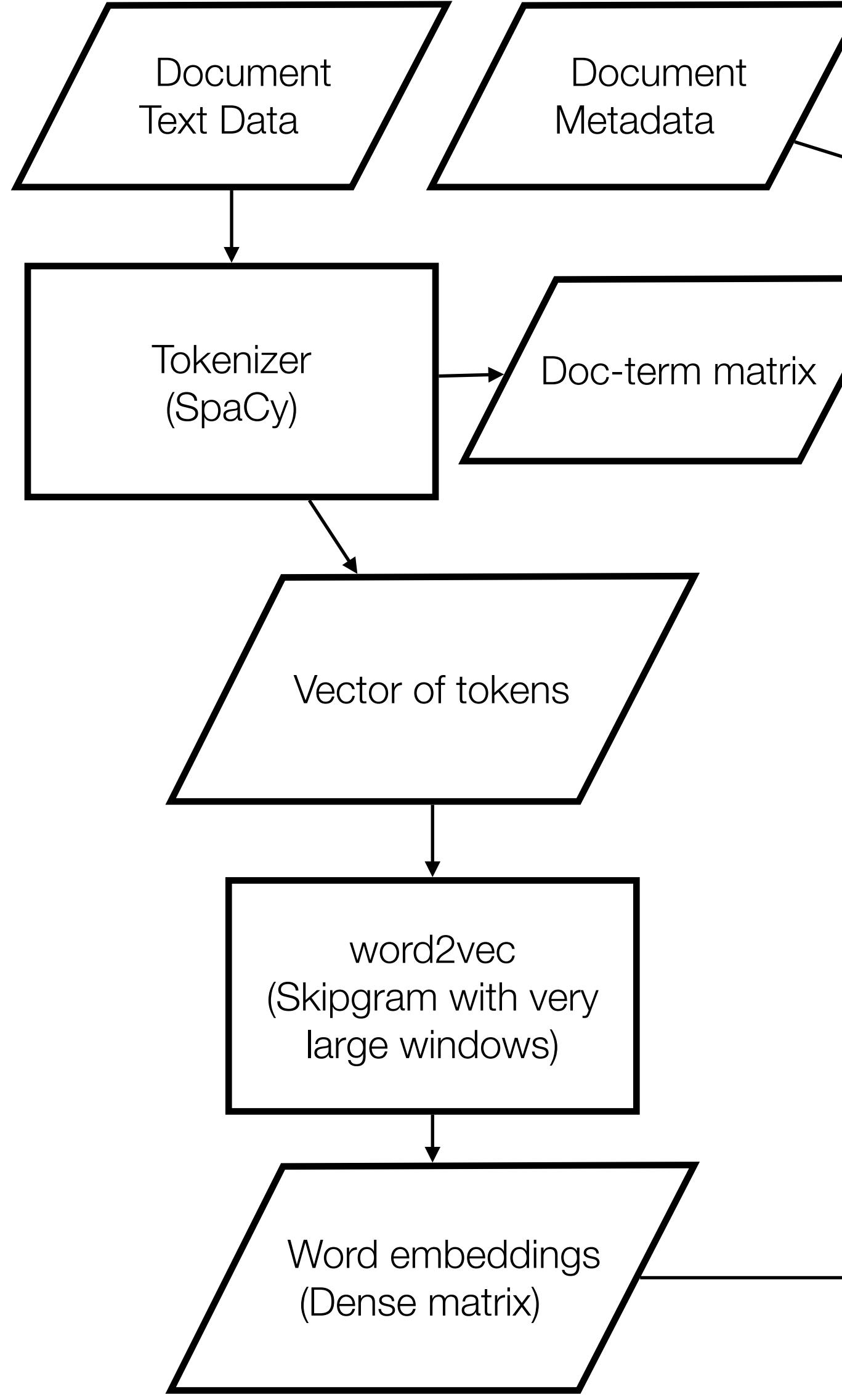
Language 1



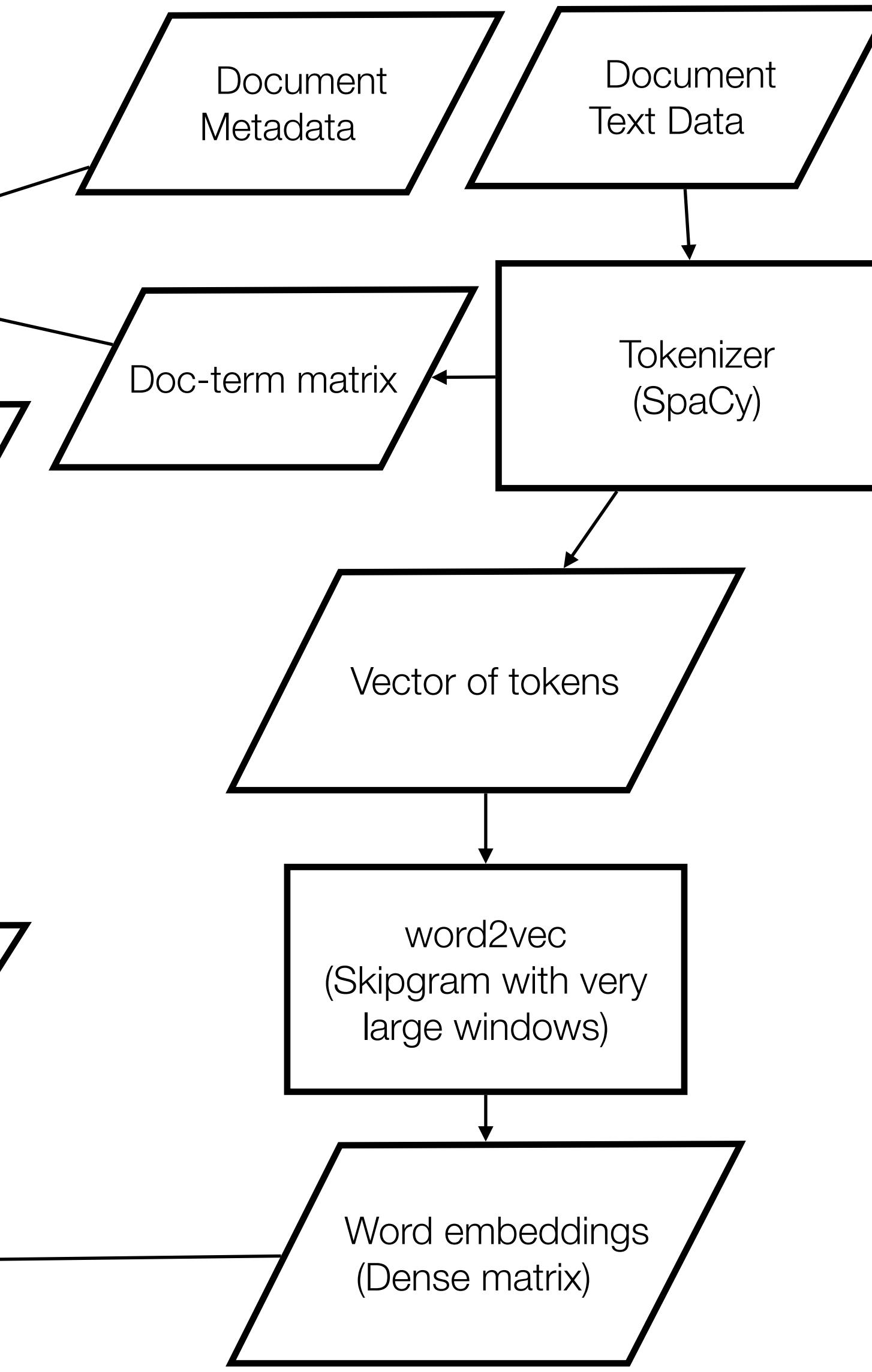
Language 2



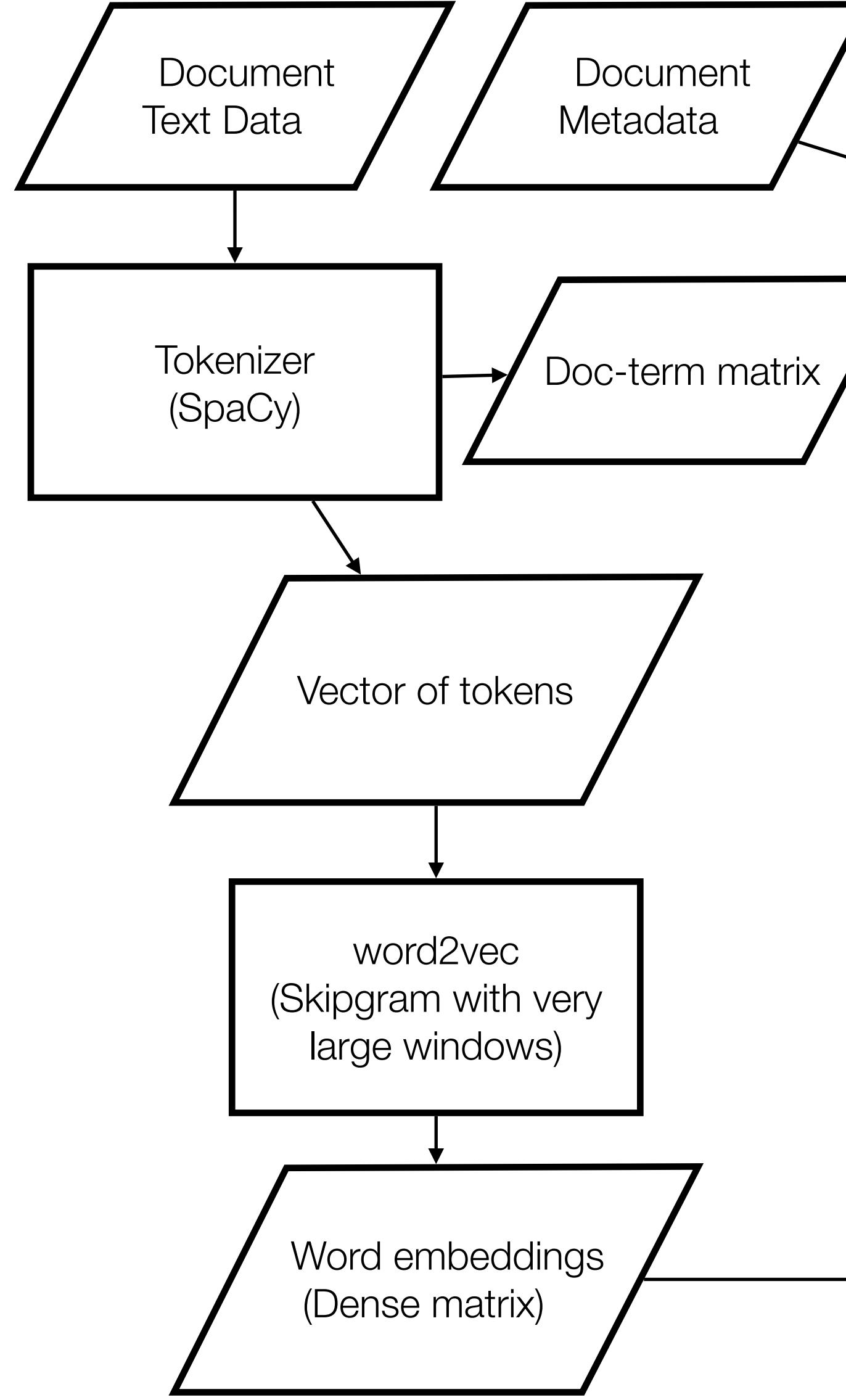
Language 1



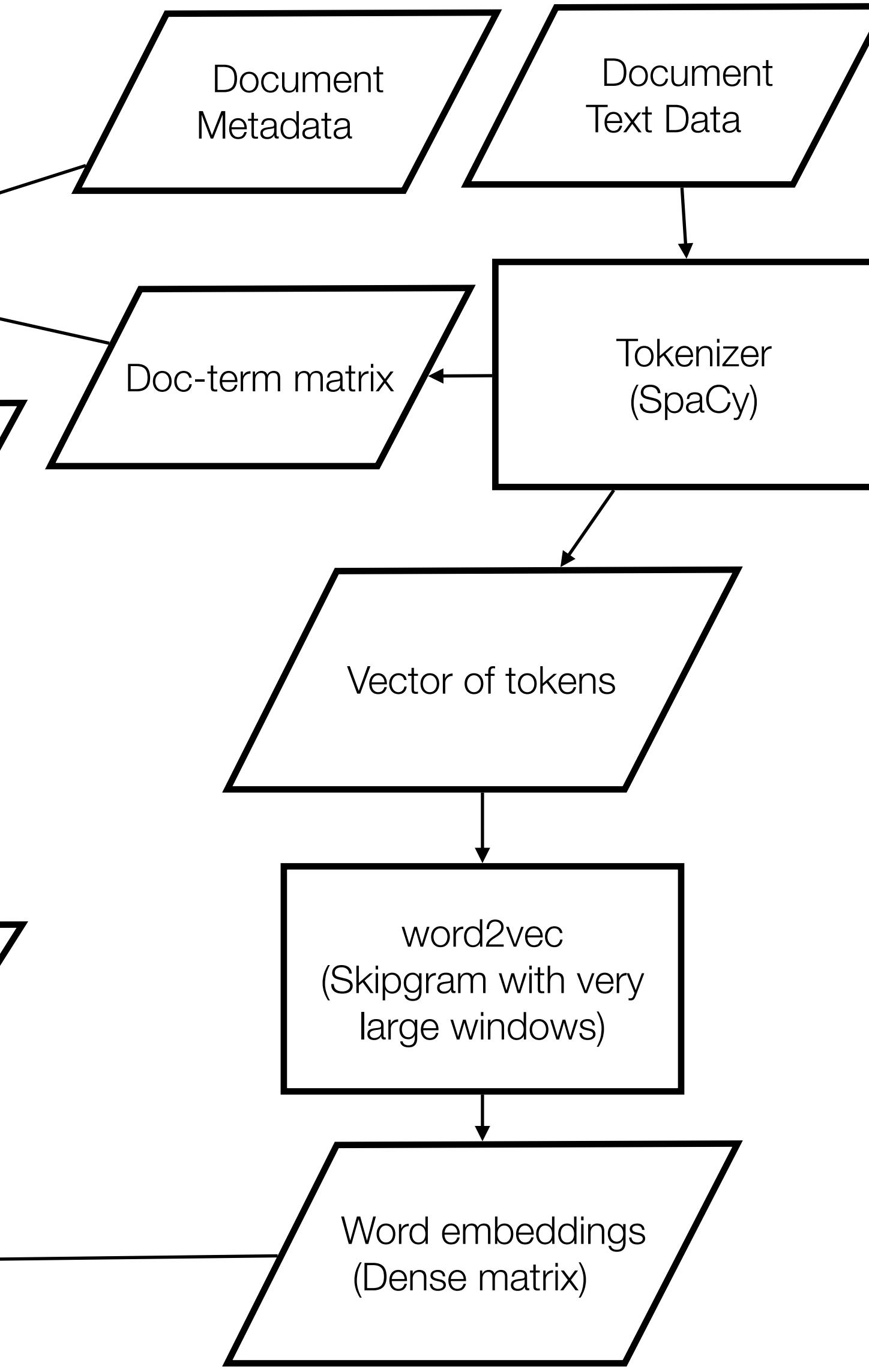
Language 2



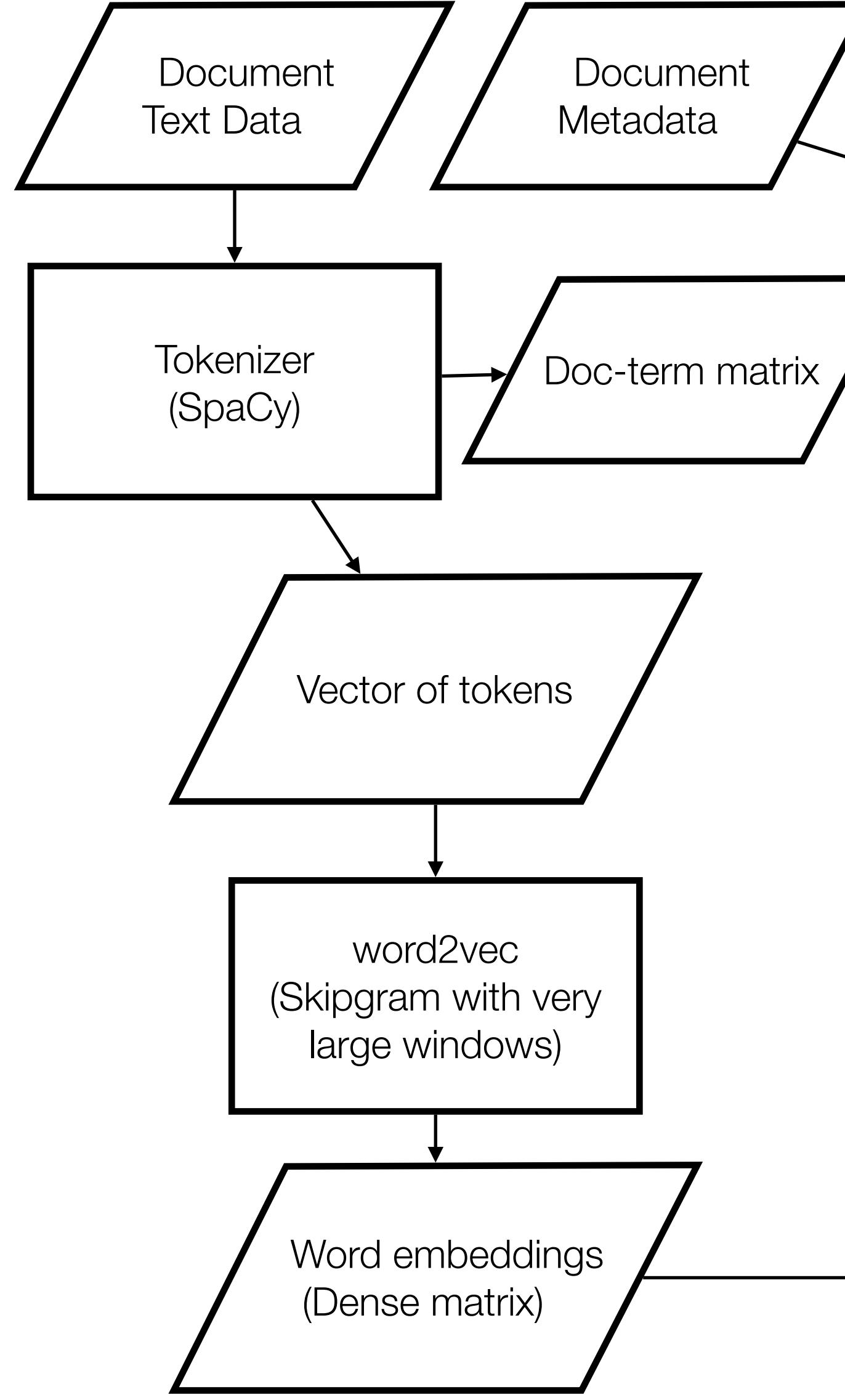
Language 1



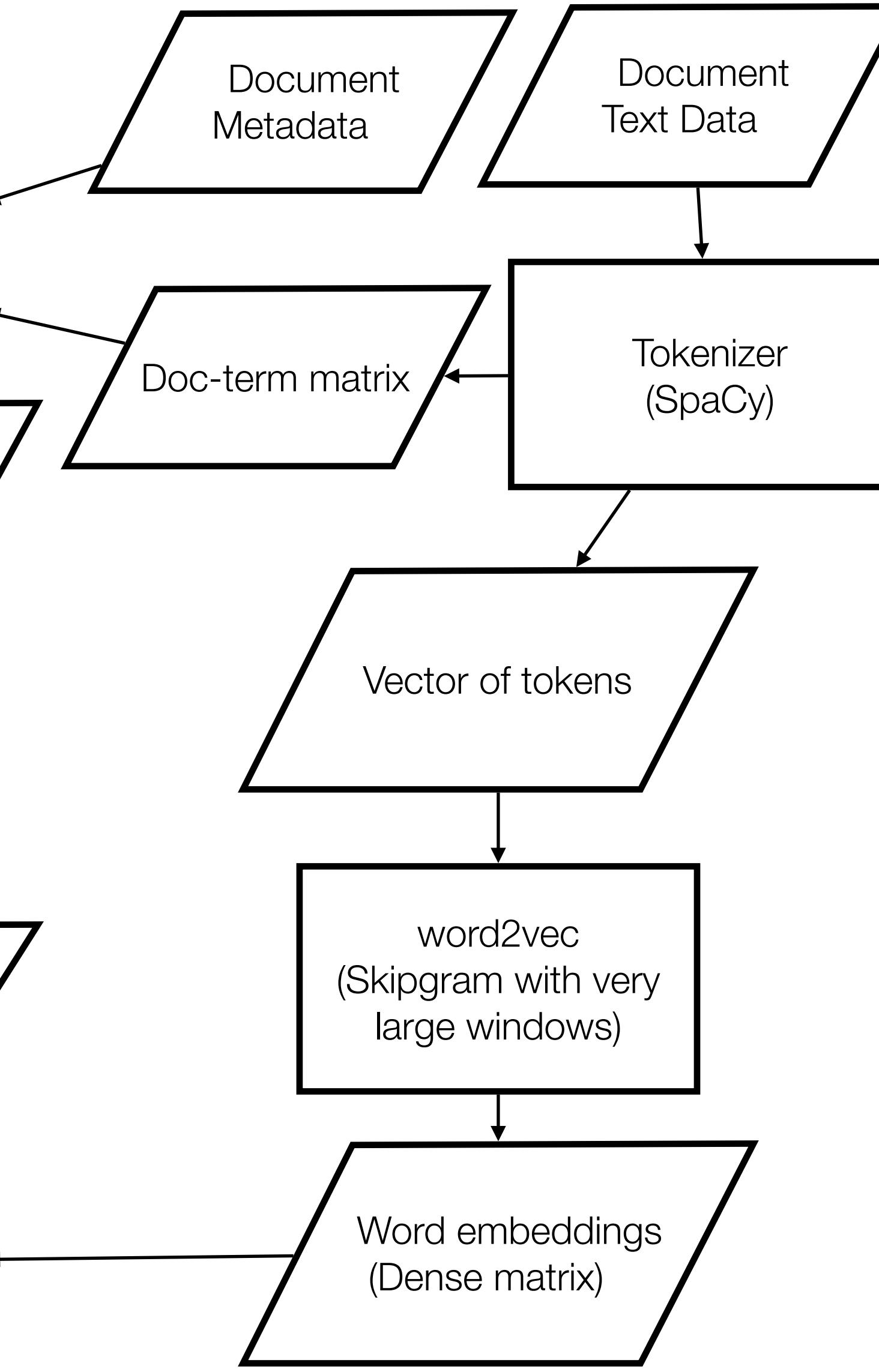
Language 2



Language 1



Language 2



Language 1

Document
Text Data

Tokenizer
(SpaCy)

Vector of tokens

word2vec
(Skipgram with very
large windows)

Word embeddings
(Dense matrix)

Document
Metadata

Doc-term matrix

Science ...

Scores
(Docs, Aggregates)

Transformations &
matrix algebra

Interpretable
multilingual word
embeddings

Rotated SVD

Multilingual word
embeddings

Procrustes alignment
using polysemic
bilingual dictionaries

Language 2

Document
Metadata

Doc-term matrix

Document
Text Data

Tokenizer
(SpaCy)

Vector of tokens

word2vec
(Skipgram with very
large windows)

Word embeddings
(Dense matrix)

Word embeddings

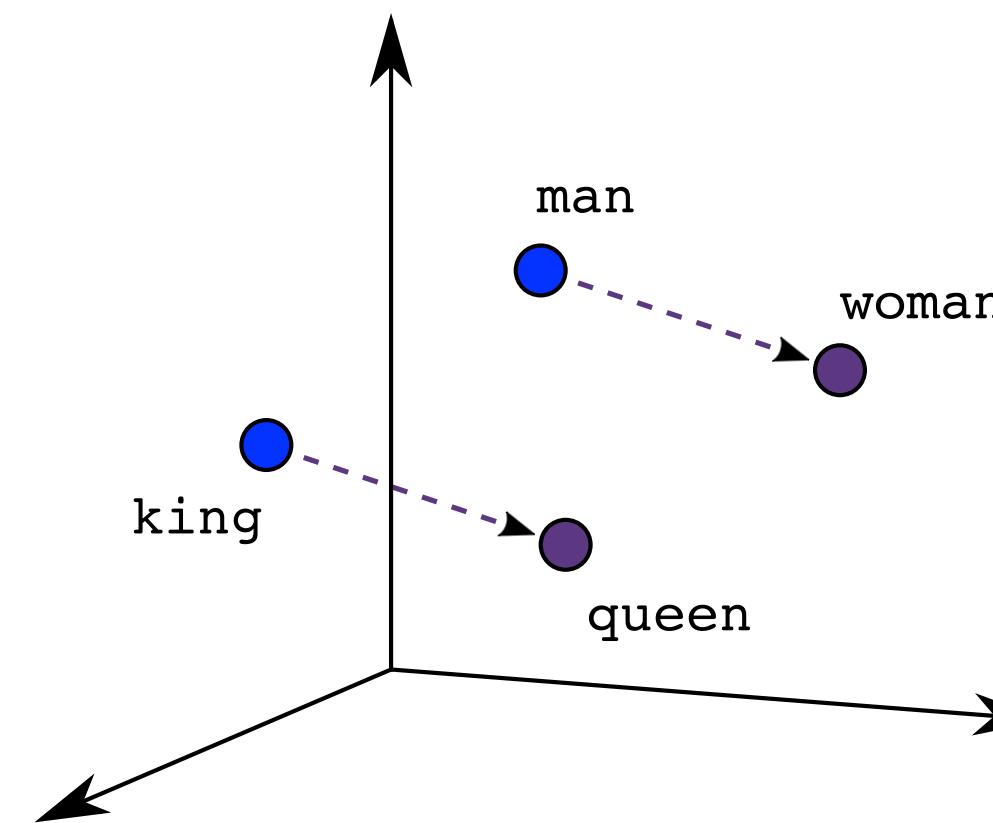
- The old (Mikolov, et al. 2013 – word2vec) new hotness in NLP / text-as-data.
- Represent each word as a vector of real numbers, a location in n -dimensional space.
- Premise: “You shall know a word by the company it keeps.” (Firth 1957)
- Goal: words that appear in the same “contexts” should be near each other, have similar vector representations.
- Context is typically defined as a window of neighboring words.

the quick brown fox jumps over the lazy dog

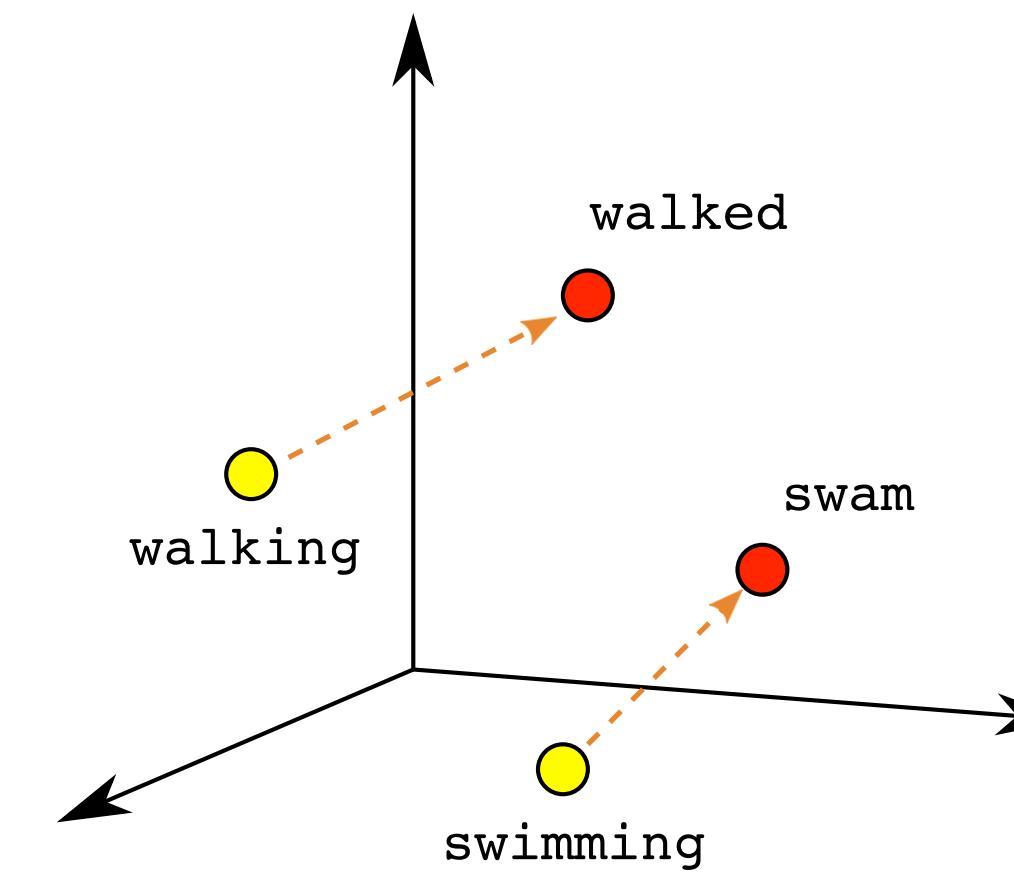
([Context words], focus word) with window size of 1:

([the, brown], quick), ([quick, fox], brown ...

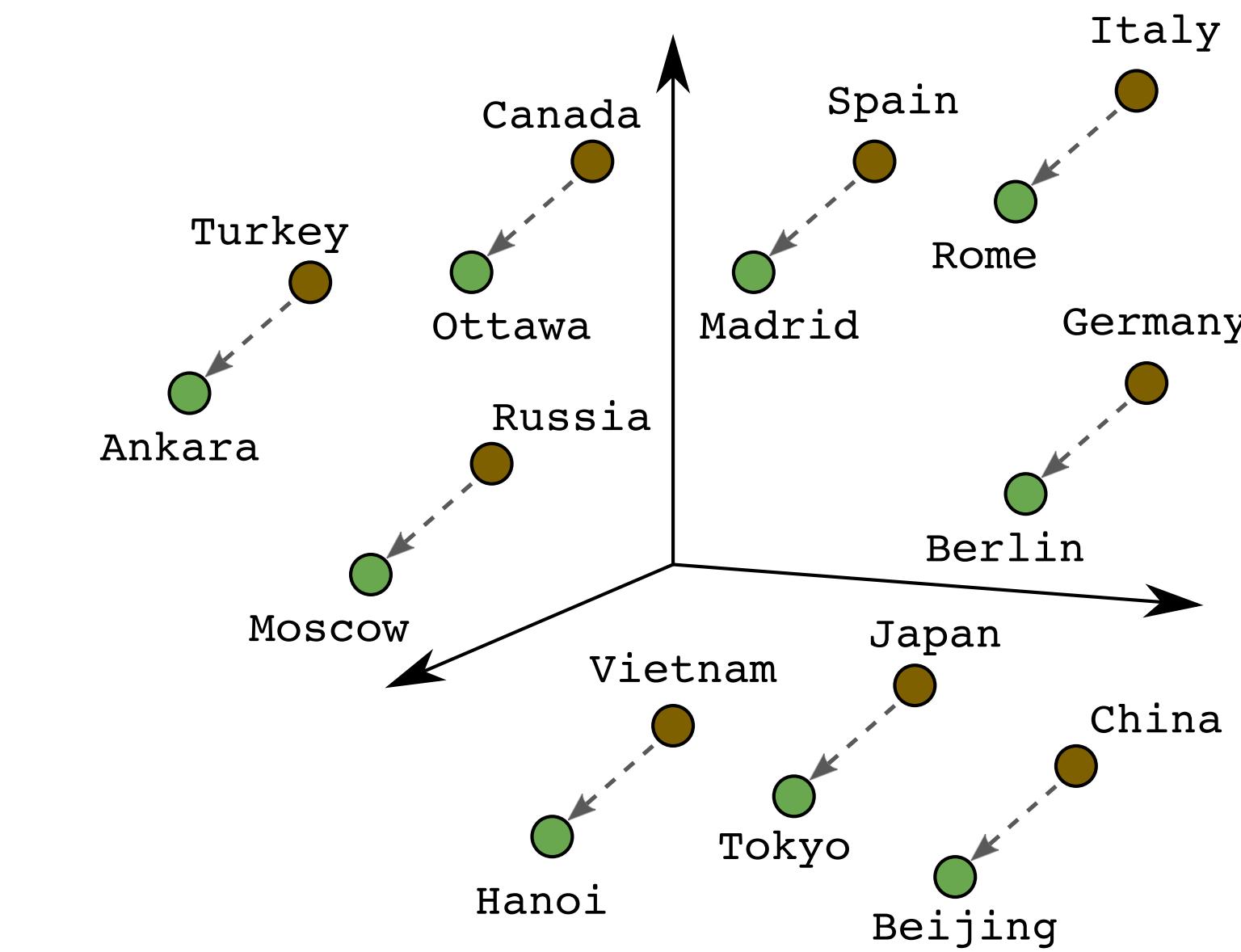
If you know one thing about word embeddings – vector semantics



Male-Female



Verb Tense



Country-Capital

Demonstrated through “Analogy completion tasks” ...

Man is to woman as king is to ??

$$V_{\text{woman}} - V_{\text{man}} + V_{\text{king}} = ?? \cong V_{\text{queen}}$$

Word2vec approximates a matrix decomposition

- Levy and Goldberg (2014) argued that word2vec is implicitly calculating a weighted singular value decomposition (SVD) of shifted/regularized pointwise mutual information (PMI) between word and context.

$$\text{pmi}(x; y) \equiv \log \frac{p(x, y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)}$$

- (Their proof has been shown to rest on a very strong assumption, but the basic intuition still seems to roughly hold.)
- Similarly, Pennington, et al. (2014) introduced GloVe embeddings, which are calculated directly as a weighted decomposition of the matrix of logged word cofrequencies in similarly defined contexts.
- Implication: If you get the features right, meaning may be embedded linearly.

The effect of window size on word embeddings

The size of the window used to collect counts can vary based on the goals of the representation, but is generally between 1 and 8 words on each side of the target word (for a total context of 3-17 words). In general, the shorter the window, the more syntactic the representations, since the information is coming from immediately nearby words, **the longer the window, the more semantic the relations.**

Jurafsky & Martin, *Speech and Language Processing*, 3rd edition (2020 draft)

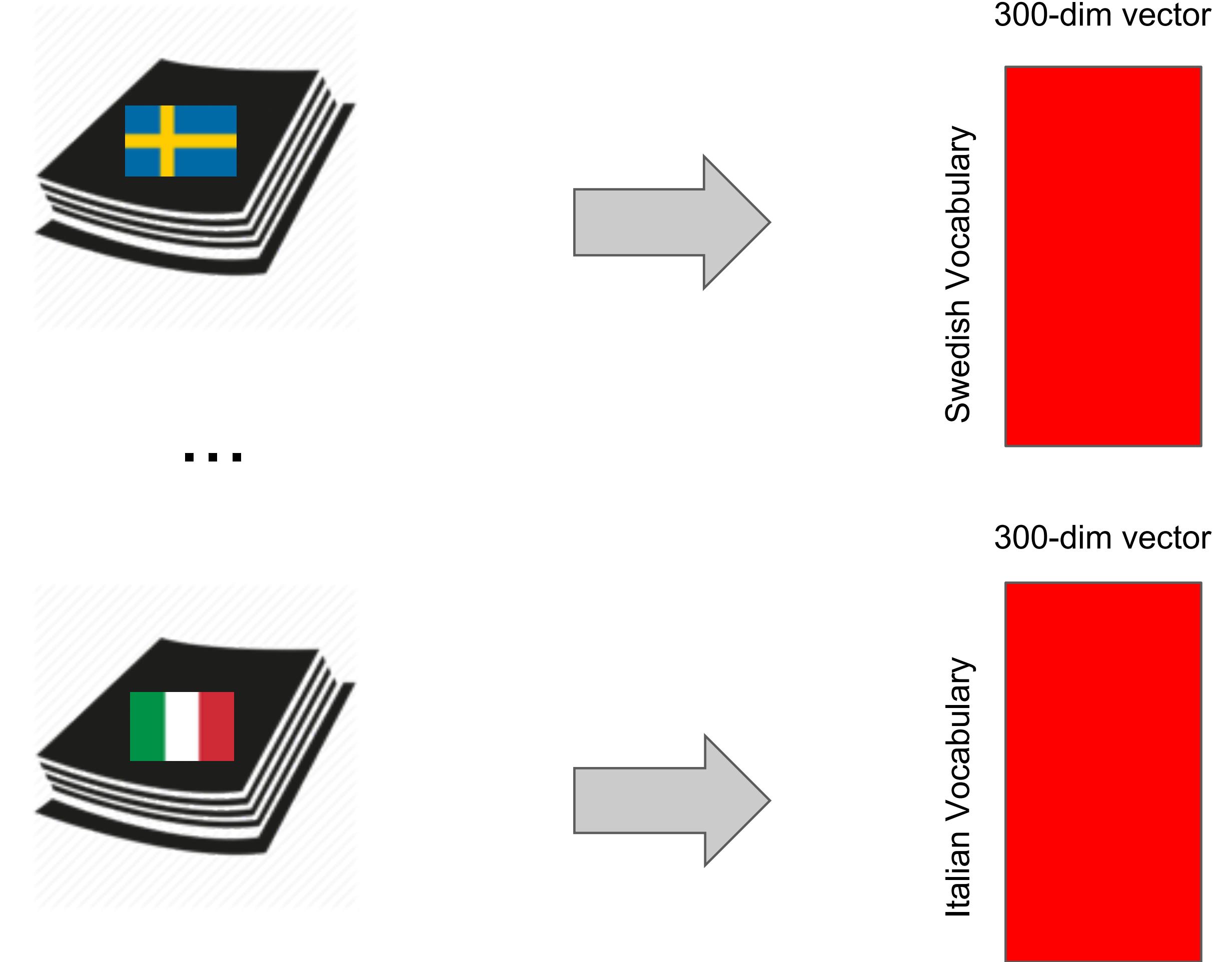
We extrapolate this wildly to capture semantics that are more topic-like, using a context window of **300** (that stops at document boundaries)

The effect of window size on word embeddings

	window size: 3		window size: 30		window size: 300
putting	able	effective	crimes	diseases	tourism
bringing	can	deliver	offences	disease	visitors
taking	unable	improve	offence	vaccine	museums
giving	trying	strengthen	prosecutions	obesity	tourist
introducing	prepared	improving	murder	screening	citizenship
providing	will	efficient	criminals	diabetes	Media
looking	willing	sustainable	crime	pregnancy	holiday
making	wants	develop	arrested	HIV	holidays
talking	want	delivering	cases	medical	music
publishing	happy	ensuring	prosecution	babies	Olympics

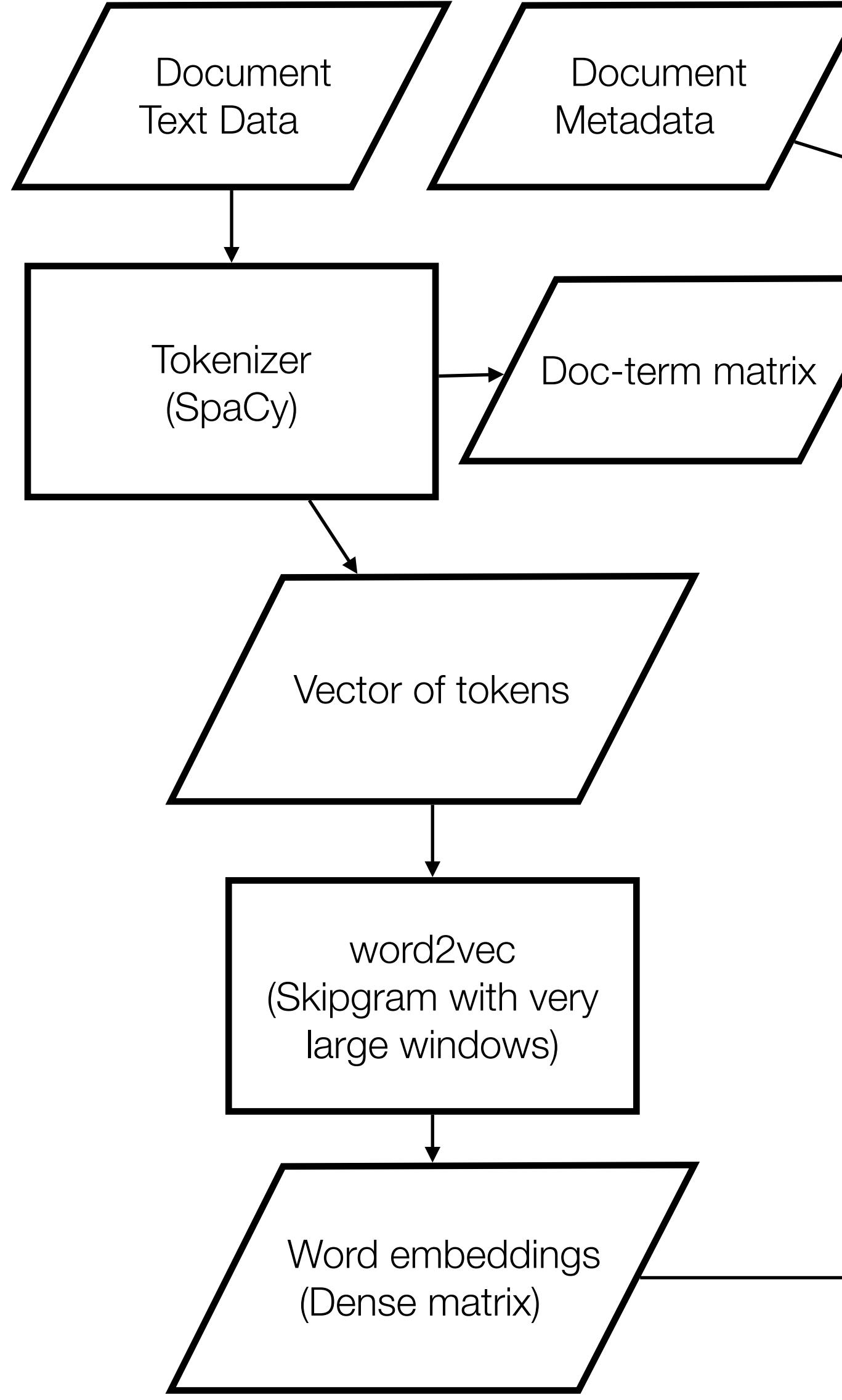
Examples of words that cluster near one another in word2vec (skip-gram) embeddings estimated on the House of Commons corpus with different windows.

Monolingual word embeddings, 300 dimensions

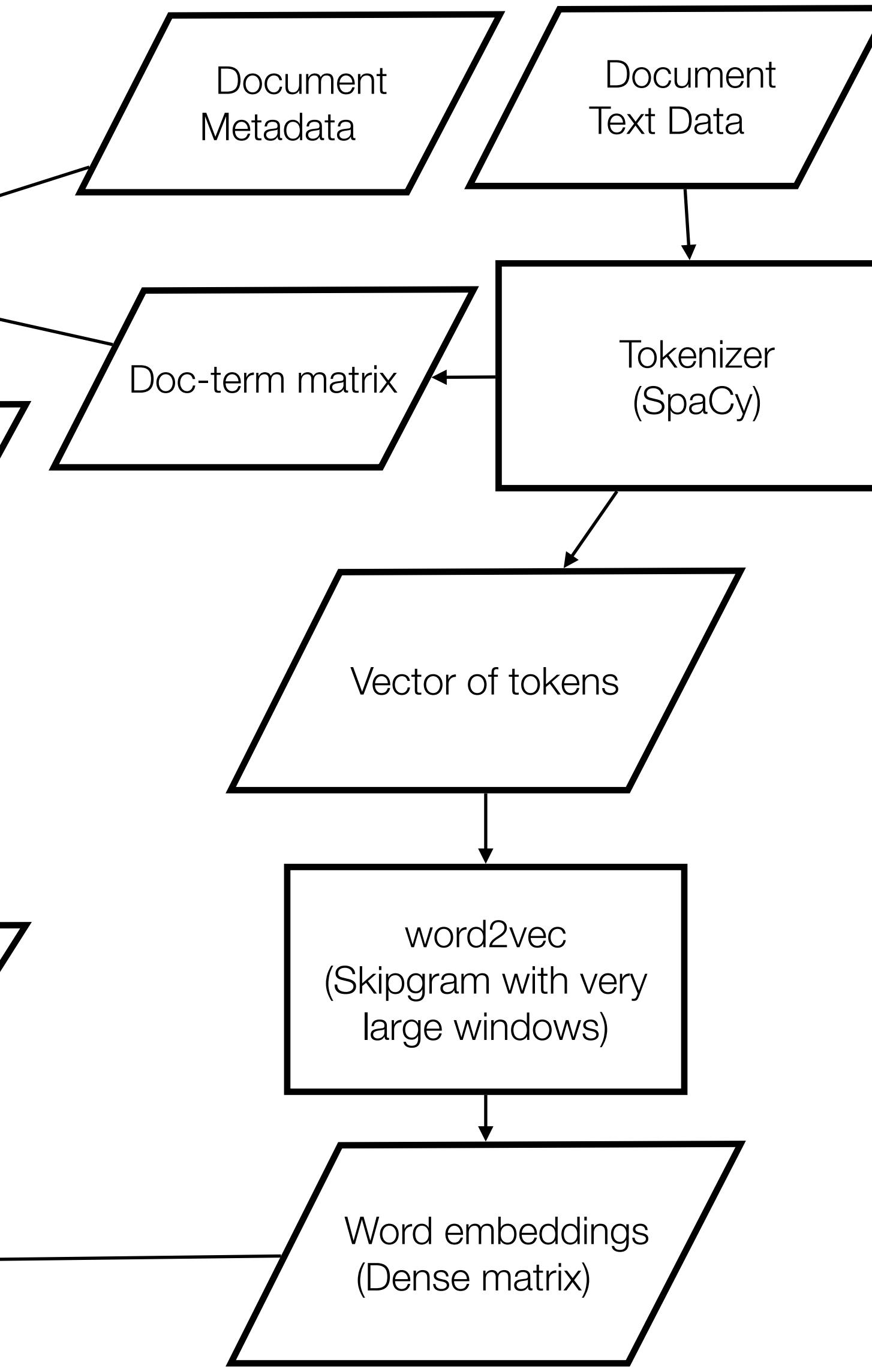


Word2vec skipgram implementation in Gensim - CPU intensive - About an hour/language.

Language 1



Language 2



Prokrustes.

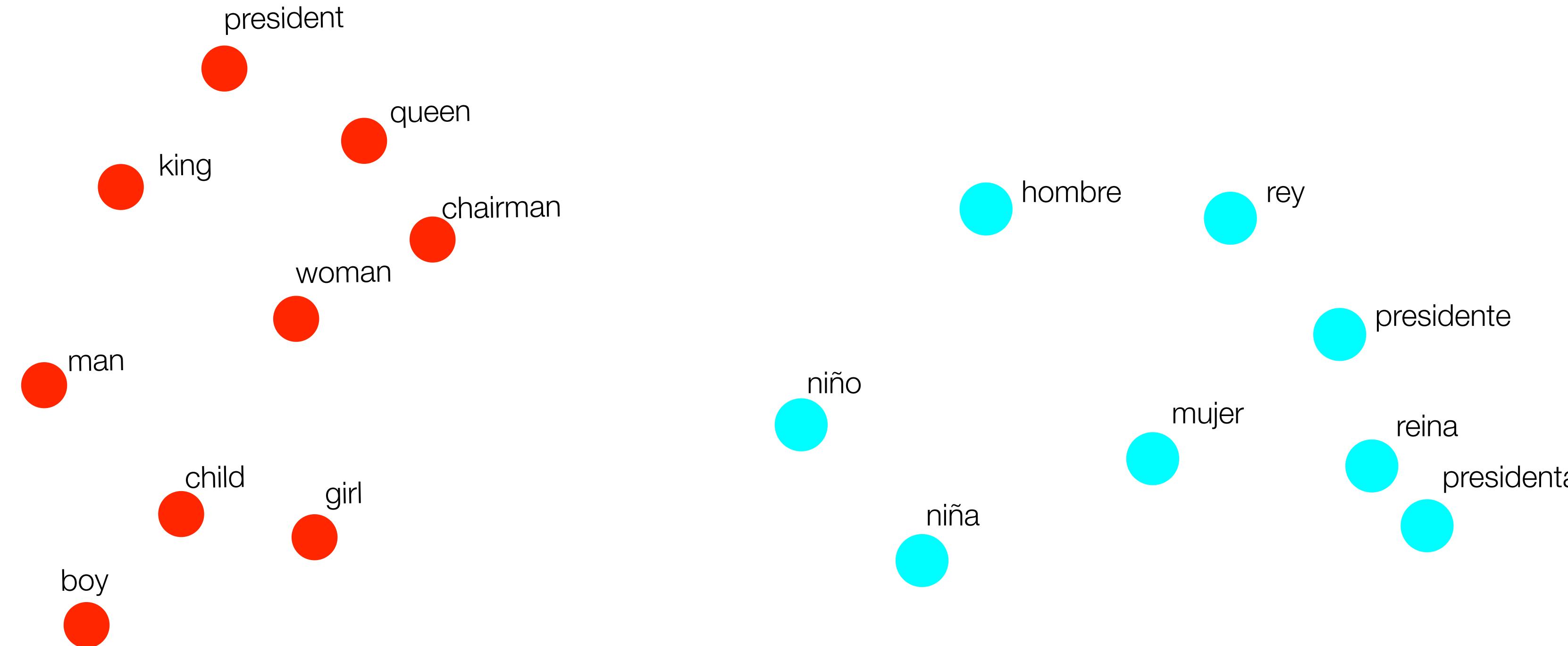


„Wie ich sehe, ist die Freiheit etwas zu groß, — daß wollen wir gleich zu ihrer Zufriedenheit abändern!“ (Er hält ihr die Beine ab.)

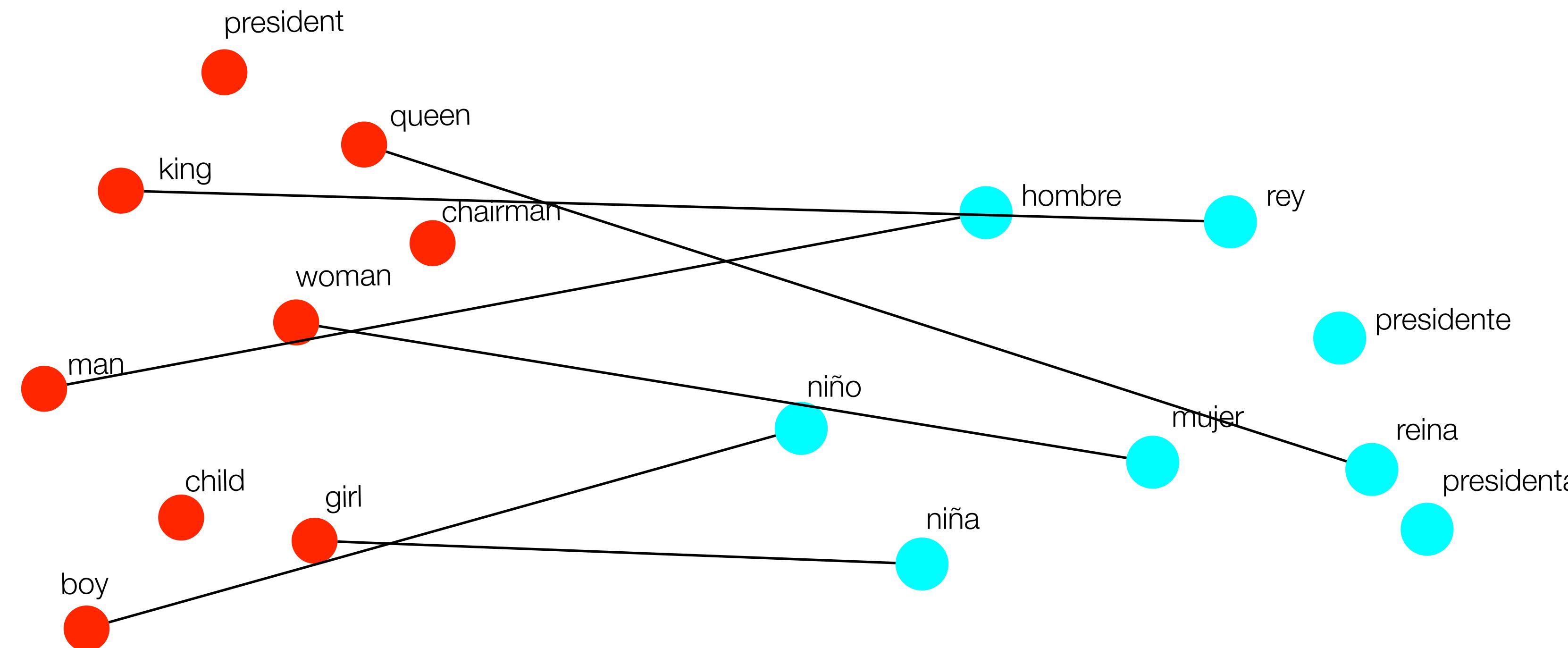
Aligning embeddings through Procrustes analysis

- Mikolov, et al. (2013) [different paper, different alia] asserted that embeddings have a universal quality ... that embeddings from different languages contain similar semantic structure.
- If that is perfectly true — if they contain exactly the same semantic relationships — and the semantic relationships are linear, embeddings in different languages are linear transformations of one another. (If they are centered and normed, the linear transformation is a rotation.)
- Conneau, et al. 2017 (Facebook MUSE Project) - get best alignment with Procrustes based on polysemic bilingual dictionaries.

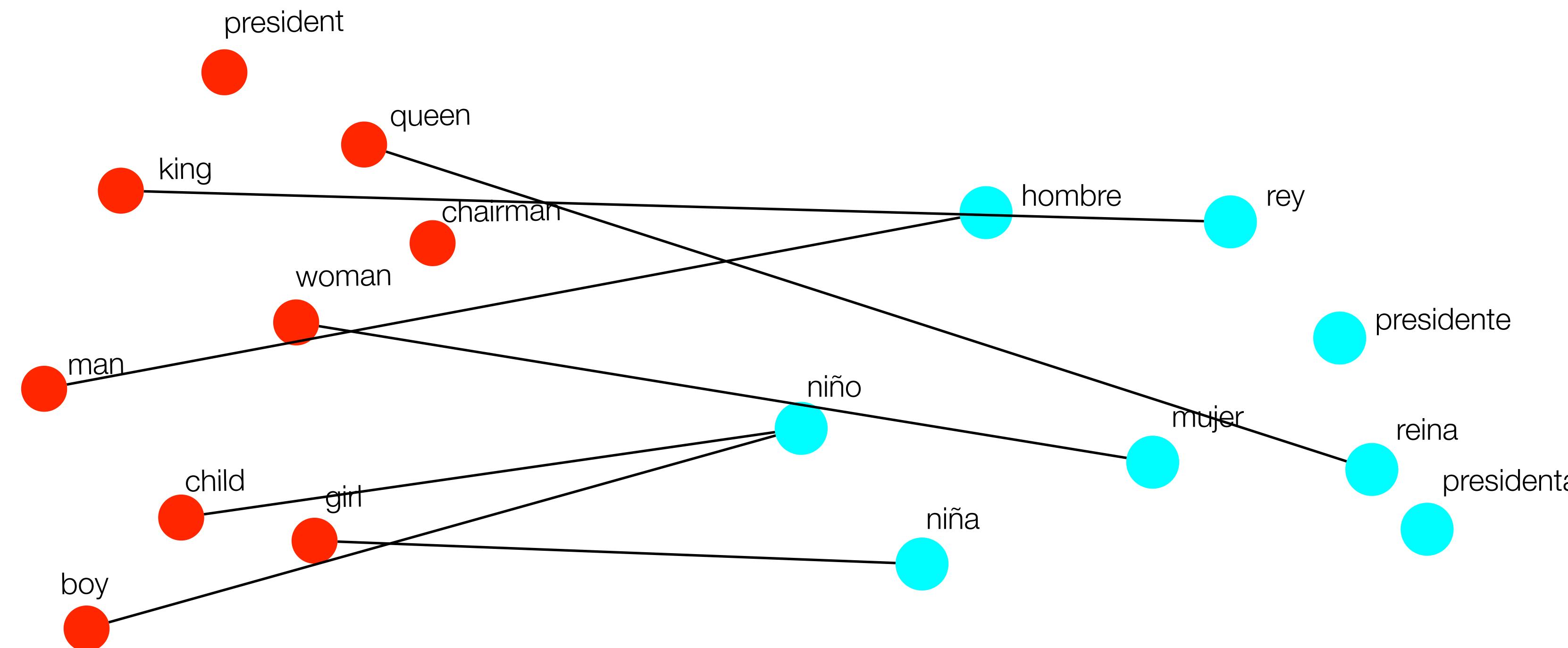
Procrustes through polysemic bilingual dictionaries



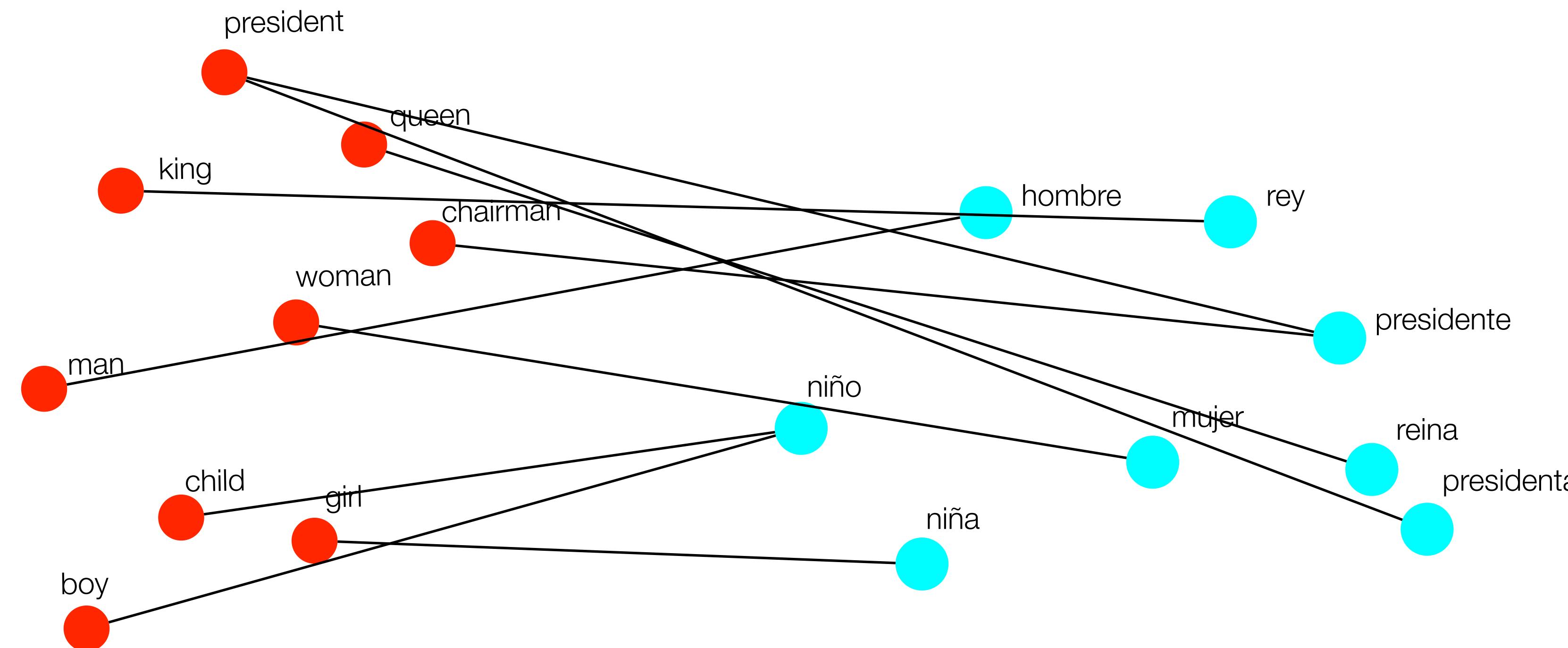
Procrustes through polysemic bilingual dictionaries



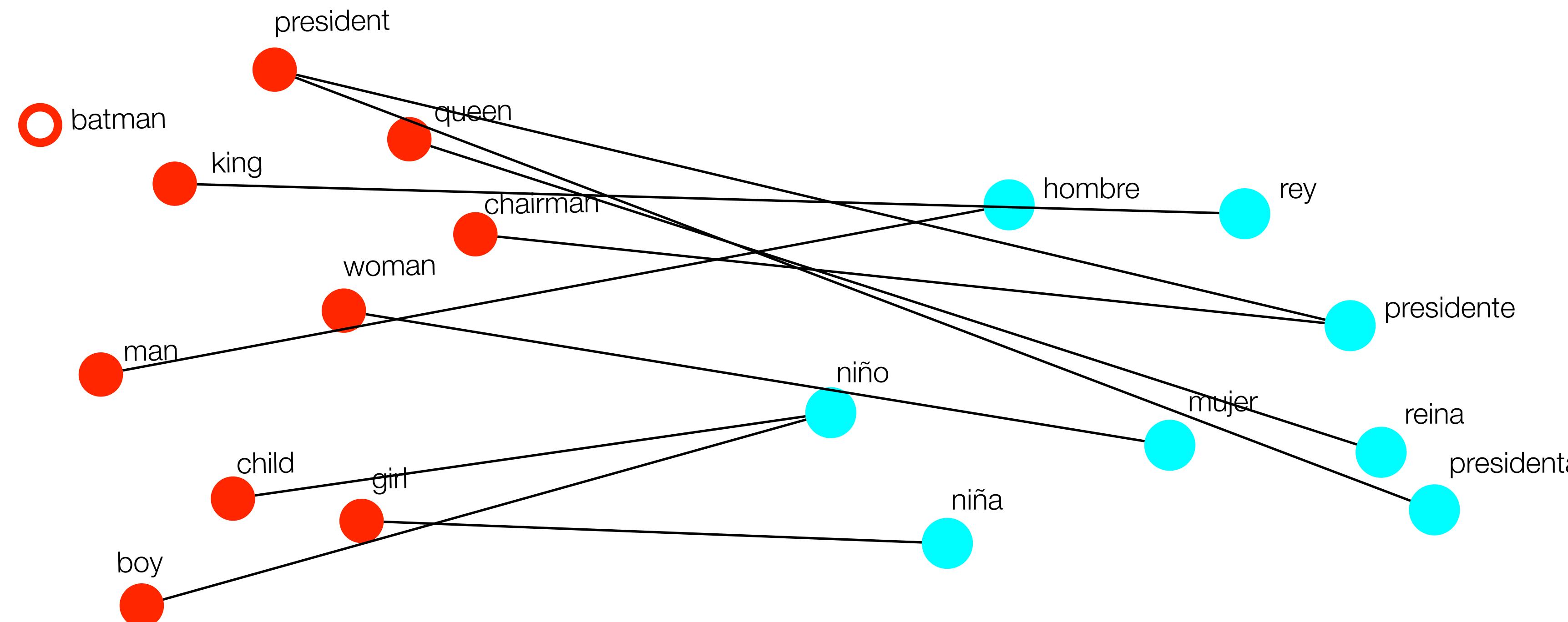
Procrustes through polysemic bilingual dictionaries



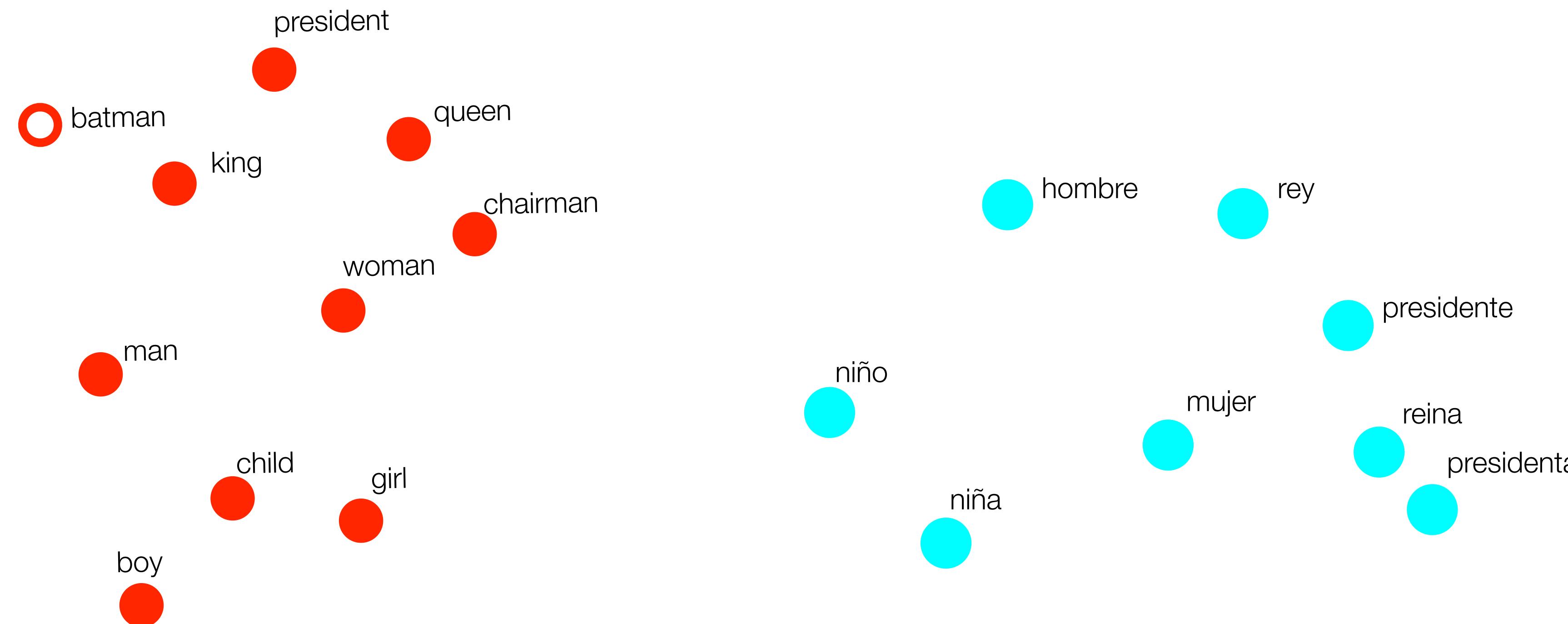
Procrustes through polysemic bilingual dictionaries



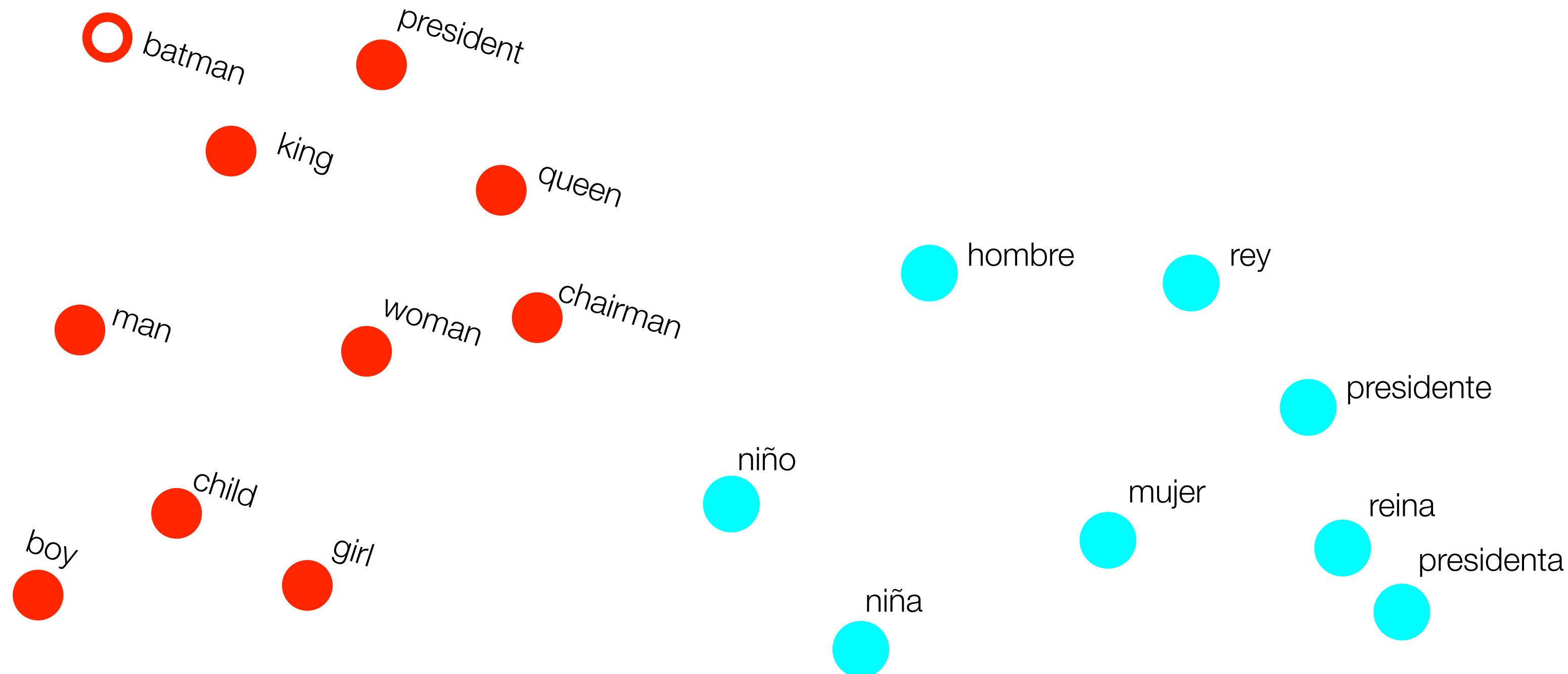
Procrustes through polysemic bilingual dictionaries



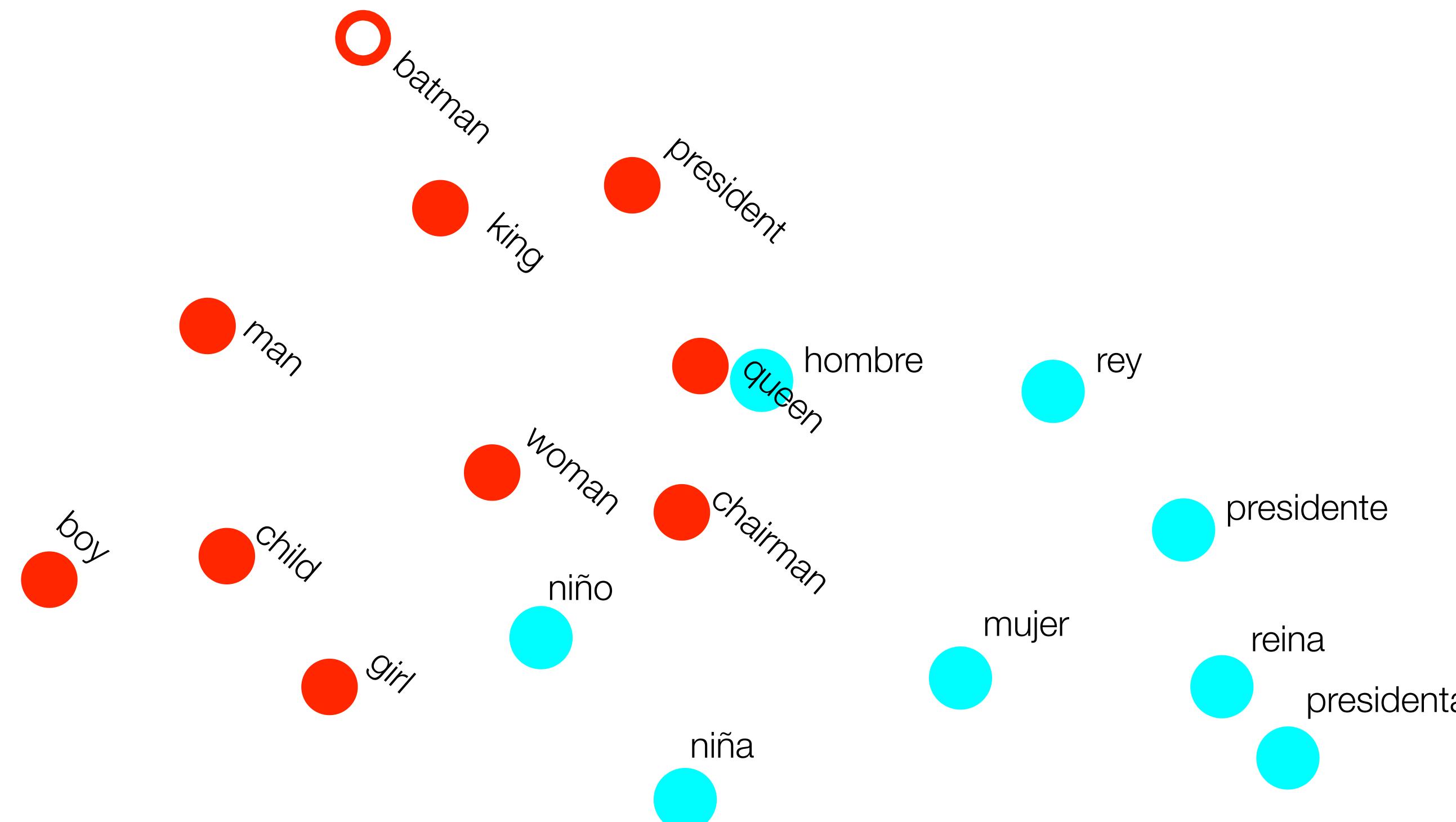
Procrustes through polysemic bilingual dictionaries



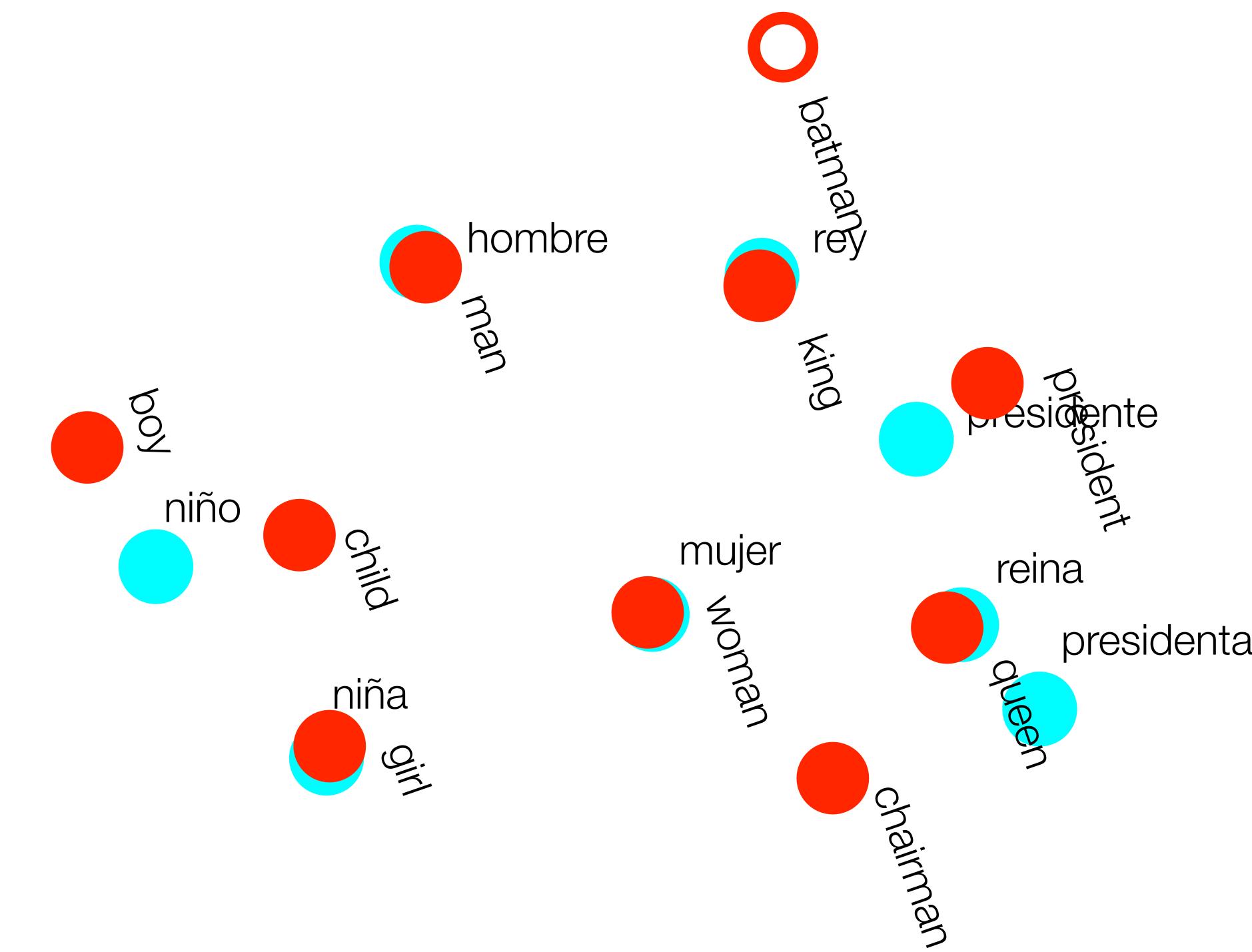
Procrustes through polysemic bilingual dictionaries



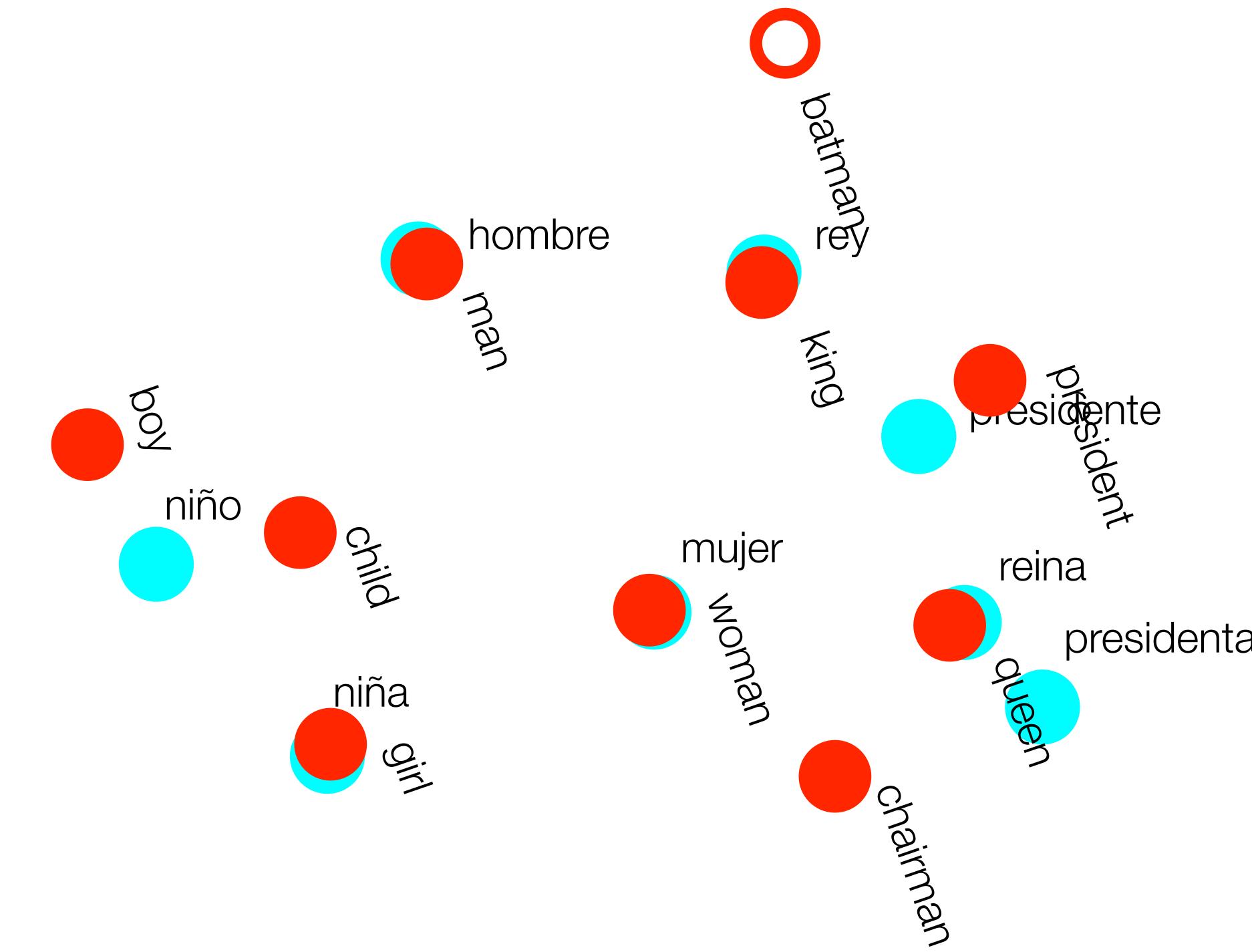
Procrustes through polysemic bilingual dictionaries



Procrustes through polysemic bilingual dictionaries

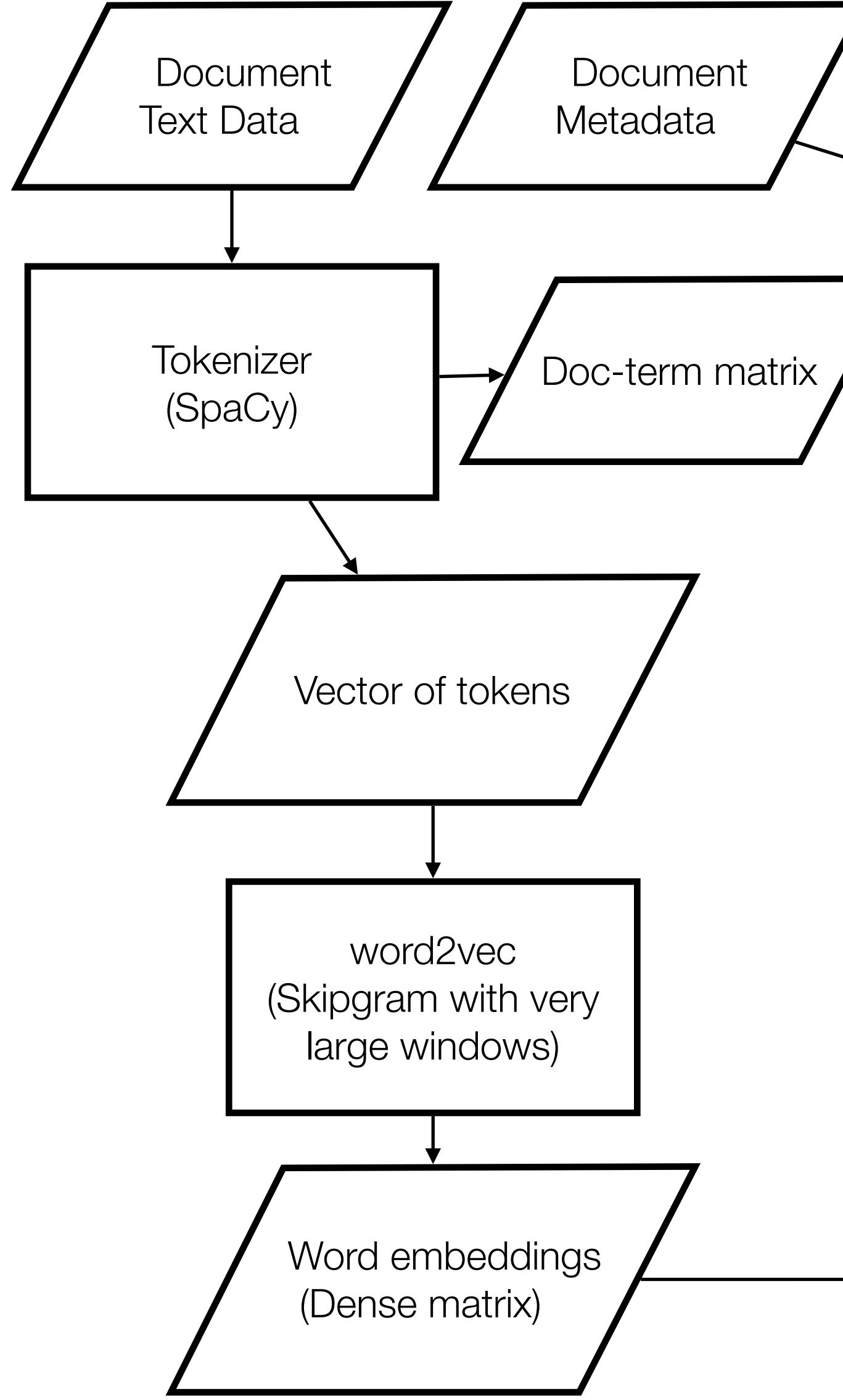


Procrustes through polysemic bilingual dictionaries

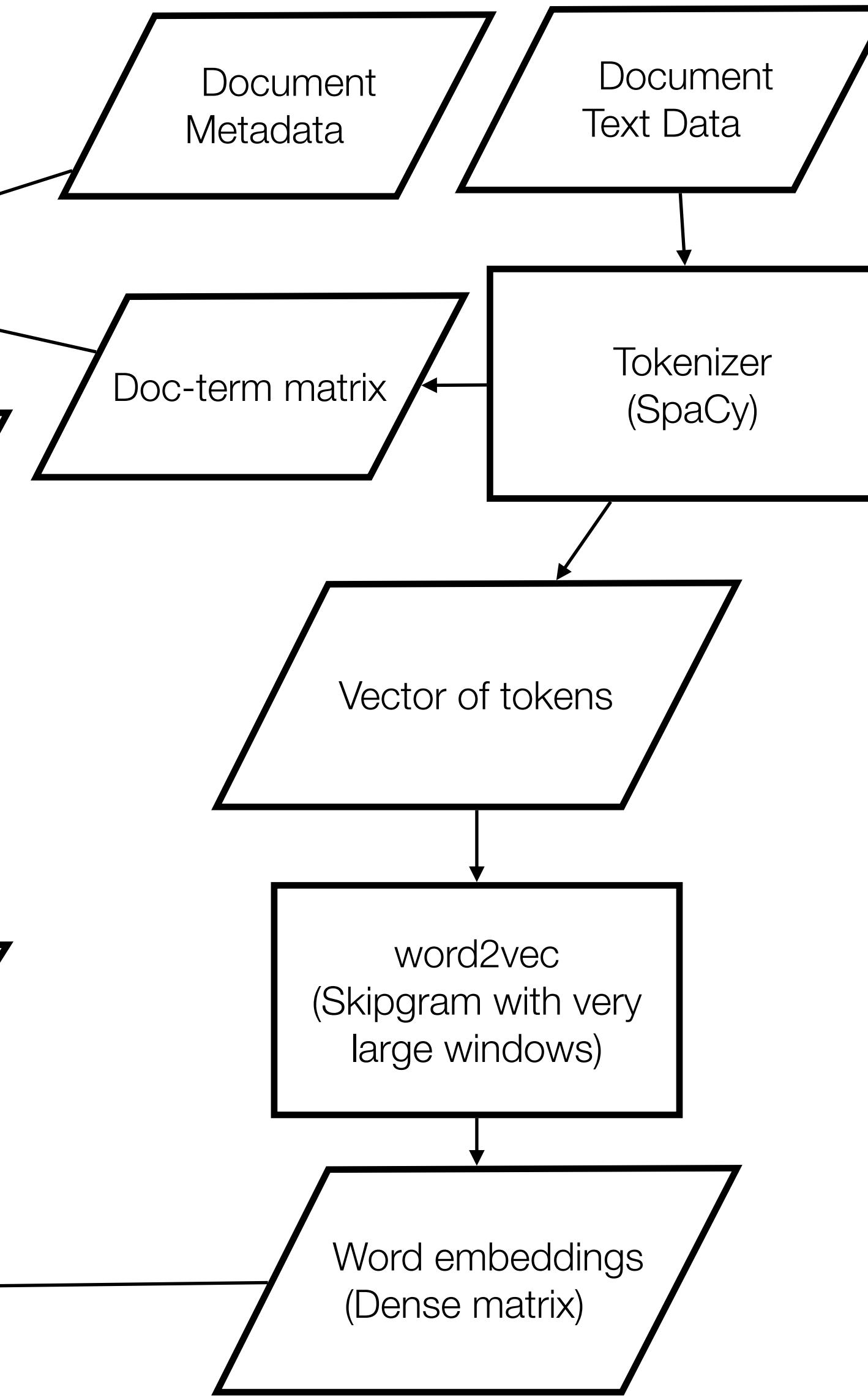


We estimate this Procrustes rotation using the implementation and bilingual dictionaries provided by the **Facebook MUSE** project (Conneau et al 2017). In the analyses reported here, the Czech, Dutch, Finnish, German, Italian, Spanish, and Swedish embeddings were “Procrusted” onto the English embeddings. (PyTorch/GPUs ~ 30 minutes/language.)

Language 1



Language 2

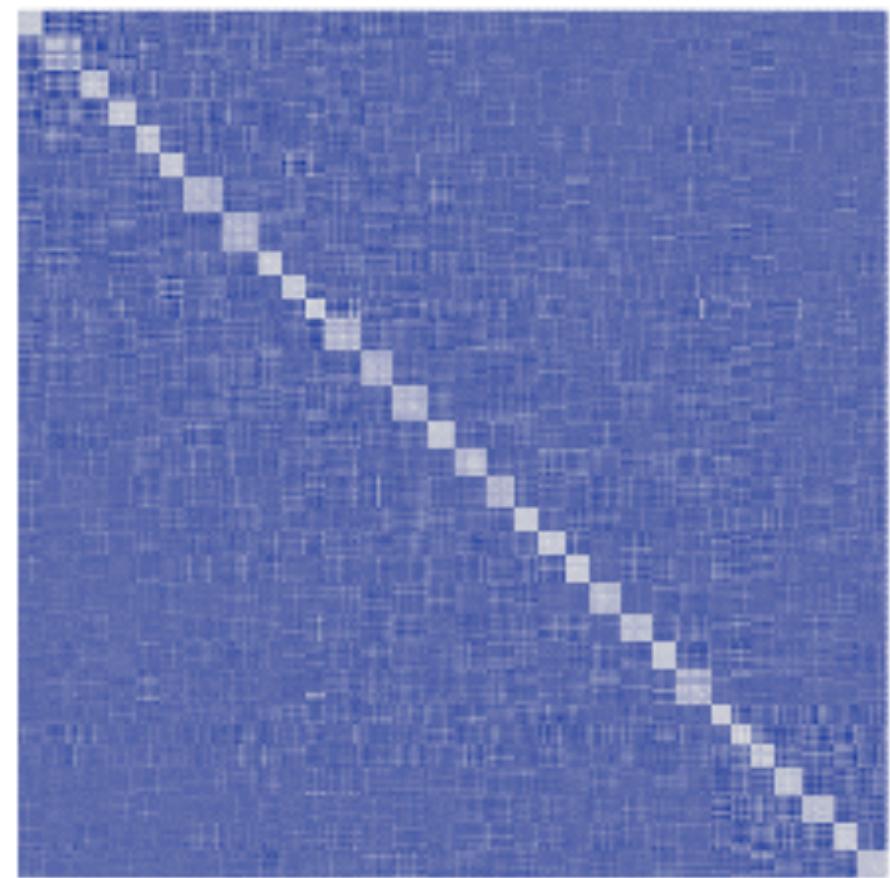


Decompose and rotate to interpretable structure

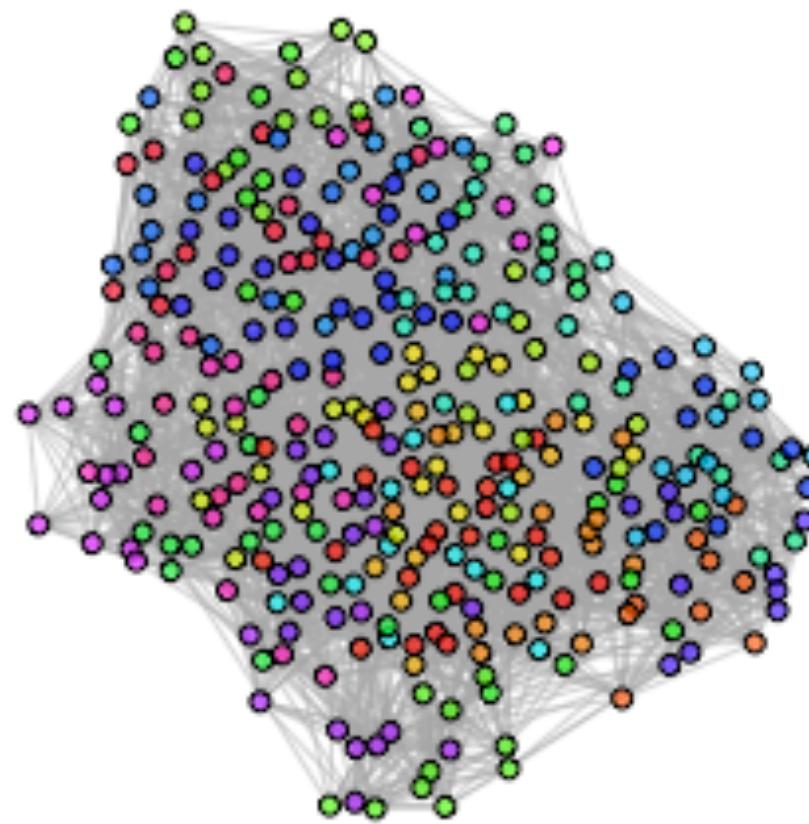
- The general idea is to calculate a Singular Value Decomposition (SVD) of the embeddings matrix, remove the highest frequency dimensions, and rotate to interpretable structure.
 - The most familiar analog is likely from factor analysis. We want to identify dimensions where a small fraction of words load highly on the dimension, and conversely for words to load highly on a small number of dimensions.
 - Topic models like LDA have an *implicit* rotation baked into their identification.
- This can take several recognizable forms, including
 - Old-school psychometrics / multivariate statistics approaches to factor analysis, aka “**Kaiser's Little Jiffy**” (Kaiser 1952) aka “principal factor analysis” (PCA, truncation below eigenvalue=1, varimax rotation to “simple structure”). For homogeneous features (all word counts, movie ratings, etc.) this is very close to standard factor analysis.
 - **Independent Component Analysis (ICA)** (whitening/PCA followed by rotation to maximal independence). Usually cast as a method for blind source separation in signal processing (e.g. the “cocktail party problem”).

Results presented here are from ICA as implemented in FastICA, ~ 1 hour.

Jiffy Communities



NCAA Basketball Games Played, Midseason



Sidebar:

The Little Jiffy is dope.

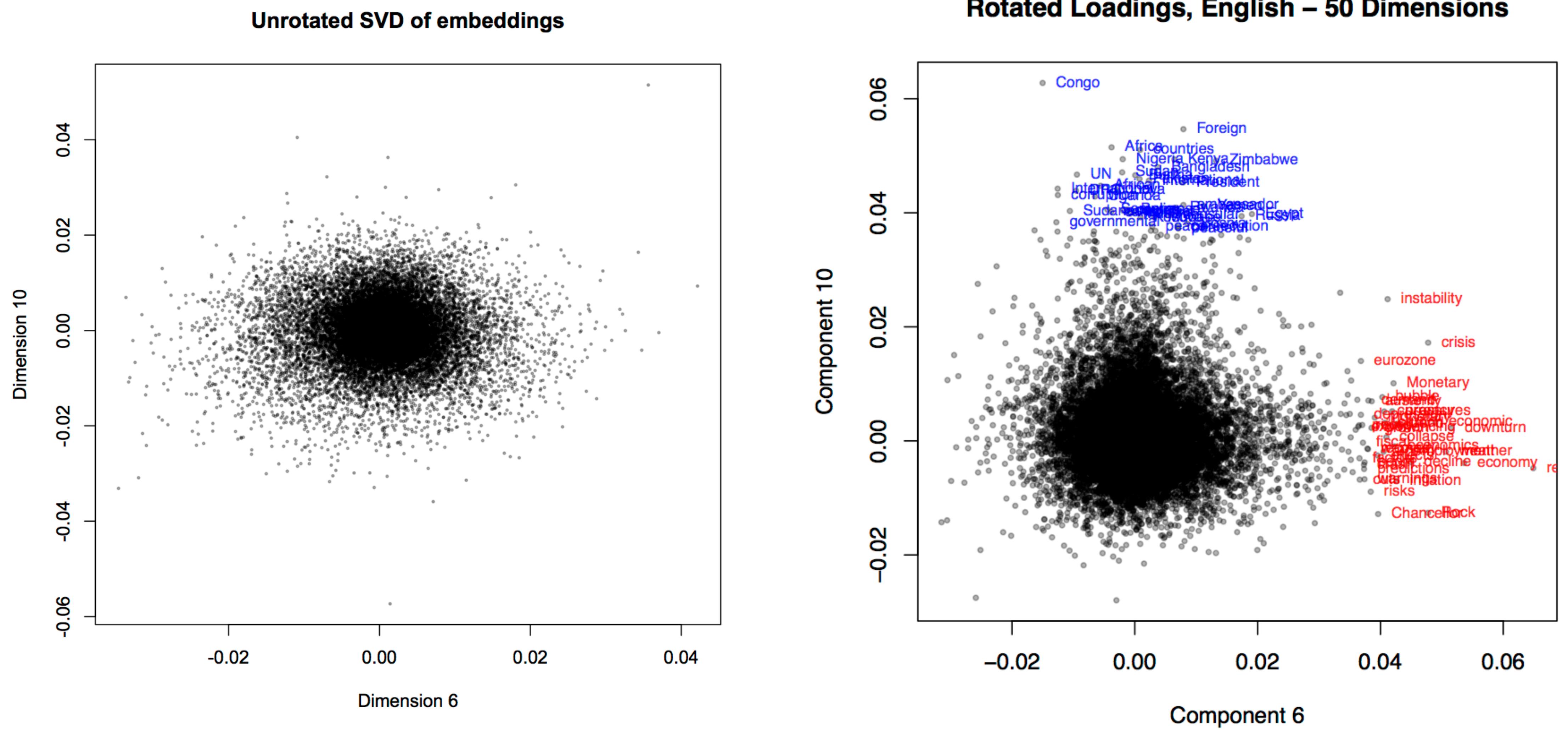
In a different project I show that the Little Jiffy applied to a network adjacency matrix provides best-in-class community detection performance on benchmarks and other examples.

Performance on Basketball Conference Detection

	NMI (Auto K)	NMI (K=32)
Jiffy*	0.990 (31)	1.000
Infomap	0.902 (21)	—
Walktrap	0.829 (13)	0.984
Spectral Partitioning*	0.518 (5)	0.959
Edge Betweenness	0.827 (19)	0.915
Label Propagation	0.767 (15)	—
Spinglass	0.767 (10)	—
Louvain (Multi-level)	0.726 (8)	—
Leading Eigenvector	0.601 (8)	igraph bug
Fast Greedy	0.466 (3)	0.511
Optimal	I should live so long.	—

* Coded by Monroe in R. Same eigengap method. All others in igraph in R.

Effect of rotation



Our large-window embeddings *contain* a topic model.

43 substantive “topics” revealed by rotation of embeddings

Top 5 loading English terms

Terrorism	terrorism; terrorist; threat; intelligence; terrorists	Rights	liberties; freedom; liberty; freedoms; rights
Representation	election; elections; electorate; elected; voters	Prisons	prison; prisons; prisoners; custody; Prison
Professions	professional; qualified; profession; trained; nurses	Maritime	sea; fishermen; fishing; vessels; maritime
Inequality	poorest; inequality; rich; fairness; society	Local/Regional	counties; constituency; regional; cities; areas
Discrimination	men; discrimination; equality; women; female	Religion	Muslim; religious; gay; religion; Christian
Health	doctor; medical; hospital; NHS; doctors	Bureaucracy	bureaucracy; tape; administrative; bureaucratic; savings
Agriculture	farmers; agricultural; farming; agriculture	Transport	Transport; railways; transport; rail; roads
Agreements	agreement; agreements; negotiations; negotiate; cross	Business	businesses; sized; tape; small; enterprises
Environment	environment; Natural; wildlife; biodiversity; beauty	Education	school; teachers; schools; teacher; Education
International Crises	Zimbabwe; Iran; Foreign; Burma; Sudan	Housing	housing; houses; Housing; house; homes
Private/Public	private; public; sector; privatisation; voluntary	Crime	offence; crime; offences; crimes; fraud
Banks	Rock; shareholders; company; RBS; assets	Science/R&D	science; investment; technology; R
Families	family; mother; father; wife; mothers	Sport	sport; Olympics; sports; sporting; games
Public Health	smoking; alcohol; substances; tobacco; products	Taxes	taxation; tax; HMRC; taxes; Paymaster
Budget	money; resources; funds; expenditure; budget	Labor	workers; employees; employer; employers; staff
Energy	energy; electricity; renewables; Energy; Ofgem	Defense	Army; Defence; defence; troops; reservists
Global (Aid/Climate)	0.7; climate; G8; DFID; International	Disabilities	disabilities; disabled; disability; Disability; DLA
Macroeconomy	deficit; fiscal; economy; borrowing; finances	OECD/Trade	Germany; France; Sweden; Canada; Australia
Media	television; broadcast; media; BBC; broadcasting	Immigration	immigration; asylum; migrants; Immigration; migration
Courts	court; courts; judicial; tribunal; appeals	Europe	European; Europe; Brussels; EU; directive
Jobs	unemployed; jobseeker; incapacity; unemployment; Work	History/Heritage	museum; museums; anniversary; memorial; heritage
		Universities	university; universities; student; students; fees

Our topic model incorporates component languages simultaneously.

Example topics, top 20 terms, all languages + English

Terrorism		Environment		Macroeconomy		Immigration	
Top terms	Top English terms	Top terms	Top English terms	Top terms	Top English terms	Top terms	Top English terms
de_Terrorismus	en_terrorism	de_Naturschutz	en_environmental	de_Konjunktur	en_deficit	es_inmigrantes	en_immigration
it_po	en_terrorist	es_ambiental	en_Natural	de_Wachstum	en_fiscal	en_immigration	en_asylum
it_terrorismo	en_threat	sv_miljöbalken	en_wildlife	it_debito	en_economy	es_extranjeros	en_migrants
en_terrorism	en_intelligence	de_Umwelt	en_biodiversity	es_déficit	en_borrowing	de_Zuwanderung	en_Immigration
sv_terrorism	en_terrorists	de_Natur	en_beauty	en_deficit	en_finances	es_inmigración	en_migration
en_terrorist	en_terror	en_environmental	en_natural	es_crecimiento	en_recession	en_asylum	en_immigrants
nl_AIVD	en_Terrorism	en_Natural	en_environment	sv_finanser	en_economic	en_migrants	en_visa
nl_terroristische	en_intercept	fi_luonnon	en_Environment	es_economía	en_recovery	de_Asylbewerber	en_citizenship
it_ndrangheta	en_Intelligence	cs_prostředí	en_conservation	es_inflación	en_euro	en_Immigration	en_refugees
en_threat	en_extremism	en_wildlife	en_Protection	sv_räntan	en_growth	en_migration	en_Border
de_Anschläge	en_counter	de_Flächen	en_fracking	de_Finanzpolitik	en_Monetary	sv_flyktingpolitik	en_visas
en_intelligence	en_proscription	es_ambiente	en_rivers	it_deficit	en_monetary	it_soggiorno	en_migrant
de_Sicherheitsbehörden	en_TPIMs	es_medioambiental	en_pollution	en_fiscal	en_eurozone	es_asilo	en_refugee
sv_terrorismen	en_ISC	es_Ambiente	en_trees	sv_ekonomi	en_inflation	de_Asylrecht	en_nationals
en_terrorists	en_orders	en_biodiversity	en_countryside	de_Stabilitäts-	en_Chancellor	es_extranjería	en_passport
de_Verfassungsschutz	en_Carlile	cs_přírody	en_reservoir	sv_finanserna	en_debt	sv_invandring	en_illegally
sv_terrorister	en_MI5	sv_biologisk	en_habitats	es_Monetario	en_currency	de_Aufenthaltsrecht	en_deportation
nl_dreiging	en_al	sv_naturen	en_Antarctic	de_Neuverschuldung	en_Budget	en_immigrants	en_deported
sv_Svar	en_7/7	en_beauty	en_species	de_konjunkturellen	en_borrow	de_Asyl	en_Borders
nl_aanslagen	en_9/11	es_Costas	en_planning	es_Estabilidad	en_billion	sv_arbetskraftsinvandring	en_UKBA

An additional 33 dimensions reveal rhetorical “topics”

Top 5 loading English terms

Issues	issues; matters; circumstances; questions; situations	Quotes	read; Times; Daily; newspaper; article
Nonsense	nonsense; misleading; somehow; neither; utterly	Problems/Solutions	problems; solve; problem; solution; solutions
Groups	Association; Society; Federation; organisations; Institute	Comparisons	rather; Rather; More; less; more
Compliance	fines; inspection; penalties; CQC; sanctions	Skepticism	might; somewhat; bit; seem; seems
Failure	promises; failed; promised; promise; Speech	Wondering	wonder; explain; how; decide; whether
Studies	analysis; study; survey; evaluation; commissioned	Questions/Answers	answer; questions; answers; question; answered
Decisions	decisions; decision; steps; account; choices	Praise	tribute; pleasure; thanks; gratitude; praise
Consequences	consequences; impact; unintended; effect; effects	Differences	difference; distinction; between; differences; relationship
Statistics	trend; numbers; figures; statistics; number	Alternatives	options; option; alternative; alternatives; choice
I Am	glad; afraid; sorry; grateful; pleased	Costs	costs; cost; taxpayer; incurred; expensive
Objectives	objectives; objective; achieve; achieving; targets	Uncertainty	uncertainty; concern; fears; risk; danger
Standards	level; threshold; high; levels; inflation	Deliberations	debate; debates; pleasure; discussion; meetings
Initiatives	pilot; awareness; engagement; learn; initiatives	Change	change; changes; changing; changed; radical
Reasons	reasons; reason; explain; why; arguments	My	Friend; Friends; constituency; constituent; constituents
Timetable	timetable; delay; wait; summer; recess	Rules	rules; guidelines; guidance; law; definition
Disaster	disaster; tragic; tragedy; incident; died	Procurement	procurement; contract; value; project; contractors
		Transparency	information; transparency; openness; disclosure; Information

Rhetorical Topics, Top 20 Keywords, All Languages + English

#Initiatives	
Top terms	Top English terms
nl_initiatieven	en_pilot
sv_åtgärder	en_awareness
cs_sněmovnou	en_engagement
nl_slot	en_learn
de_Projekte	en_initiatives
de_Initiativen	en_strategy
nl_voorlichting	en_initiative
it_iniziative	en_active
es_campañas	en_launched
nl_stimuleren	en_behaviour
it_promozione	en_approach
de_Kampagne	en_pilots
cs_prevence	en_principles
en_pilot	en_co
fi_piirtein	en_prevention
es_colaboración	en_curriculum
de_Initiative	en_relationships
nl_comunicatie	en_practice
nl_cetera	en_activity
en Awareness	en_plan

#Failure	
Top terms	Top English terms
de_fehlt	en_promises
de_Ankündigungen	en_failed
de_Statt	en_promised
en_promises	en.promise
de_Fehlanzeige	en_Speech
de_versprochen	en_rhetoric
sv_vackra	en_failure
de_Versprechen	en_lack
en_failed	en_Yet
es_promesas	enfails
en_promised	en_failing
en.promise	en_yet
it_promesse	en_nothing
sv_handling	en_supposed
de_Versprechungen	en_incompetence
en_Speech	en_ignored
de_Taten	en_spin
nl_beloft	en_Instead
es_falta	en_missed
es_intenciones	en_?

Why we estimate 150 topics and cull from there by hand

- Other “topics” capture idiosyncrasies of specific corpora.
- For example, several components are lists of Dutch names, exposing the fact that the ParlSpeech data for the Netherlands incorrectly includes recorded vote lists as “speech.”
- To a “topic” model this is a topic. These words co-occur together in a lot of documents.
- To a human, it’s noise we wish had been removed in the preprocessing stage.
- By making the embeddings interpretable, we can remove them at this stage.

Component 27

nl_Noorman

nl_Ortega

nl_Halsema

nl_Karimi

nl_Kant

nl_Spies

nl_Snijder

nl_Gerkens

nl_Scheltema

nl_Giskes

nl_Arib

nl_Wiegman

nl_Gesthuizen

nl_Ferrier

nl_Griffith

nl_Sipkes

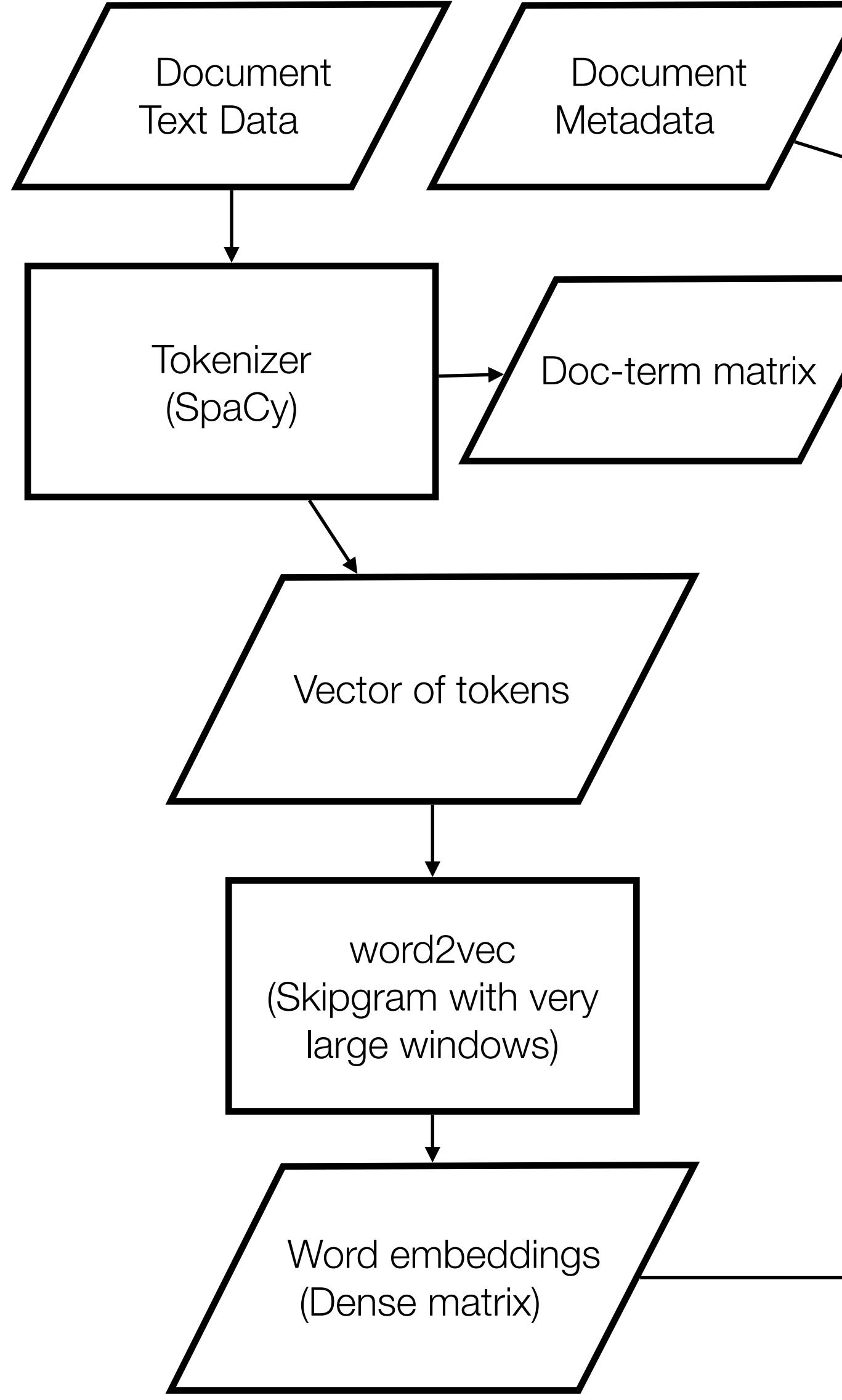
nl_Kalsbeek

nl_Ouwéhand

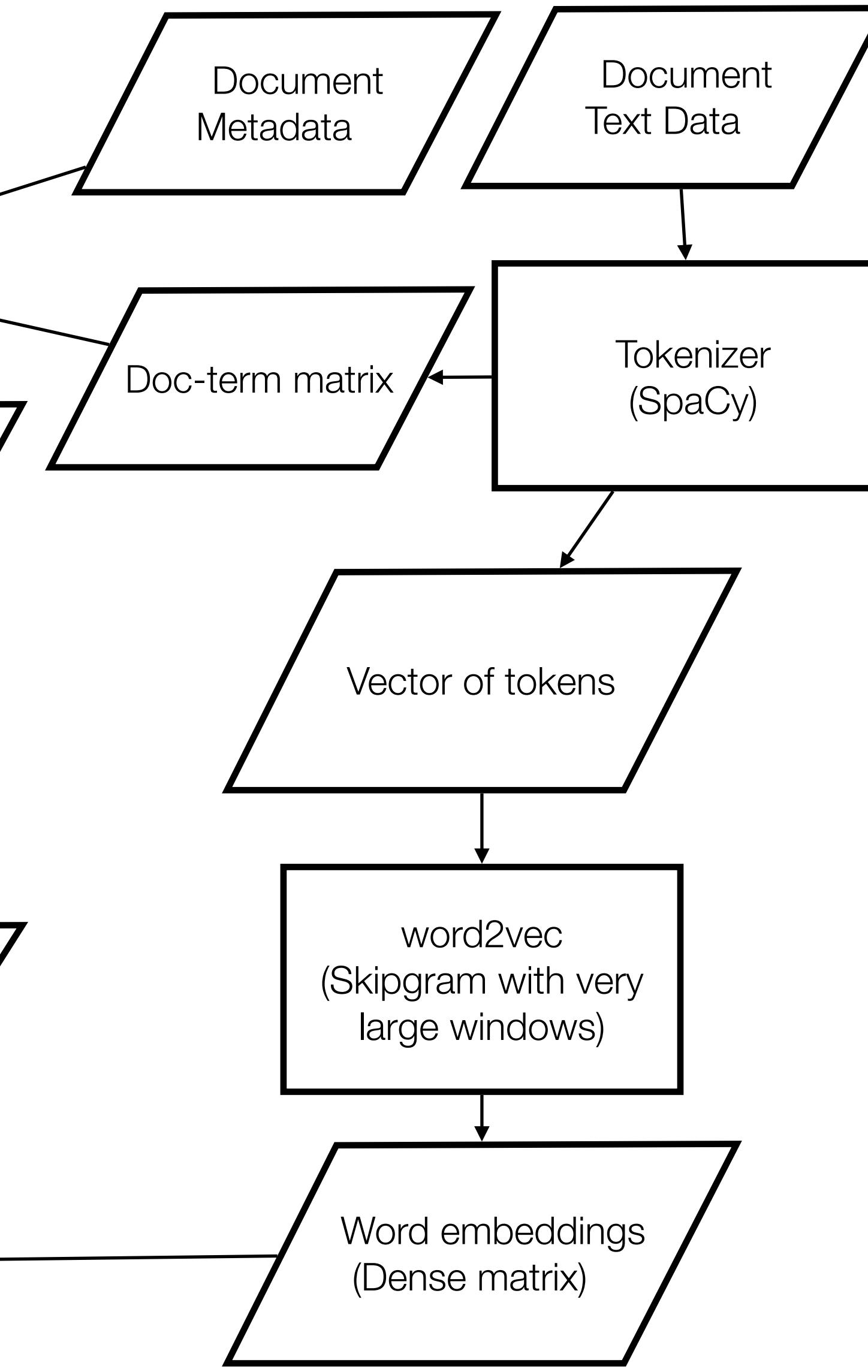
nl_Leijten

nl_Sterk

Language 1



Language 2

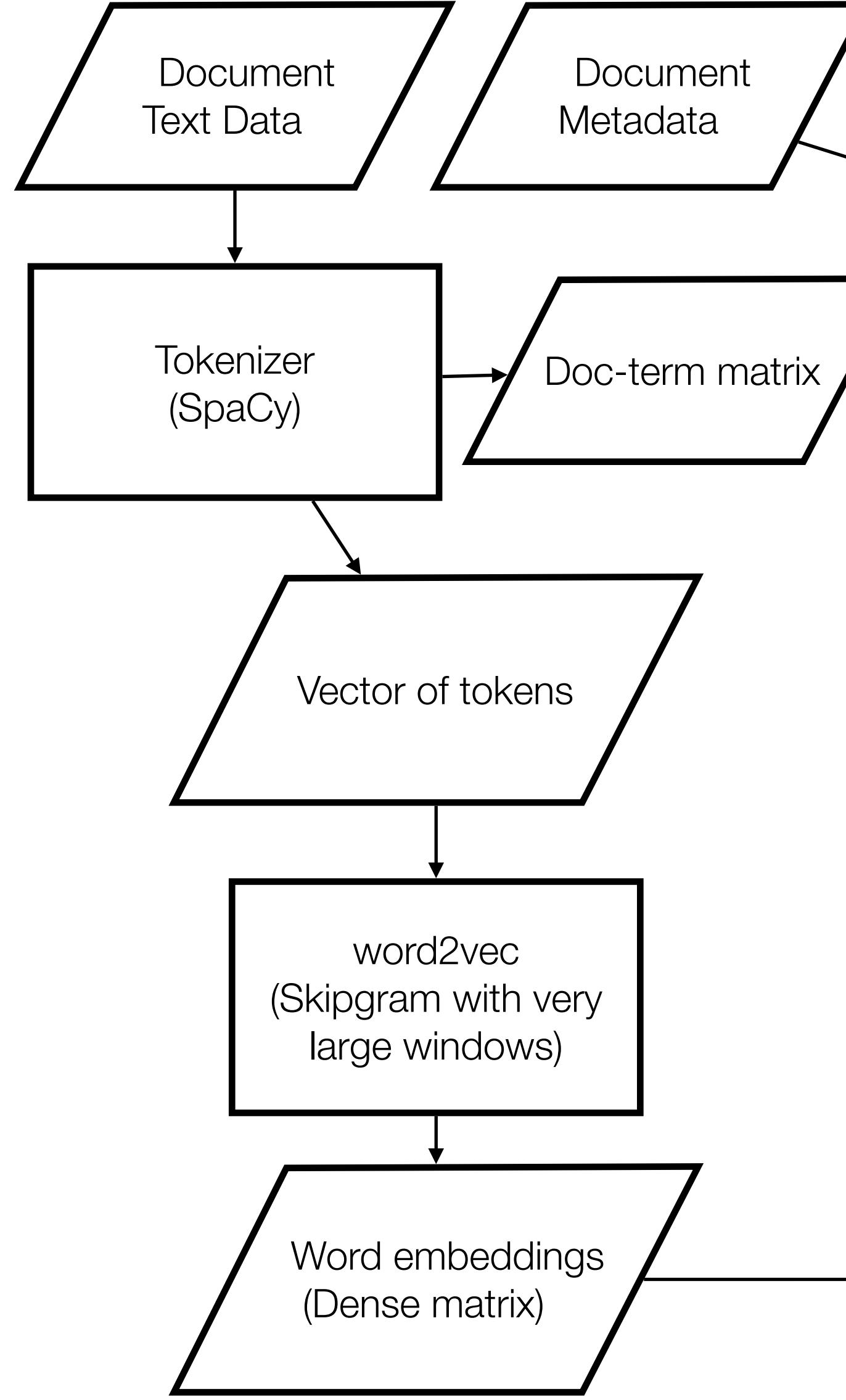


Scoring documents and aggregates of documents (like parties)

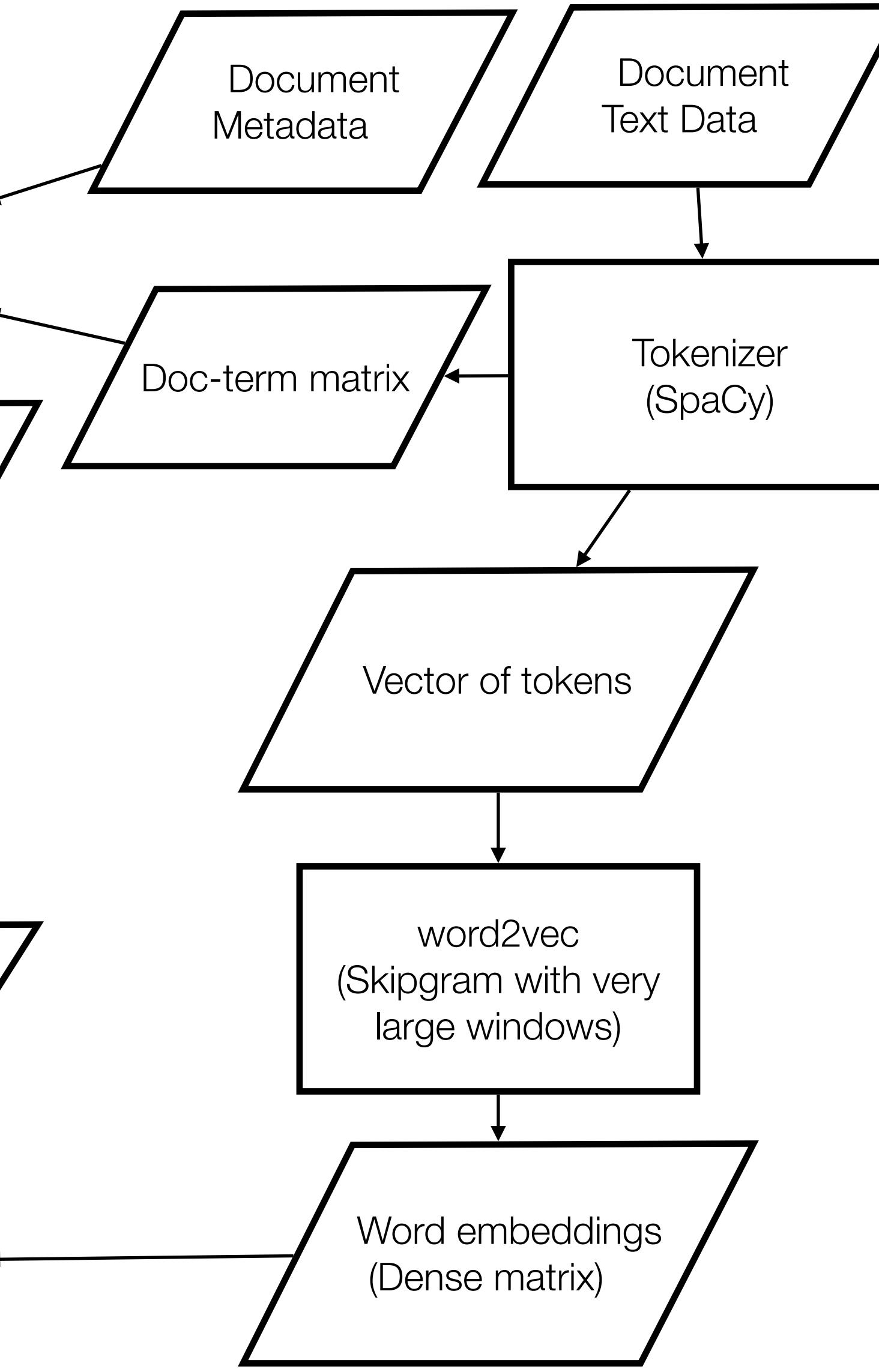
- Represent documents with regularized and weighted doc-token PPMI. This is in essence, “Fightin’ Words” (Monroe, et al. 2008).
- Raw topic score = cosine (doc-vector, topic-vector).
- Norm (L2) scores across topics/dimensions of interest (e.g., substantive topics).
- For groupings like party or party-year, weighted / centered / normed average of scores for documents produced by affiliated speakers.

For scoring individual documents, this took 6-8 hours / corpus on a 512GB machine. (Massive, dense matrix operations.)

Language 1



Language 2



All graphics color parties by Comparative Manifesto Project family



Size is proportional to share of seats in parliament for time period.

Party-year graphics plot parties that were in government for more than six months with a black border.

*Remember: party only enters the model as a “group_by” key for speeches.
Party family and government/opposition status don’t enter at all.*

We can track attention to issues by parties over time.

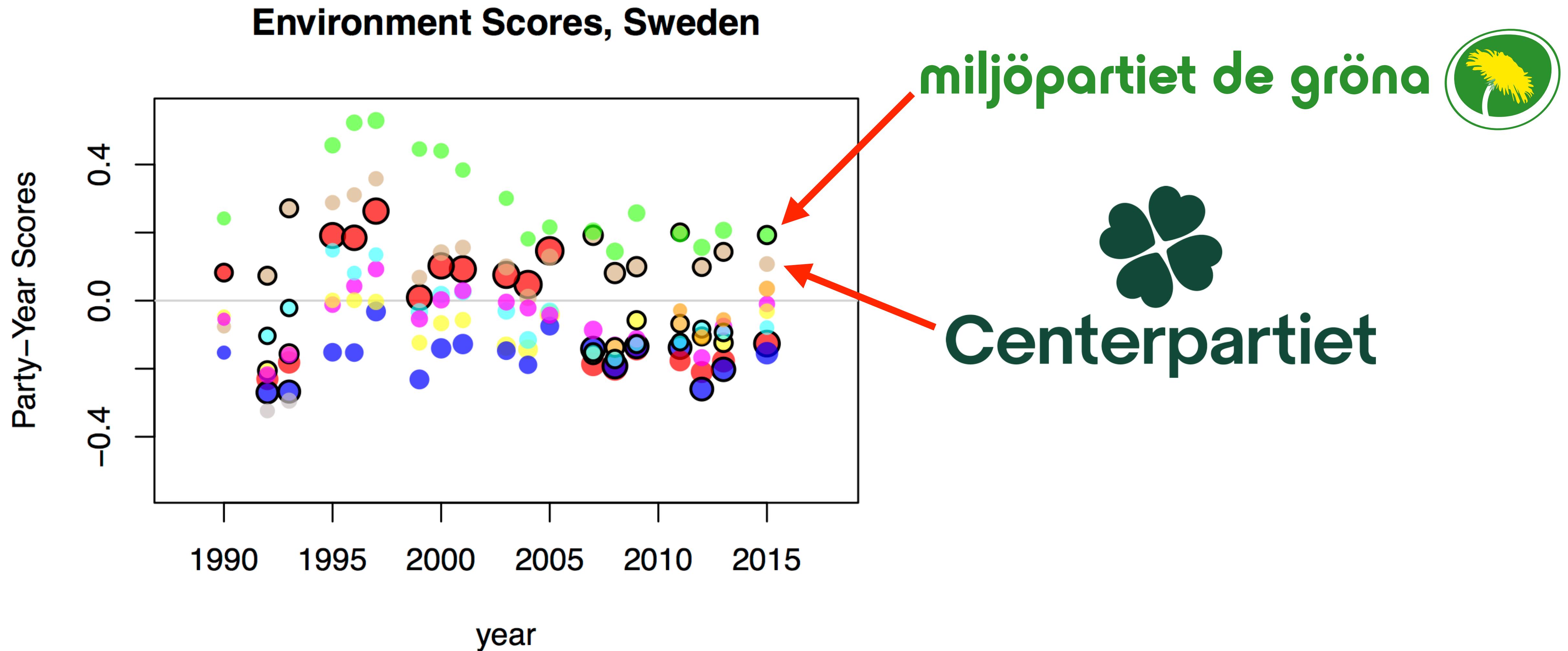
Sanity checks –

Do parties dominate the issues they are expected to?

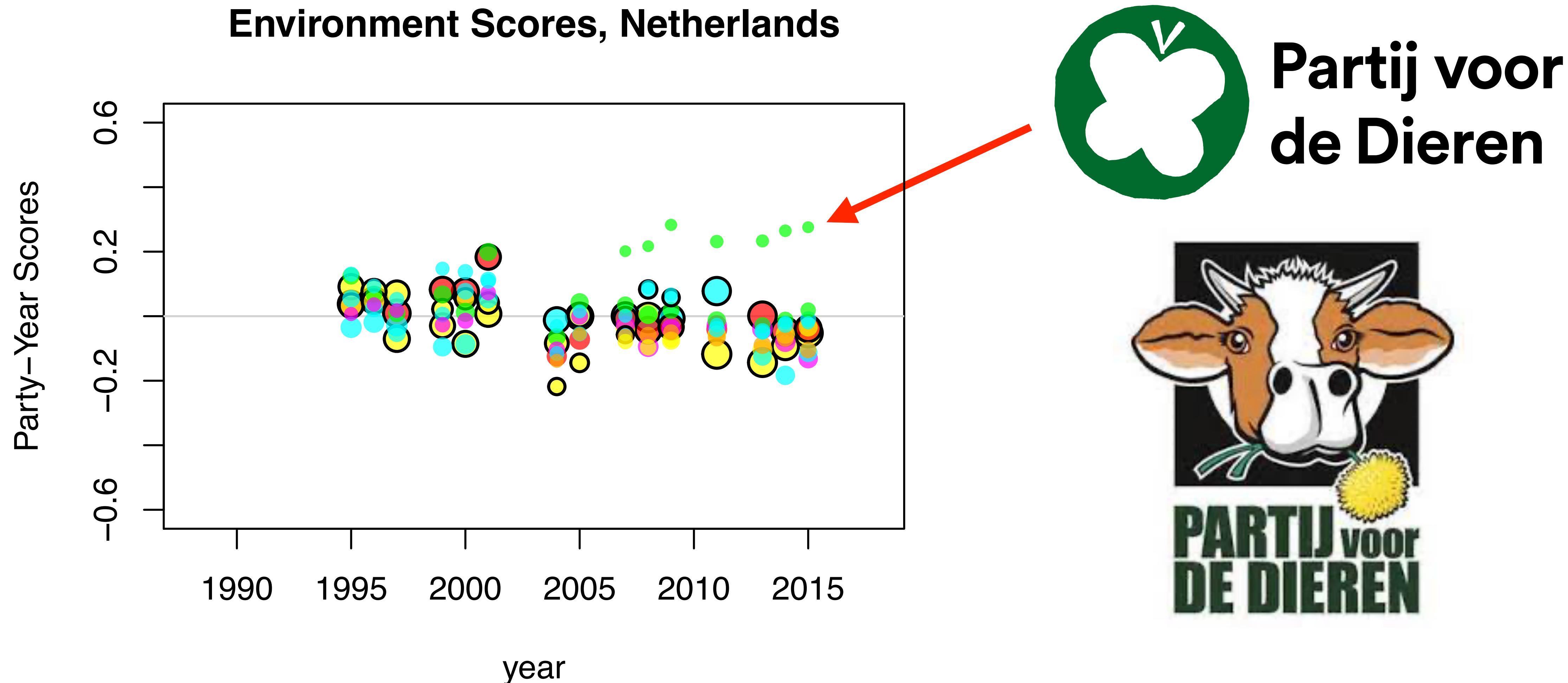
Do movements over time correspond to events?

Are estimates stable over time, even though years are scored independently?

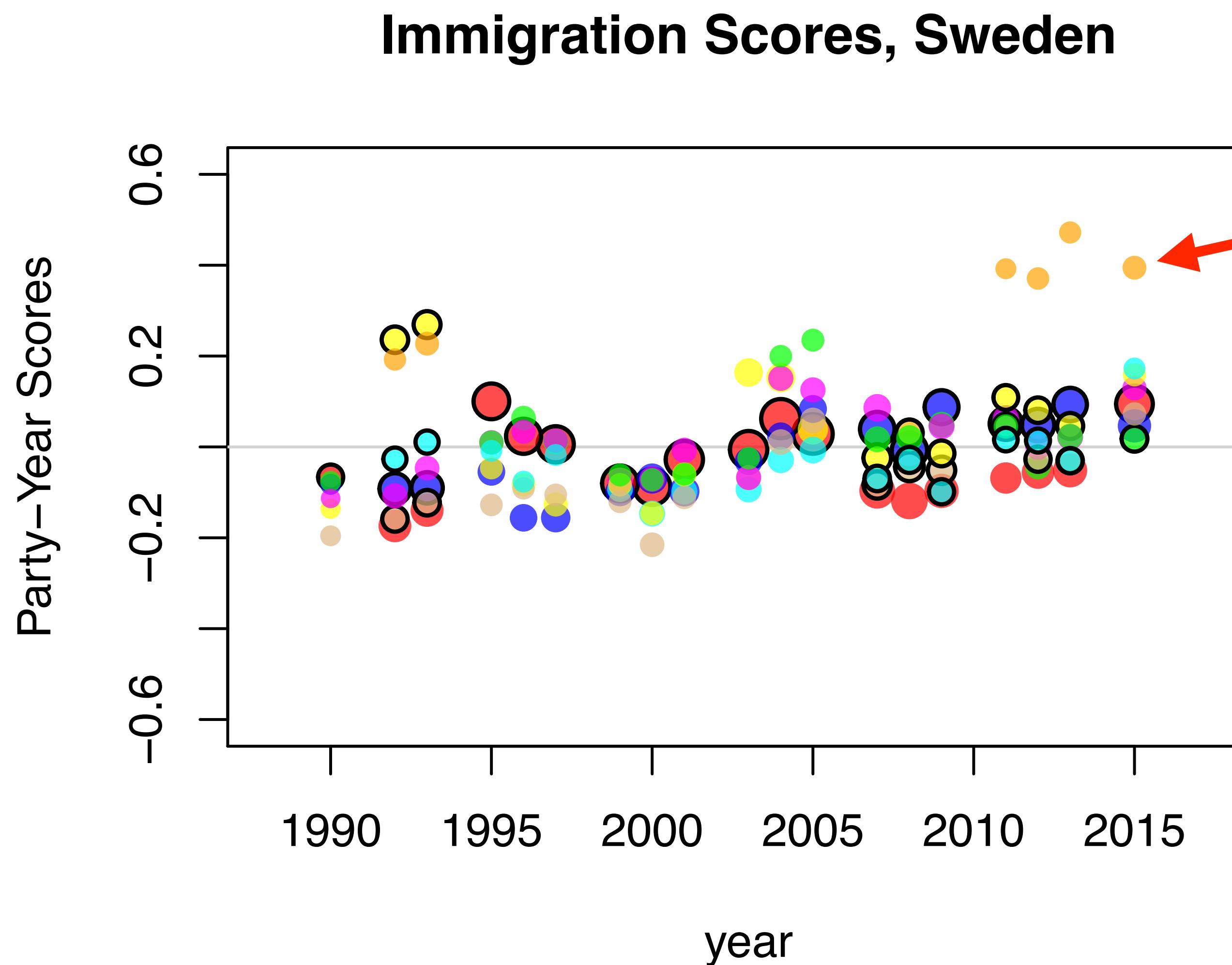
Scores of Swedish parties on the “Environment” topic over time



Scores of Dutch parties on the “Environment” topic over time

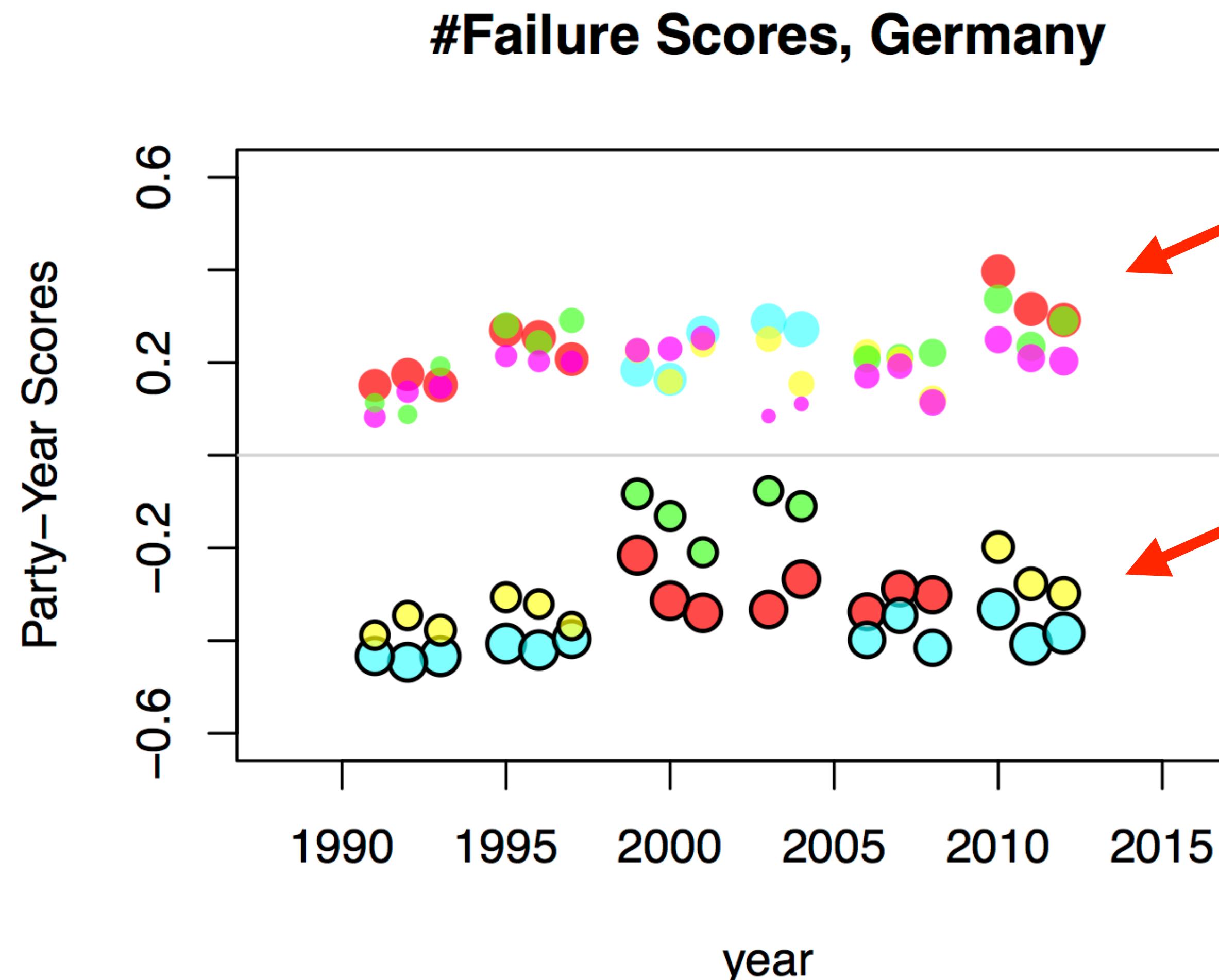


Scores in Sweden on the “Immigration” topic over time



We can characterize how oppositions and governments behave

Where you stand is where you sit ...



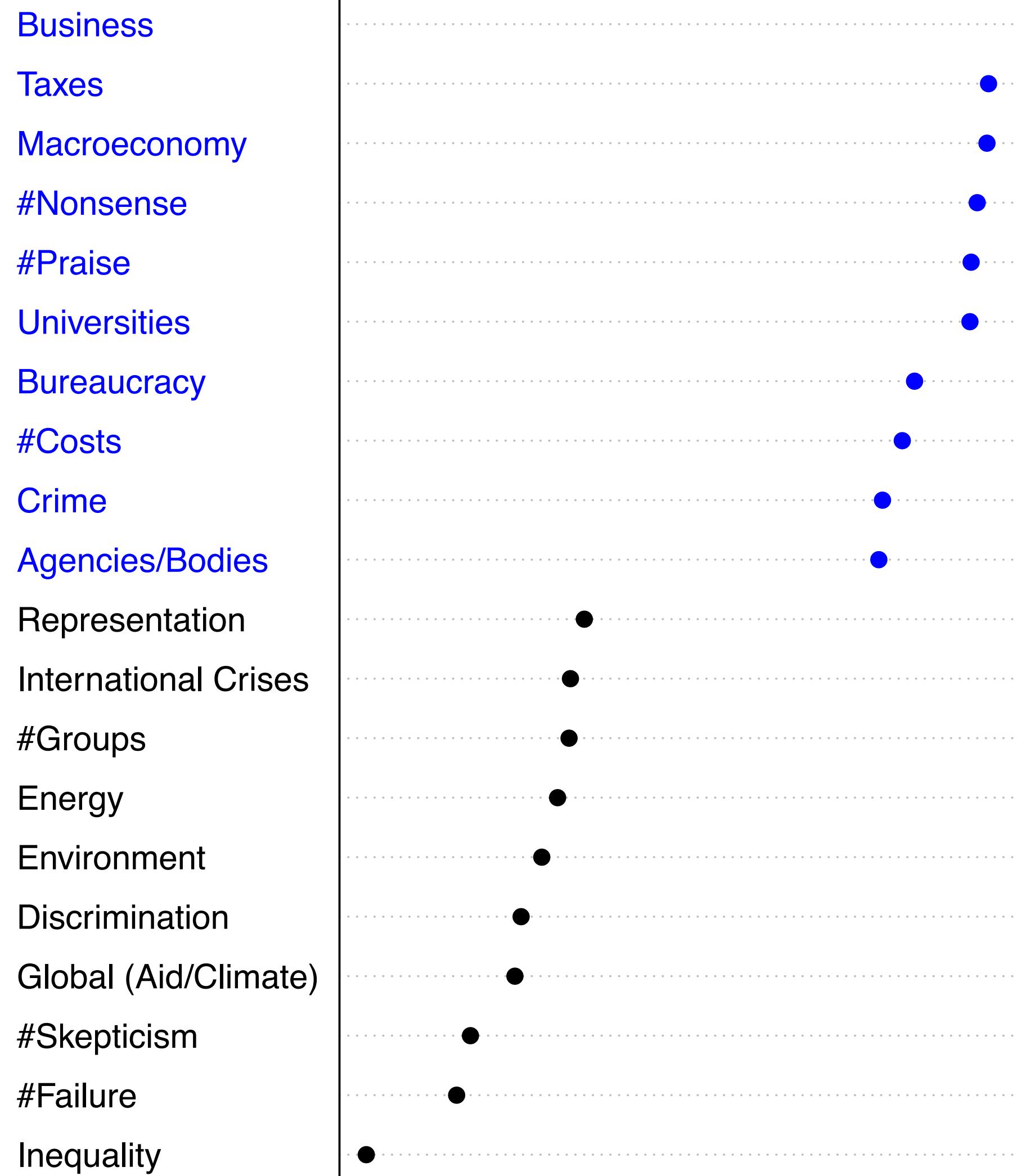
Every party talks about failure when it is out of government.

No party talks about failure when it is in government.

This is true of every party in our data — all years, all countries.

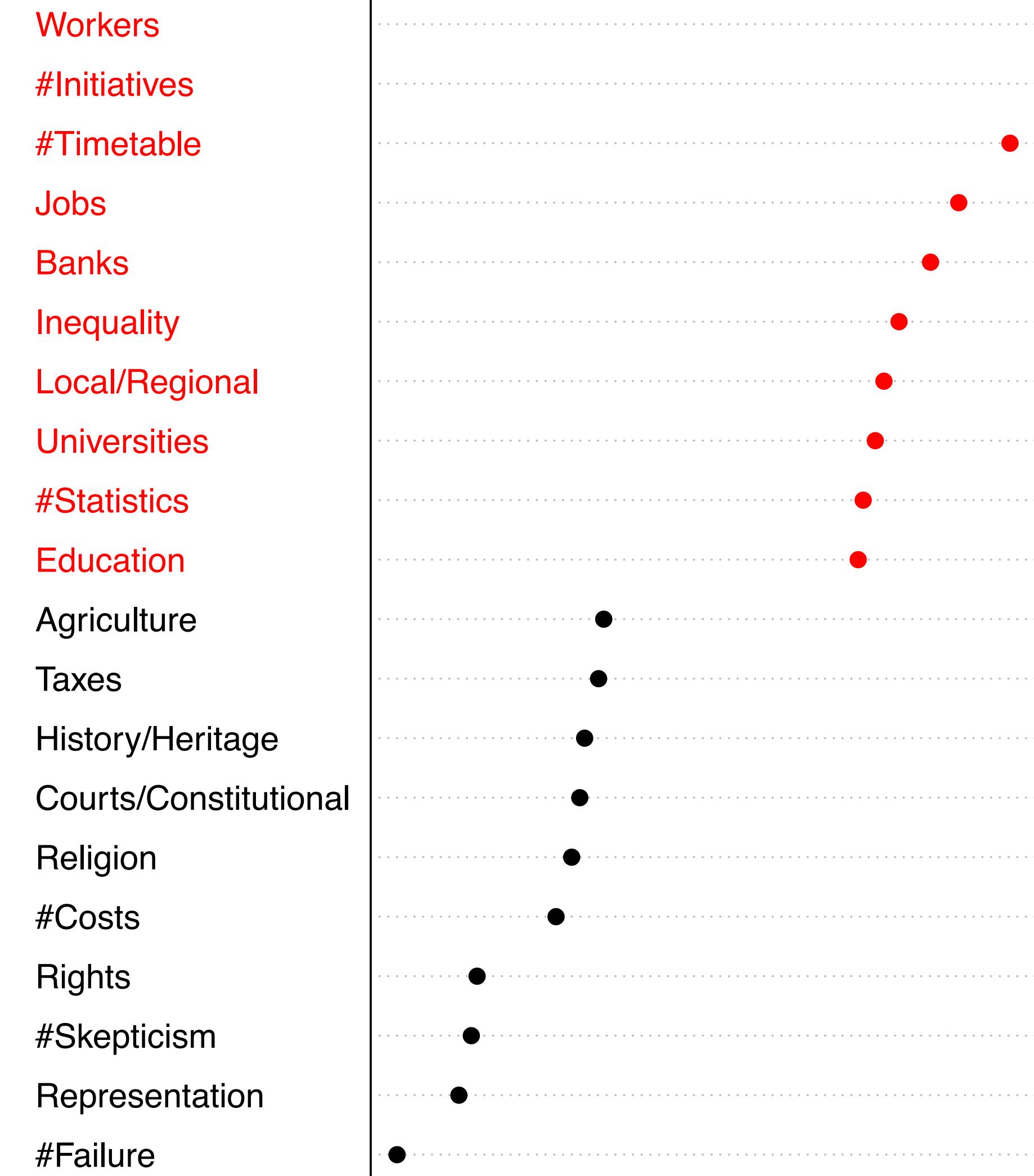
What topics do parties of particular families give most emphasis,
relative to other parties?

Conservative parties vs. others



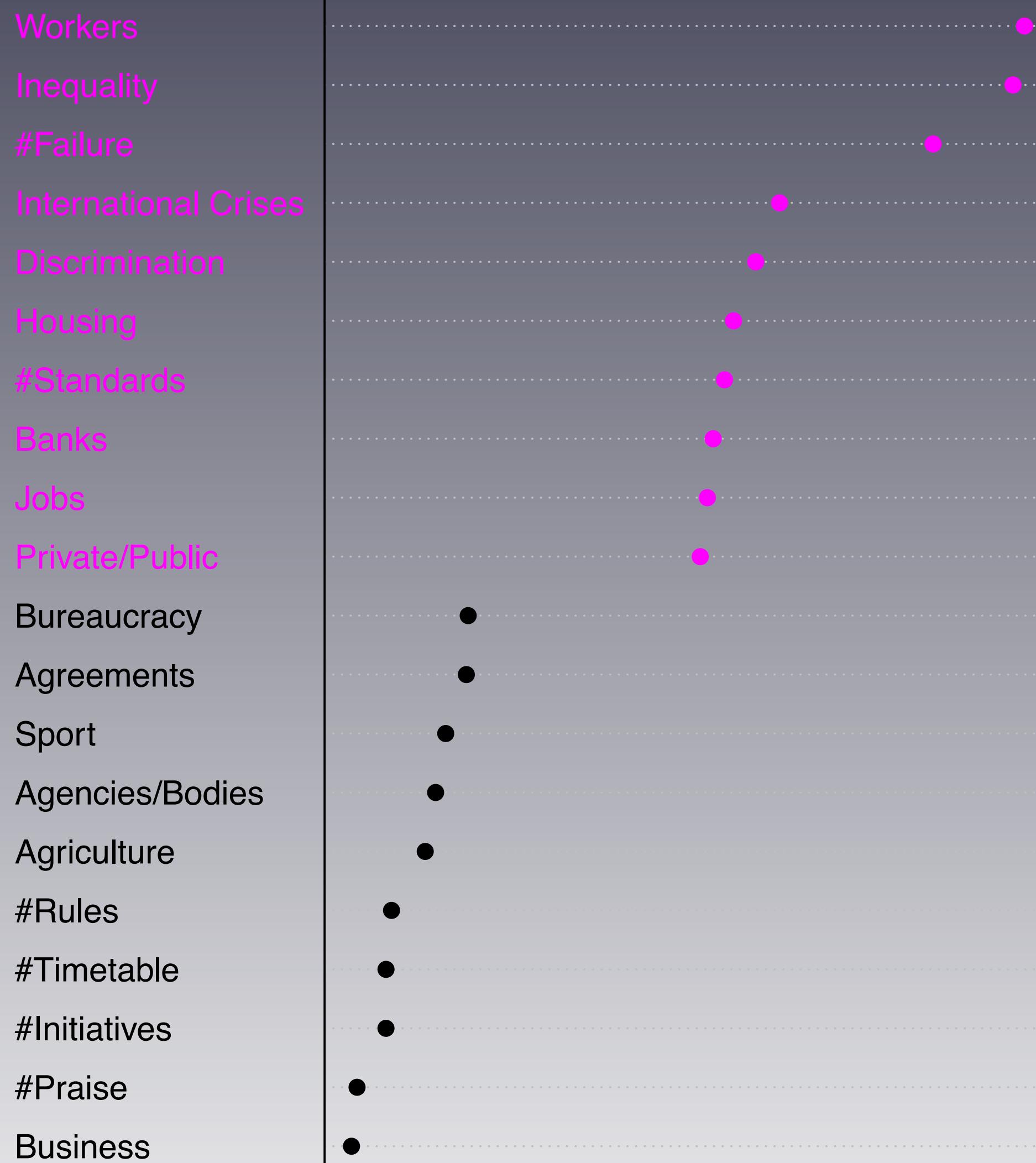
-0.05 0.00 0.05
Less attention ~ More attention

Social Democratic parties vs. others

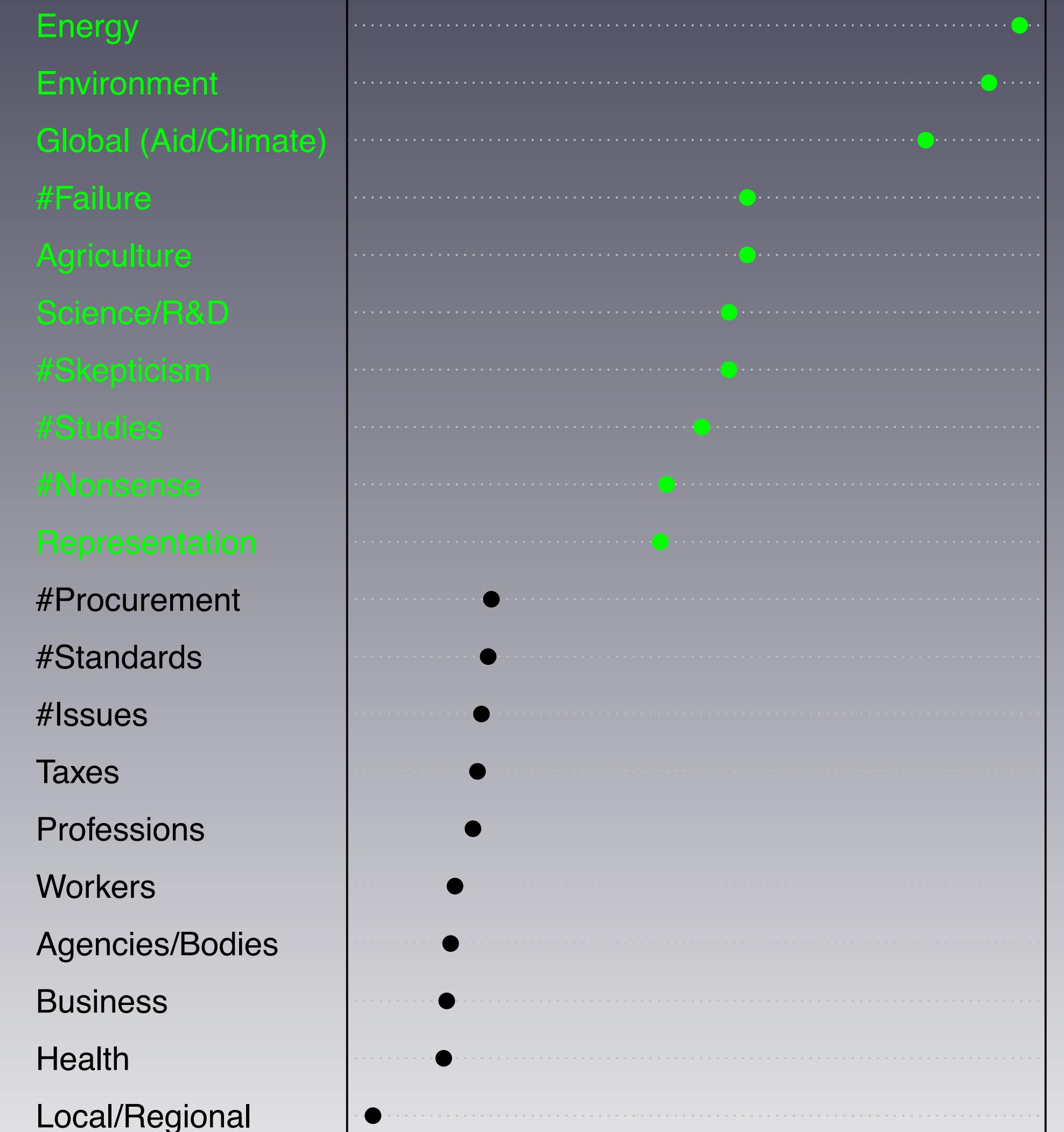


-0.05 0.00 0.05
Less attention ~ More attention

Left parties vs. others



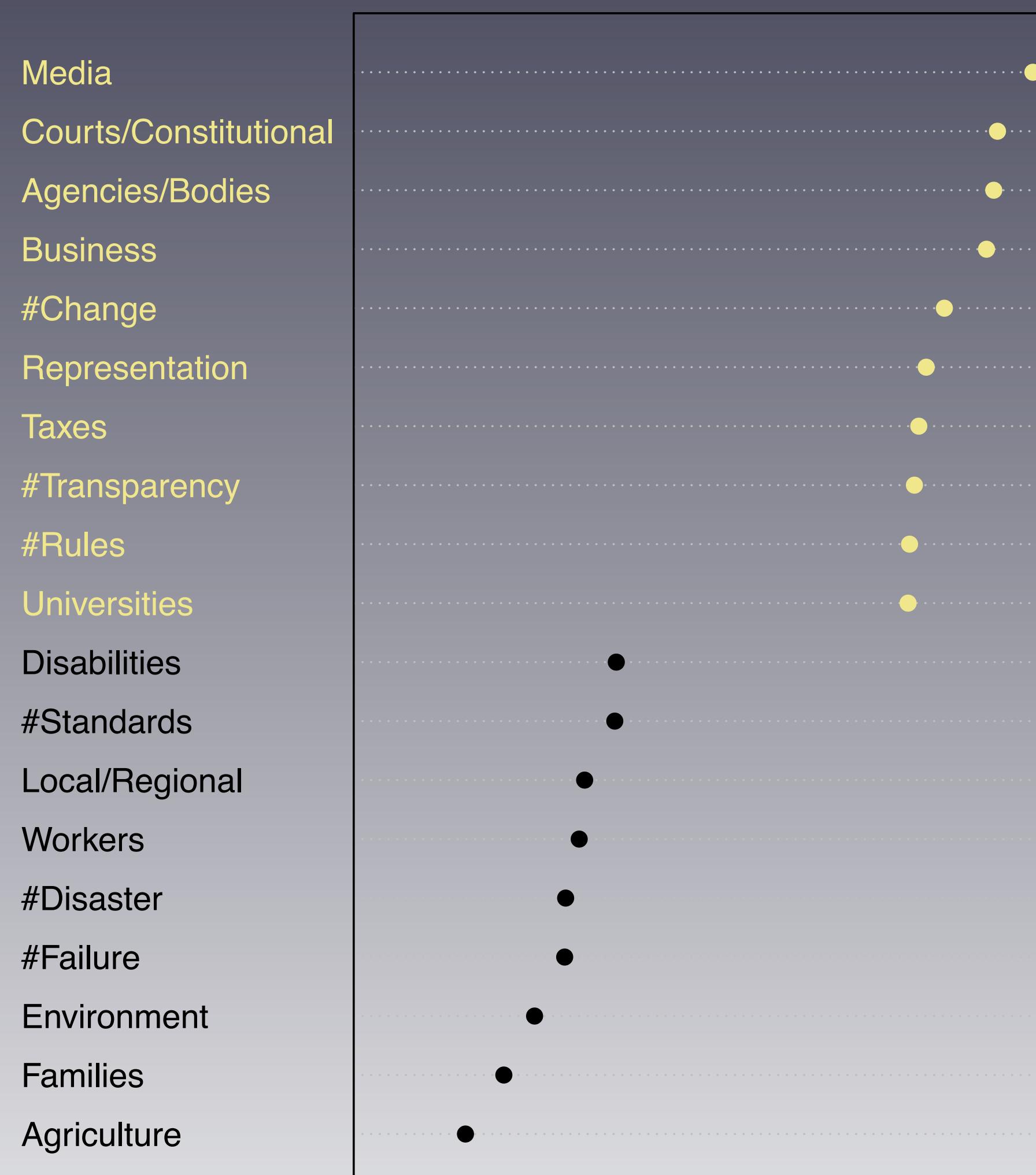
Ecological parties vs. others



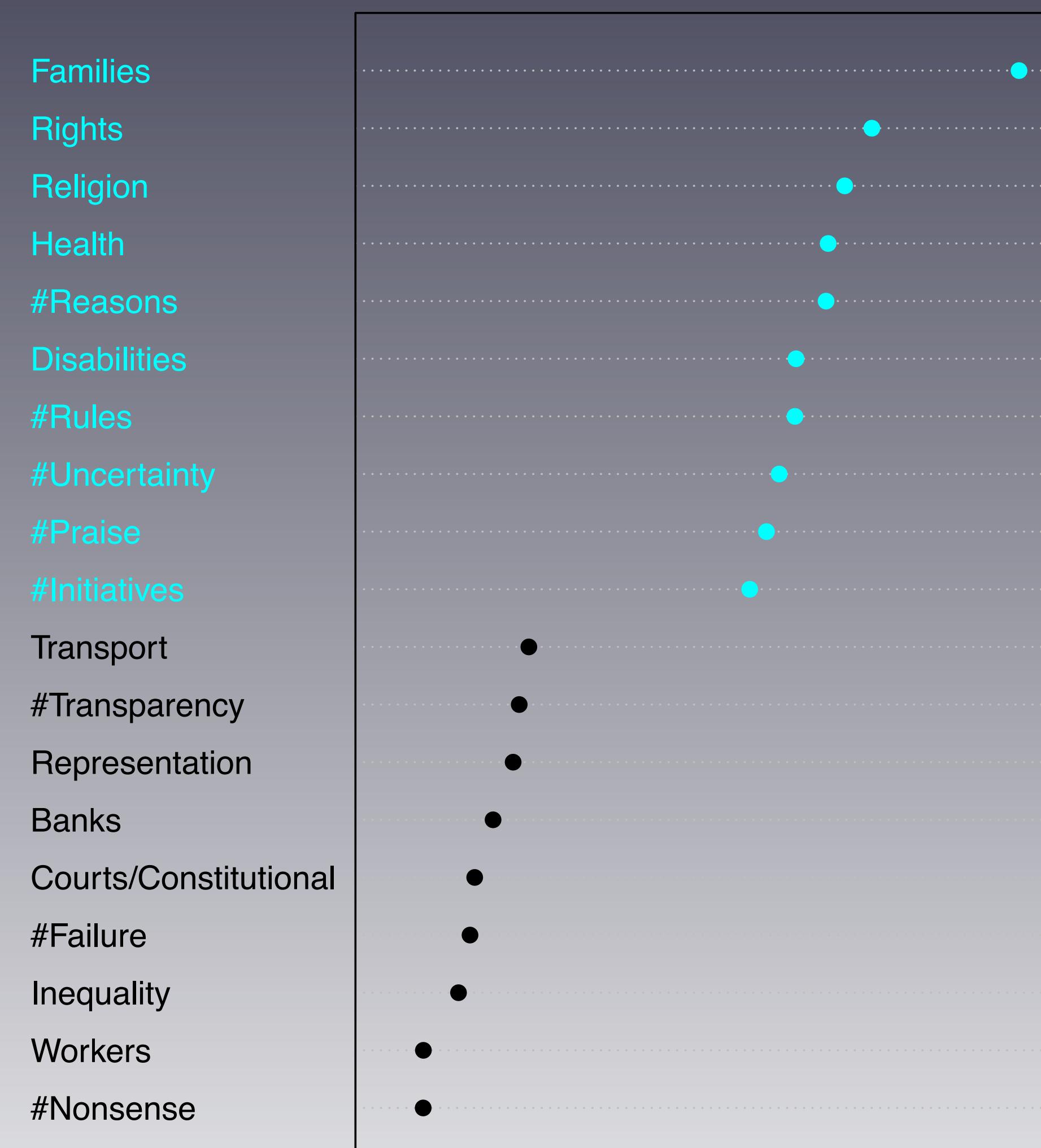
Less attention ~ More attention

Less attention ~ More attention

Liberal parties vs. others



Christian Democratic parties vs. others



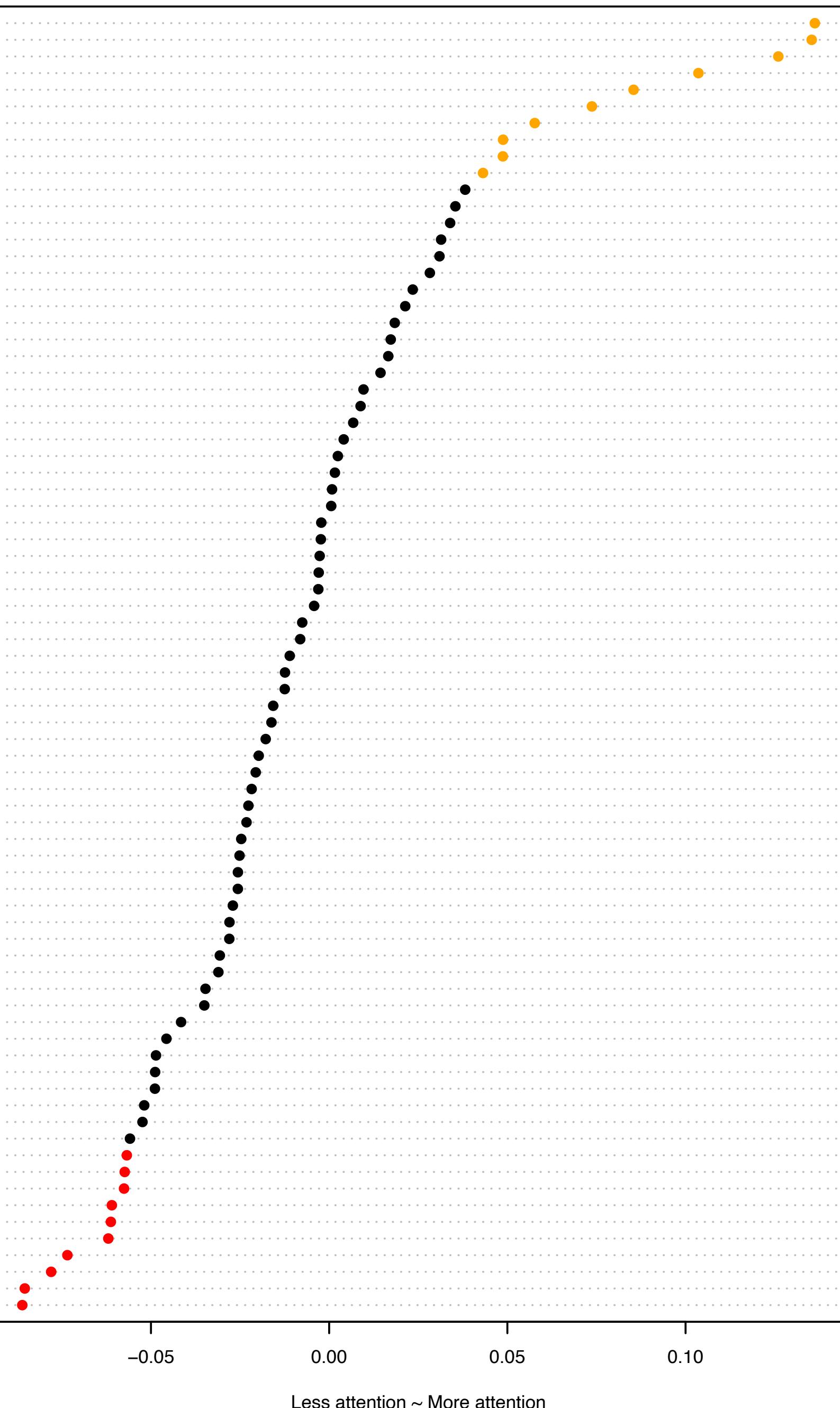
Less attention ~ More attention

Less attention ~ More attention

Nationalist

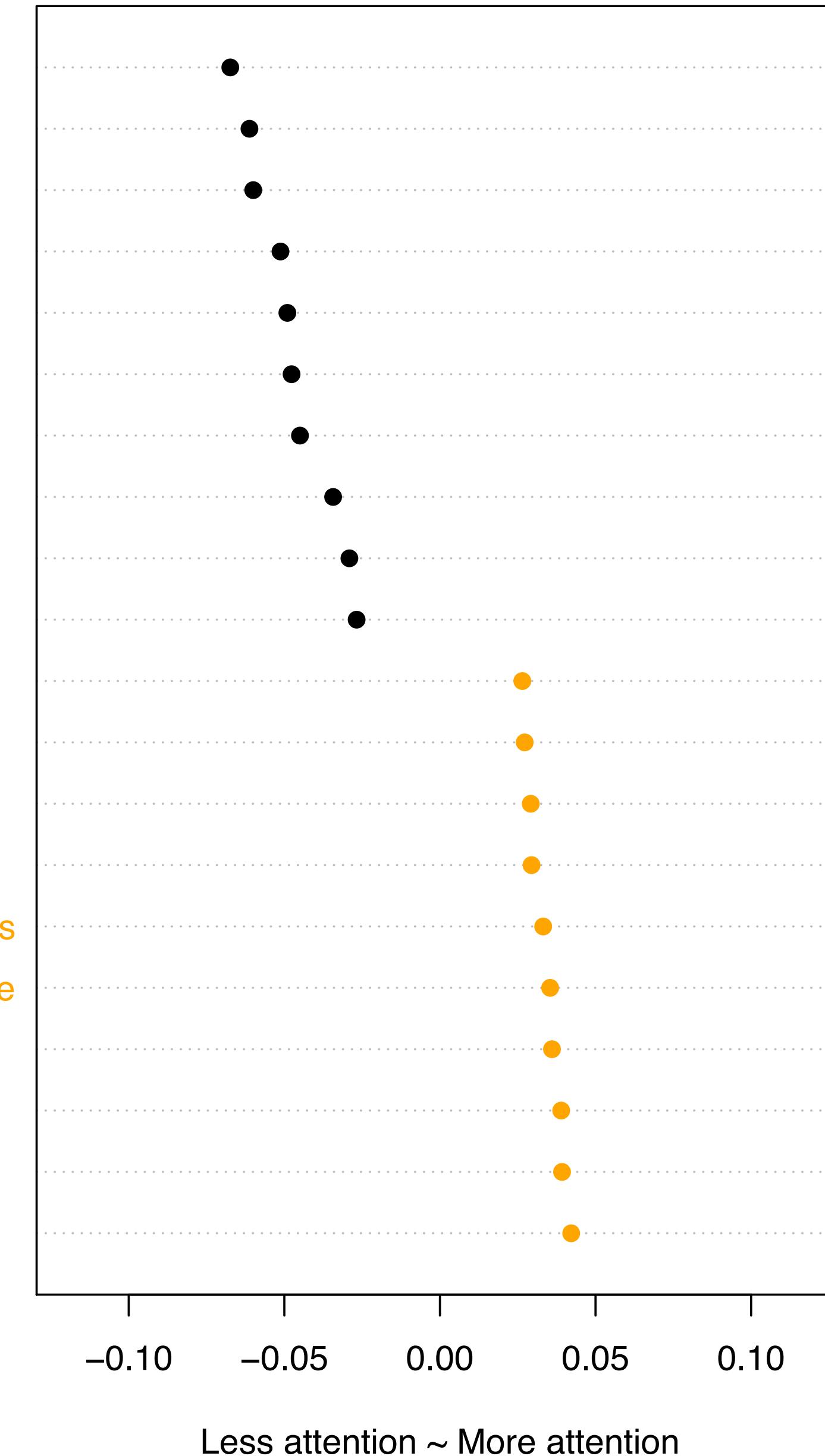
Bonus slide!

Representation
Immigration
#Failure
#Nonsense
Crime
Religion
OECD/Trade
Budget
#Quotes
#Disaster
#Skepticism
International Crises
Maritime
Professions
Public Health
Prisons
Inequality
Europe
#I_am_
#My_
Local/Regional
Housing
Defense
#Standards
Bureaucracy
Courts/Constitutional
#Uncertainty
#Consequences
#Costs
Transport
Agriculture
#Transparency
#Comparisons
Energy
Sport
Pensions
Telecom
Macroeconomy
Media
Rights
#Decisions
#Groups
Banks
#Praise
#Questions/Answers
History/Heritage
#Procurement
#Compliance
Taxes
#Issues
Terrorism
#Reasons
Private/Public
Workers
Business
Disabilities
#Problems/Solutions
Agencies/Bodies
Universities
#Statistics
Jobs
Science/R&D
Education
Families
#Change
#Rules
#Timetable
Agreements
#Deliberations
#Objectives
#Alternatives
Health
#Initiatives
Discrimination
Global (Aid/Climate)
#Studies
Environment
#Differences



Changes in Conservative party attention
when there are nationalist parties present

Conservative



We can complete cross-national party analogies

(The embeddings capture linear ideological semantics.)

Party analogies and ideological semantics (Center-right to Center-left)

		Cosine
	$V_{Lab} - V_{Con} + V_{CDU/CSU}$	is closest to  0.60
	$V_{Lab} - V_{Con} + V_{Kok}$	is closest to  0.59
	$V_{Lab} - V_{Con} + V_{CDA}$	is closest to  0.57
	$V_{Lab} - V_{Con} + V_M$	is closest to  0.49
	$V_{Lab} - V_{Con} + V_{PP}$	is closest to  0.38
	$V_{Lab} - V_{Con} + V_{FI/PdL}$	is closest to  0.27
	$V_{Lab} - V_{Con} + V_{ODS}$	is closest to  0.23

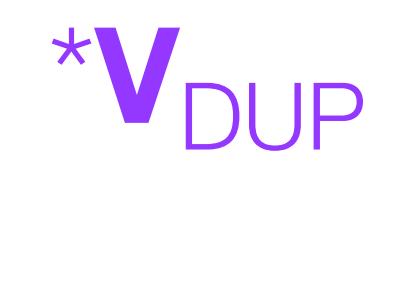
Party analogies and ideological semantics (Center-left to Left)

		Cosine
	$V_{Vas} - V_{SDP} + V_{PSOE}$	is closest to  0.63
	$V_{Vas} - V_{SDP} + V_{SAP}$	is closest to  0.49
	$V_{Vas} - V_{SDP} + V_{CSSD}$	is closest to  0.48
	$V_{Vas} - V_{SDP} + V_{PvdA}$	is closest to  0.46
	$V_{Vas} - V_{SDP} + V_{PD}$	is closest to  0.30
	$V_{Vas} - V_{SDP} + V_{SPD}$	is closest to  0.26
	$V_{Vas} - V_{SDP} + V_{Lab}$	is closest to  0.28

Party analogies and ideological semantics (Center-left to Green)

		Cosine
	$\mathbf{V}_{\text{Gruene}} - \mathbf{V}_{\text{SPD}} + \mathbf{V}_{\text{PD}}$	is closest to  0.68
	$\mathbf{V}_{\text{Gruene}} - \mathbf{V}_{\text{SPD}} + \mathbf{V}_{\text{SAP}}$	is closest to  0.50
	$\mathbf{V}_{\text{Gruene}} - \mathbf{V}_{\text{SPD}} + \mathbf{V}_{\text{Lab}}$	is closest to  0.46
	$\mathbf{V}_{\text{Gruene}} - \mathbf{V}_{\text{SPD}} + \mathbf{V}_{\text{SDP}}$	is closest to  0.24
	$\mathbf{V}_{\text{Gruene}} - \mathbf{V}_{\text{SPD}} + \mathbf{V}_{\text{PvdA}}$	is closest to  0.13
<hr/>		
	$\mathbf{V}_{\text{Gruene}} - \mathbf{V}_{\text{SPD}} + \mathbf{V}_{\text{PSOE}}$	is closest to  0.36
	$\mathbf{V}_{\text{Gruene}} - \mathbf{V}_{\text{SPD}} + \mathbf{V}_{\text{CSSD}}$	is closest to  0.34

Party analogies and ideological semantics (Center-right to Nationalist)

		Cosine
	$V_{LPF} - V_{CDA} + V_{FI/PdL}$	is closest to  0.67
	$V_{LPF} - V_{CDA} + V_M$	is closest to  0.31
	$V_{LPF} - V_{CDA} + V_{Con}$	is closest to  *0.28
	$V_{LPF} - V_{CDA} + V_{Kok}$	is closest to  *0.24
	$V_{LPF} - V_{CDA} + V_{ODS}$	is closest to  0.20
<hr/>		
	$V_{LPF} - V_{CDA} + V_{CDU/CSU}$	is closest to  0.27
	$V_{LPF} - V_{CDA} + V_{PP}$	is closest to  **0.24

Closest Party Family Mean

		ECO	LEF	SOC	LIB	CHR	CON	NAT	ETH
		sv_MP, it_FdV, fi_Vihr, uk_Green, nl_PvdD, de_Gruene	nl_GL						
ECO									
LEF		sv_V, es_IU, it_PRC, it_PdCl, de_Linke, it_SEL, fi_Vas, nl_SP, cs_KSCM, it_RnP							
SOC			uk_Lab, fi_SDG, de_SPD, sv_SAP, nl_PvdA, it_MDP, es_PSOE, cs_CSSD, it_PD						
LIB	cs_CMUS	it_P-UDEUR		it_M5S, fi_LKP, it_IdV, es_UPyD, nl_D66, cs_VV, uk.LibDem, cs_US, fi_NUOR	sv_L	nl_VVD, it_SC, de_FDP, cs_ANO		cs_ODA	
CHR				it_AP/NCD, v_KD, nl_SGP, it_PI/DES- CD, cs_KDU-CSL, fi_KD, nl_CU, it_UdC, cs_TOPO9, cs_KDS		nl_CDA, de_CDU/CSU, es_CDS	it_DCA-NPSI		
CON						es_PP, uk_Con, it_FI/PdL, sv_M, fi_Kok, cs_ODS, it_FLI, it_PT			
NAT				uk_UKIP			nl_PVV, nl_LPF, sv_SD, sv_NyD, it_AN/Fdl, it_LN, fi_PS, cs_SPR-RSC, cs_Usvit, cs_LIDEM		
ETH				es_CiU		uk_SNP		uk_SDL, uk_DUP, es_PNV, fi_RKP, es_ERC, uk_Plaid	

Coherence of Party Families

“Liberal” is by far the least coherent family, with six members estimated by our data to be closer to five other party families.



The center-right “Conservative” grouping absorbs six parties from the Liberal and Christian Democratic families.

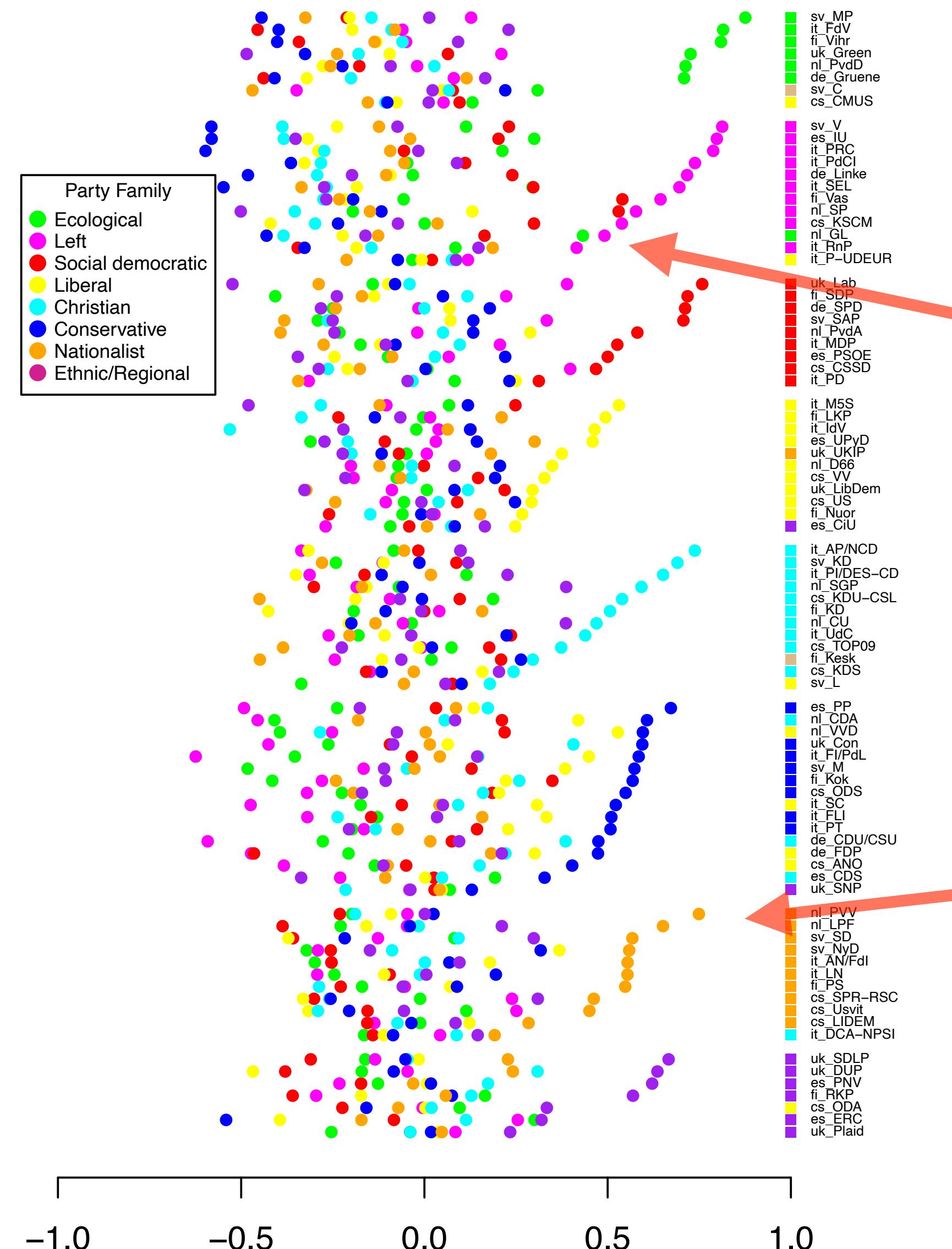
Coherence of Party Families

Each party is listed down the right with a dot colored according to its CMP party family label.

Dots in graph indicate cosine similarity of each party vector with each party family mean.

Dots out of place in the color groupings of dots are closer to a different party family than their label.

Cosine Similarities with Party Family Means



Dutch GL (Green Left) is a close call between “Green” and “Left.”

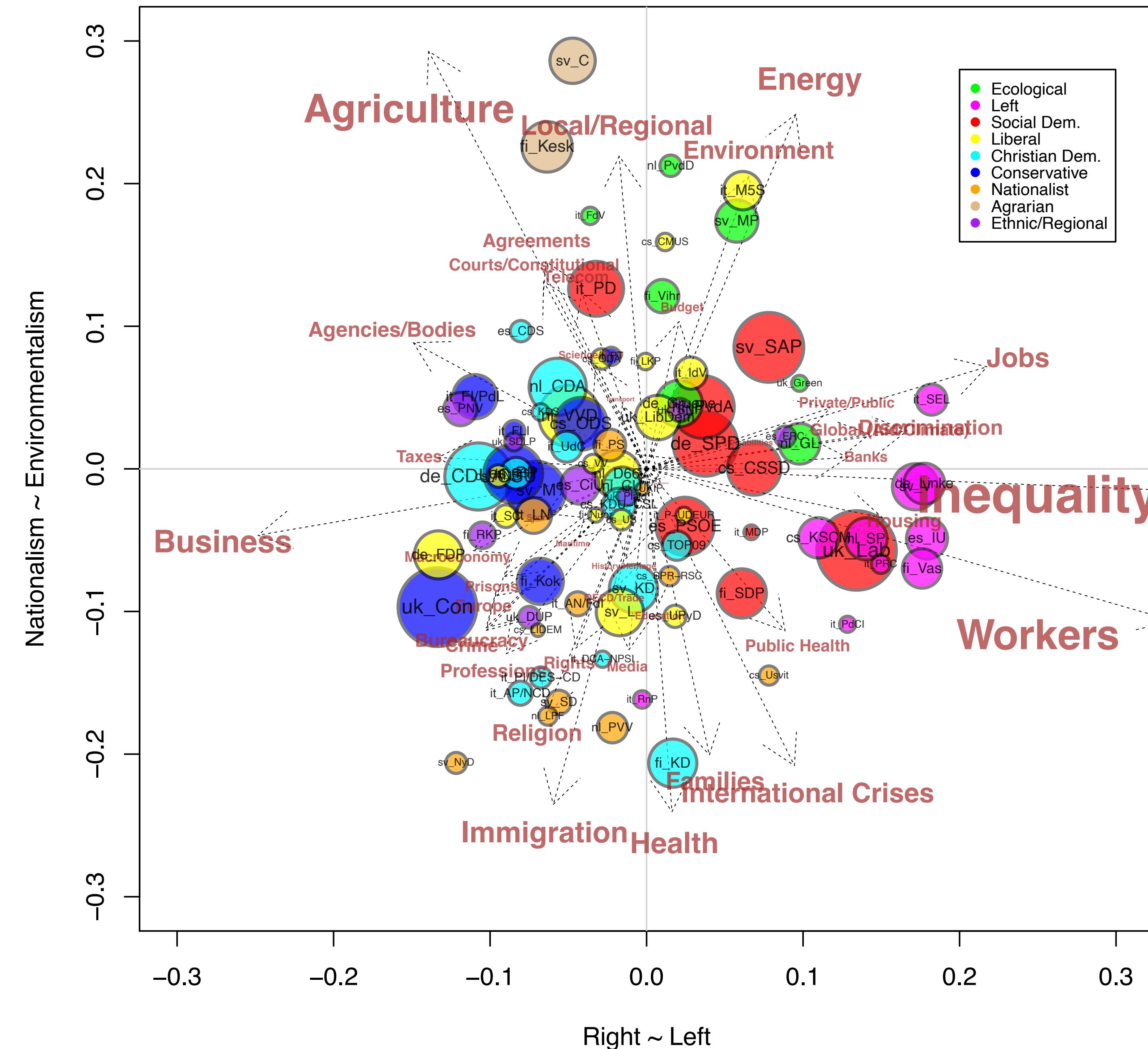
Dutch PVV (Party for Freedom) is *clearly* in the nationalist family.

Linear ideological semantics implies a meaningful ideological space.

Our embeddings contain a topic model *and* a spatial scaling model.

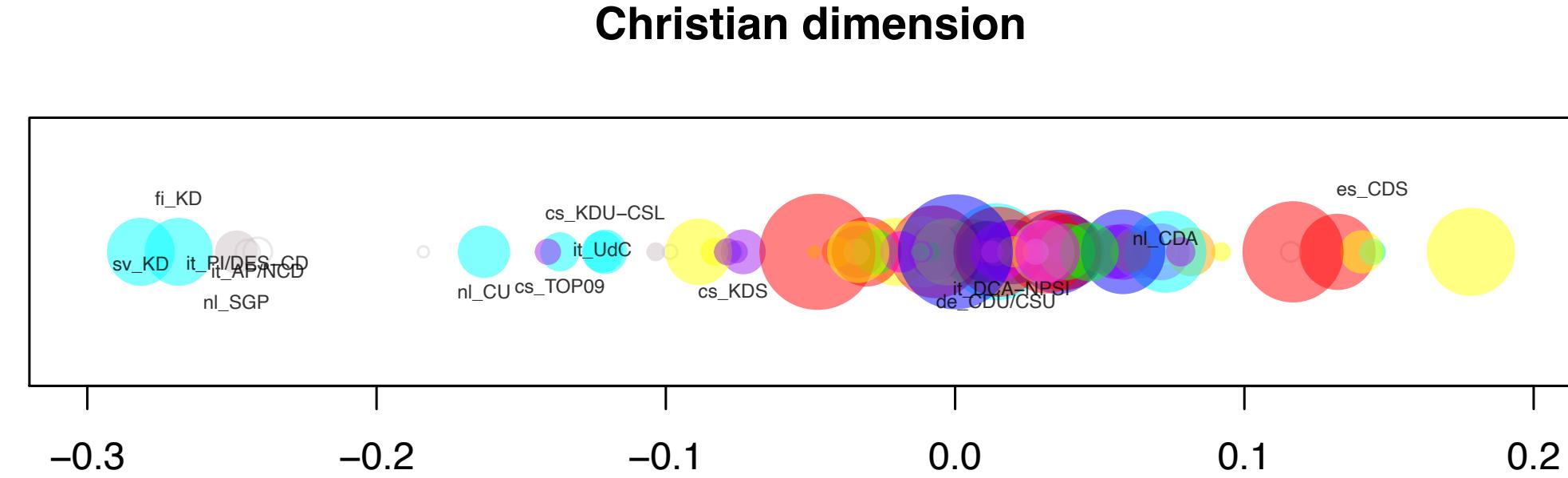
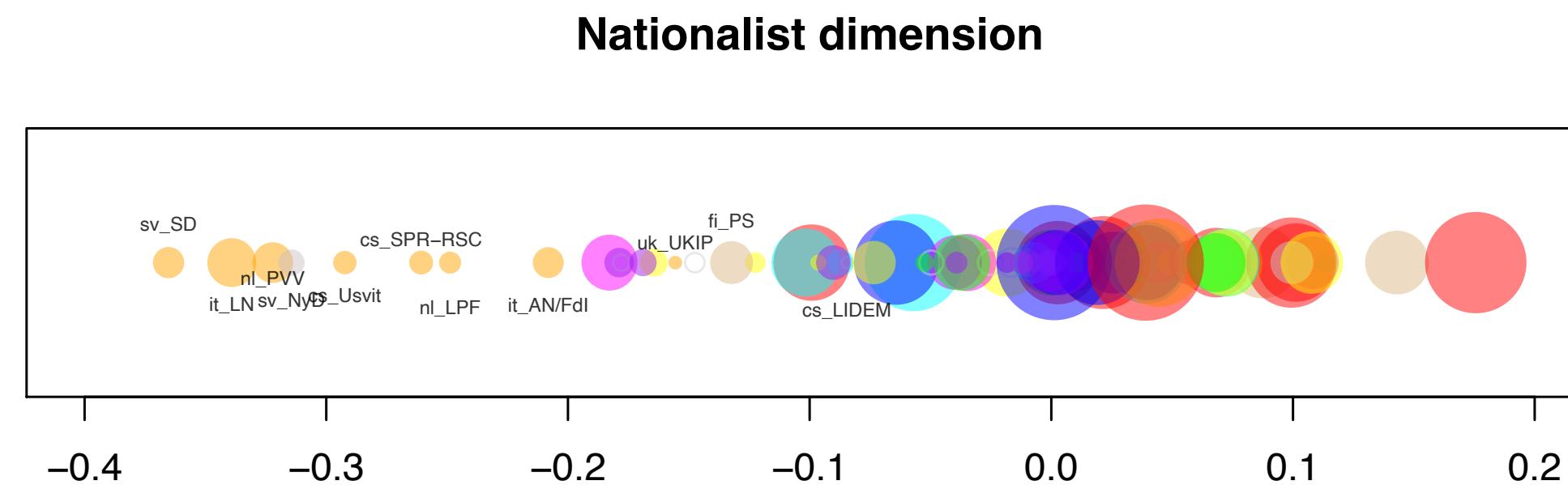
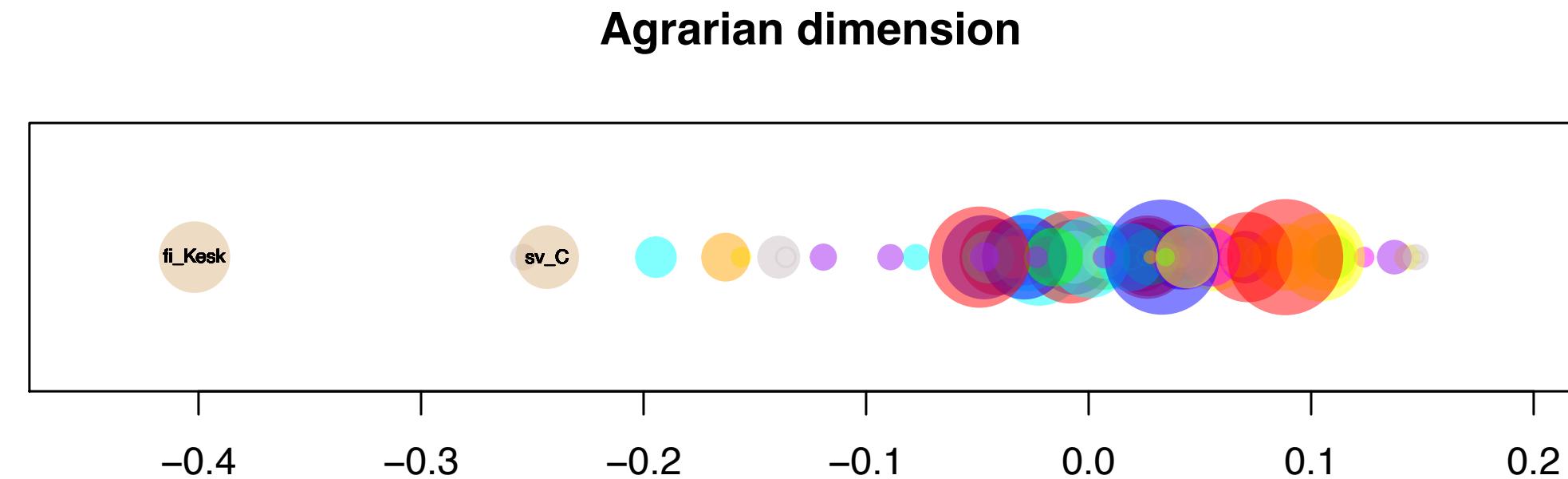
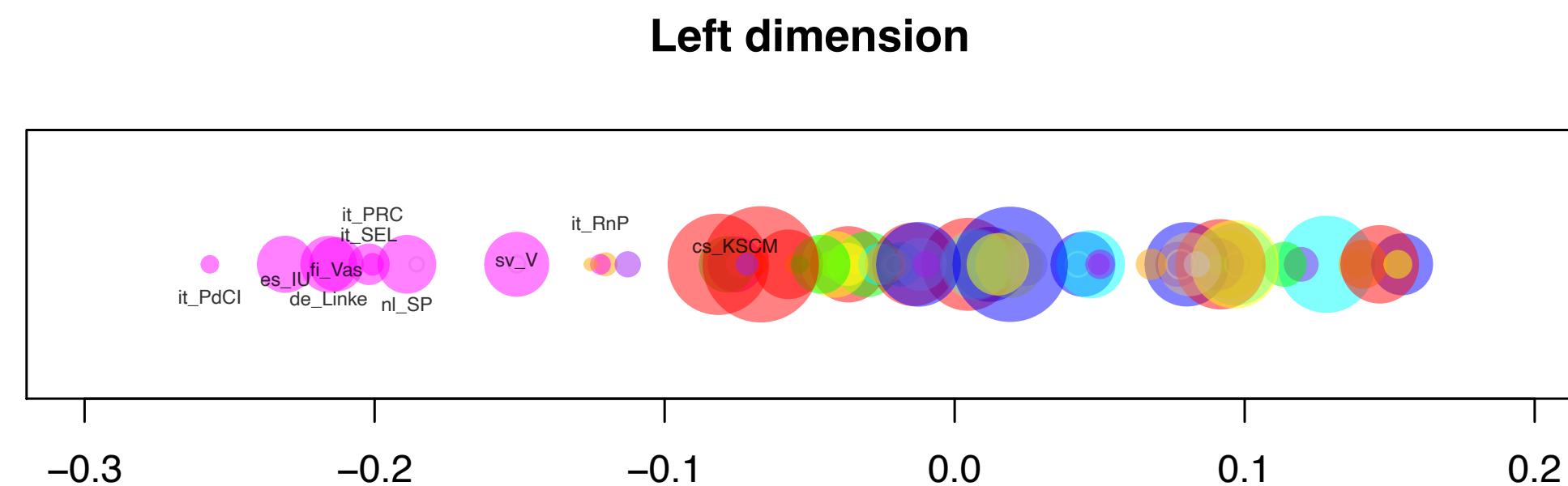
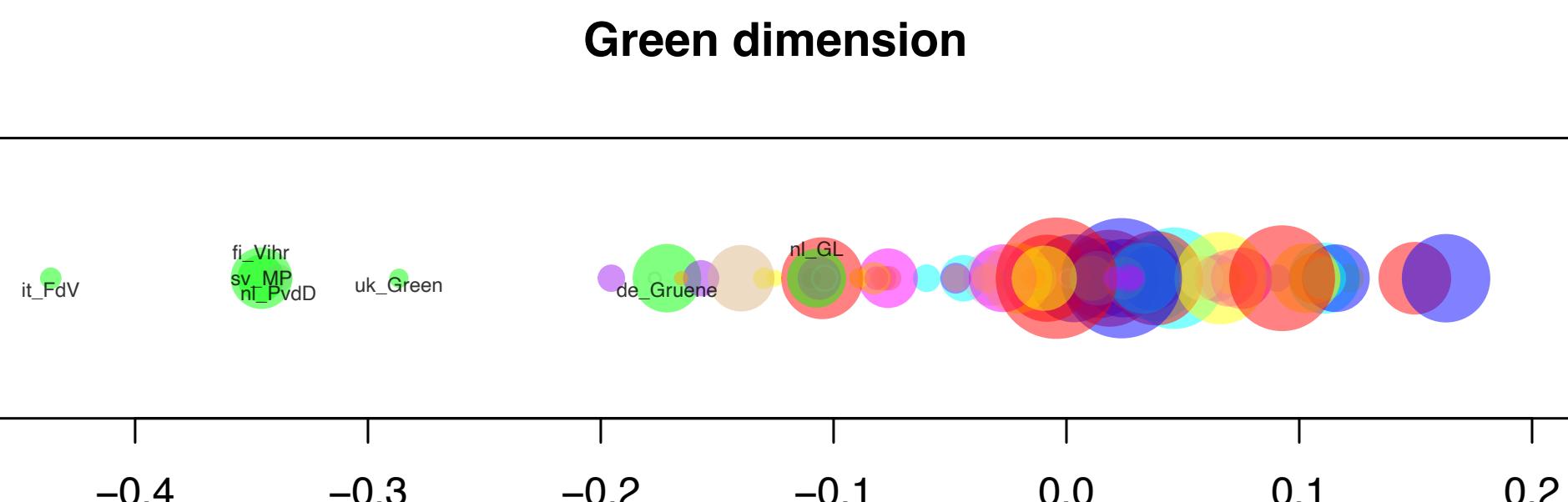
Weighted SVD of the Party Scores matrix reveals a scaling mode

Weighted SVD of Mean Party Scores



Varimax rotation isolates the most distinctive ideological party families in the first five dimensions

Projections onto rotated dimensions



Caveats

- The setting of parliamentary speech, as we saw, shares a lot of common structure, both in topical scope and procedural formality. It remains an open question how well this will generalize to other settings (like social media).
- This worked with eight European languages. Finnish is a semi-outlier in this set (Uralic rather than Indo-European), but presumably the commonality of structure would degrade if attempted across more disparate language sets.
- As with factor analysis, topic modeling, etc., there are identifying assumptions. We must be cautious in particular about contrasting relative dispersion in party scores across countries. (E.g., the UK Conservative and Labour Parties are estimated as the most extreme “conservative” and “social-democratic” parties, but this may be partially due to the preponderance of seats being held by these two parties, a situation not observed elsewhere.)

Thanks!

burtmonroe@psu.edu

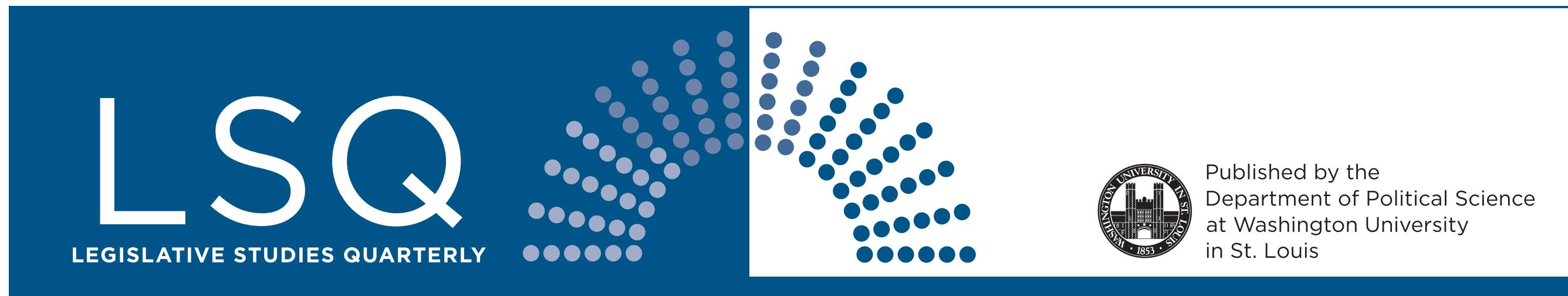
Machine Translation

Single language topic models

Top keywords for example topics, Knesset

Negotiations	Jobs	Religion / Identity	Education	Housing/Immigration
אוסלו	העבודה	בשבת	חינוך	הקליטה
ערפאת	וهرווחה	המשטרה	הספר	דירות
הפלשטים	עובדים	הדותות	והתרבות	דיור
ירושלים	התעסוקה	השואה	המורים	העלייה
השלום	שכר	היהדות	התלמידים	העלולים
Oslo	work	Saturday	education	absorption
Arafat	welfare	police	school	apartments
Palestinians	workers	religions	culture	housing
Jerusalem	employment	Holocaust	teachers	immigration
peace	wages	Judaism	students	immigrants

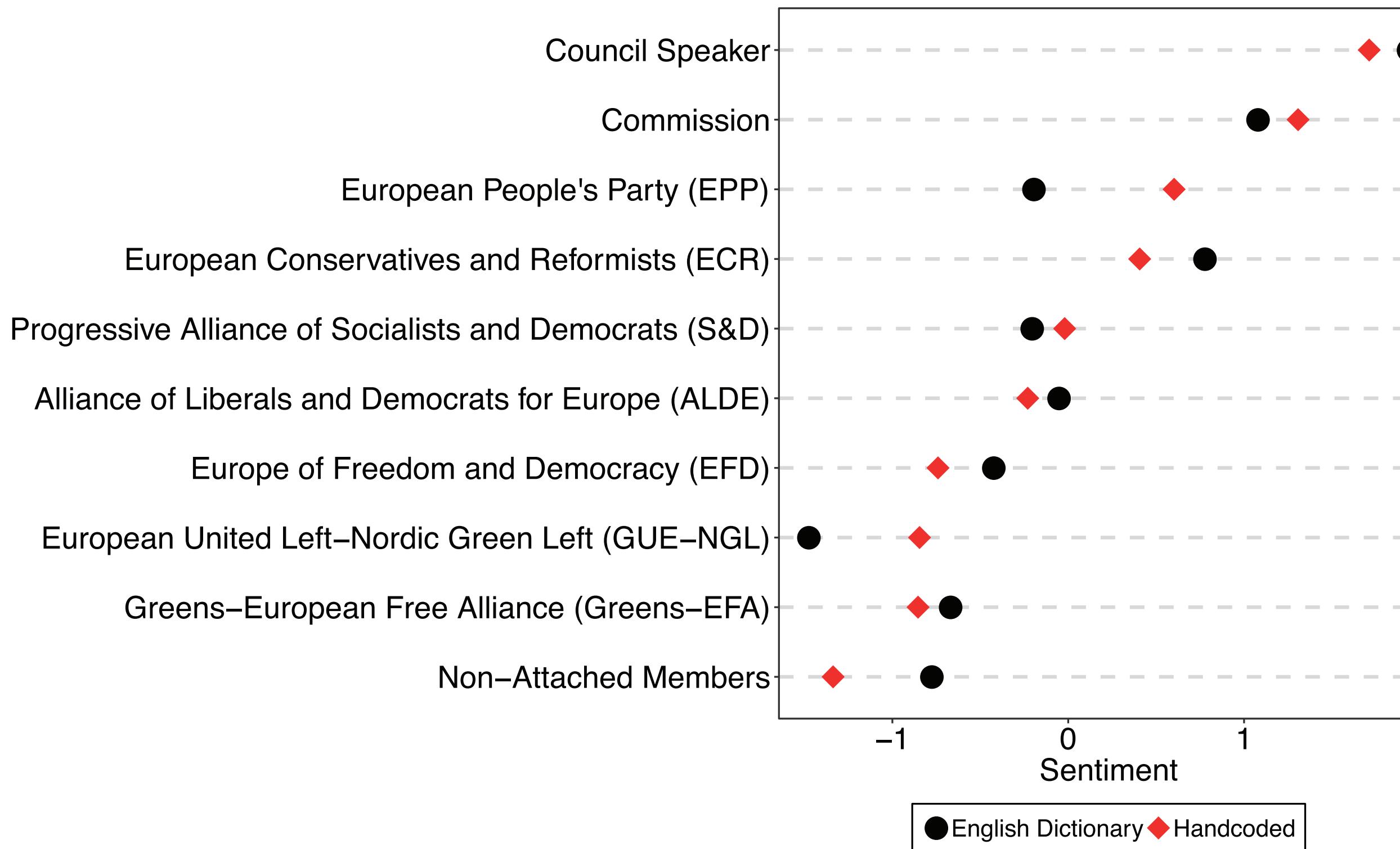
Translating sentiment dictionaries



SVEN-OLIVER PROKSCH
University of Cologne
WILL LOWE
Princeton University
JENS WÄCKERLE
University of Cologne
STUART SOROKA
University of Michigan

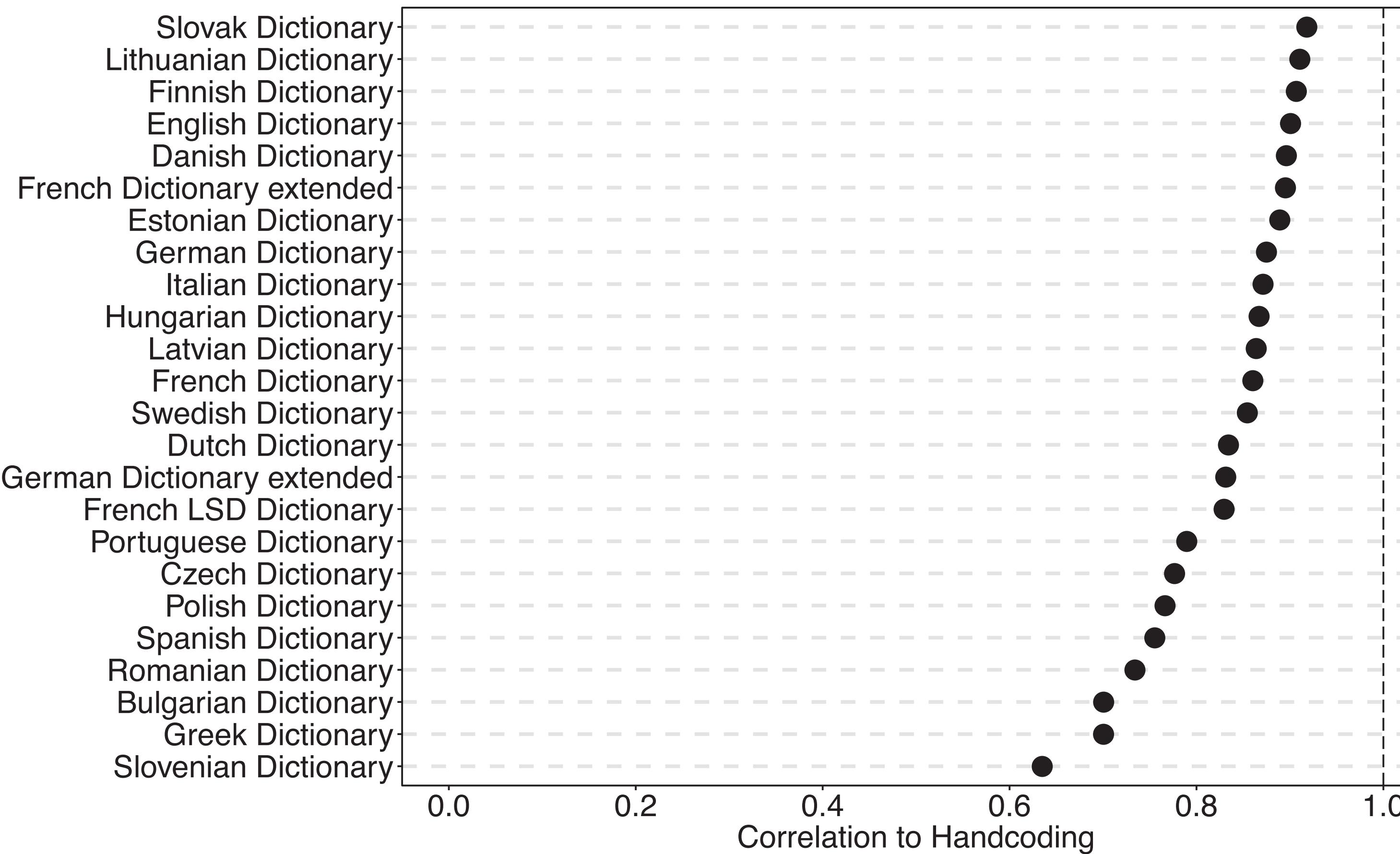
*Multilingual Sentiment Analysis: A
New Approach to Measuring Conflict
in Legislative Speeches*

FIGURE 1
 Expressed Sentiment in the State of the Union Debate in the European Parliament (2010) [Colour figure can be viewed at wileyonlinelibrary.com]



Note. The plot shows the expressed sentiment on the automated approach (English Lexicoder dictionary) and on hand-coded estimates for the relevant speaker groups (Council, Commission, and EP political groups). Because the two measures are on different scales (log ratio of counts and five-point Likert scale, respectively), we display z-transformed scores.

FIGURE 2
Correlation of Sentiment to Hand Coding in State of the Union
Debate in the European Parliament 2010 Using Translated
Lexicoder Dictionaries



Note. The “French LSD Dictionary” is from Duval and Petry (2016).

FIGURE 3
Correlation of Sentiment to Hand Coding in State of the Union
Debate in the European Parliament 2010 Using Translated
SentiWS Dictionaries

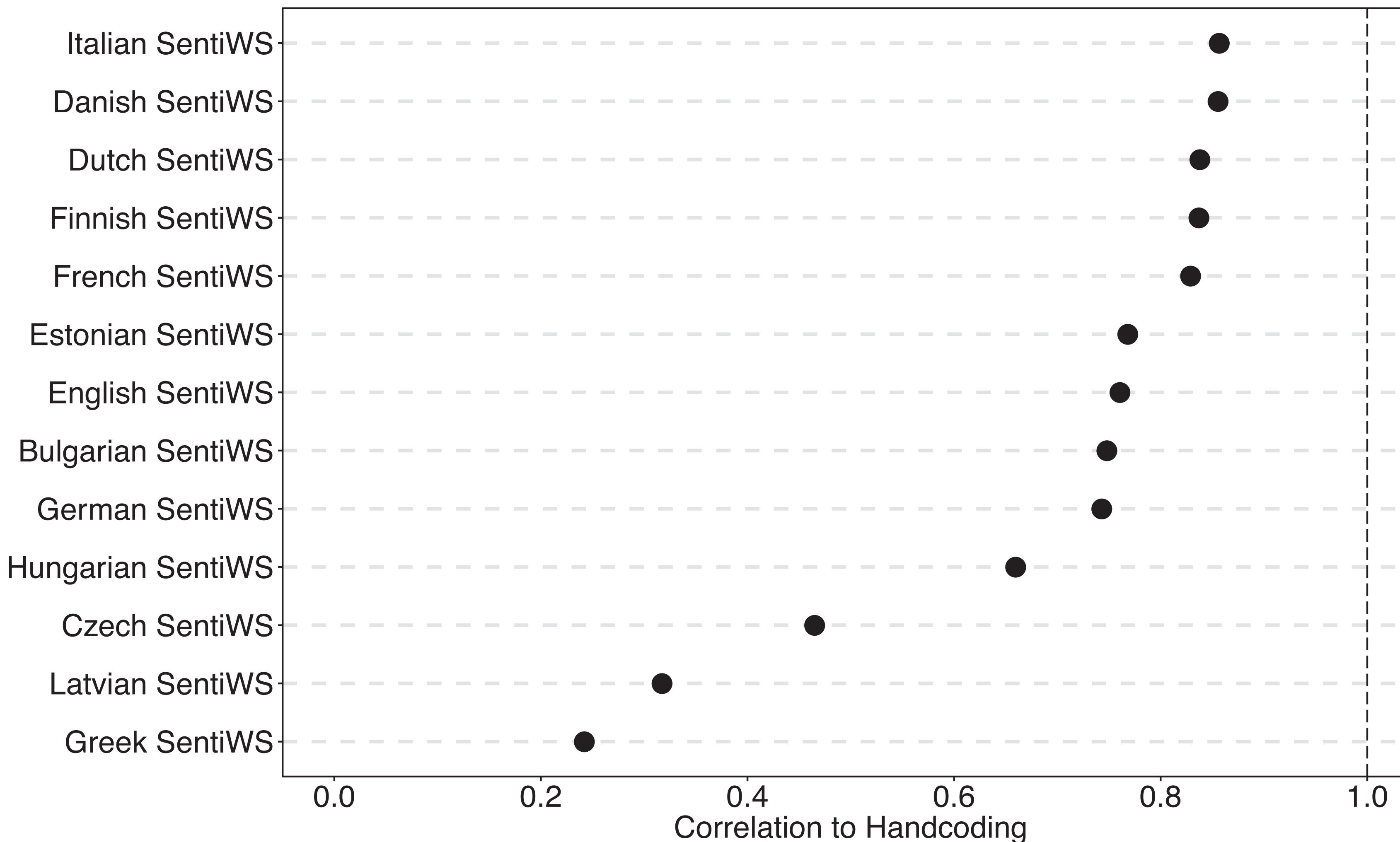


FIGURE 4
 Sentiment in Irish Budget Debates, 2008–12 [Colour figure can
 be viewed at wileyonlinelibrary.com]

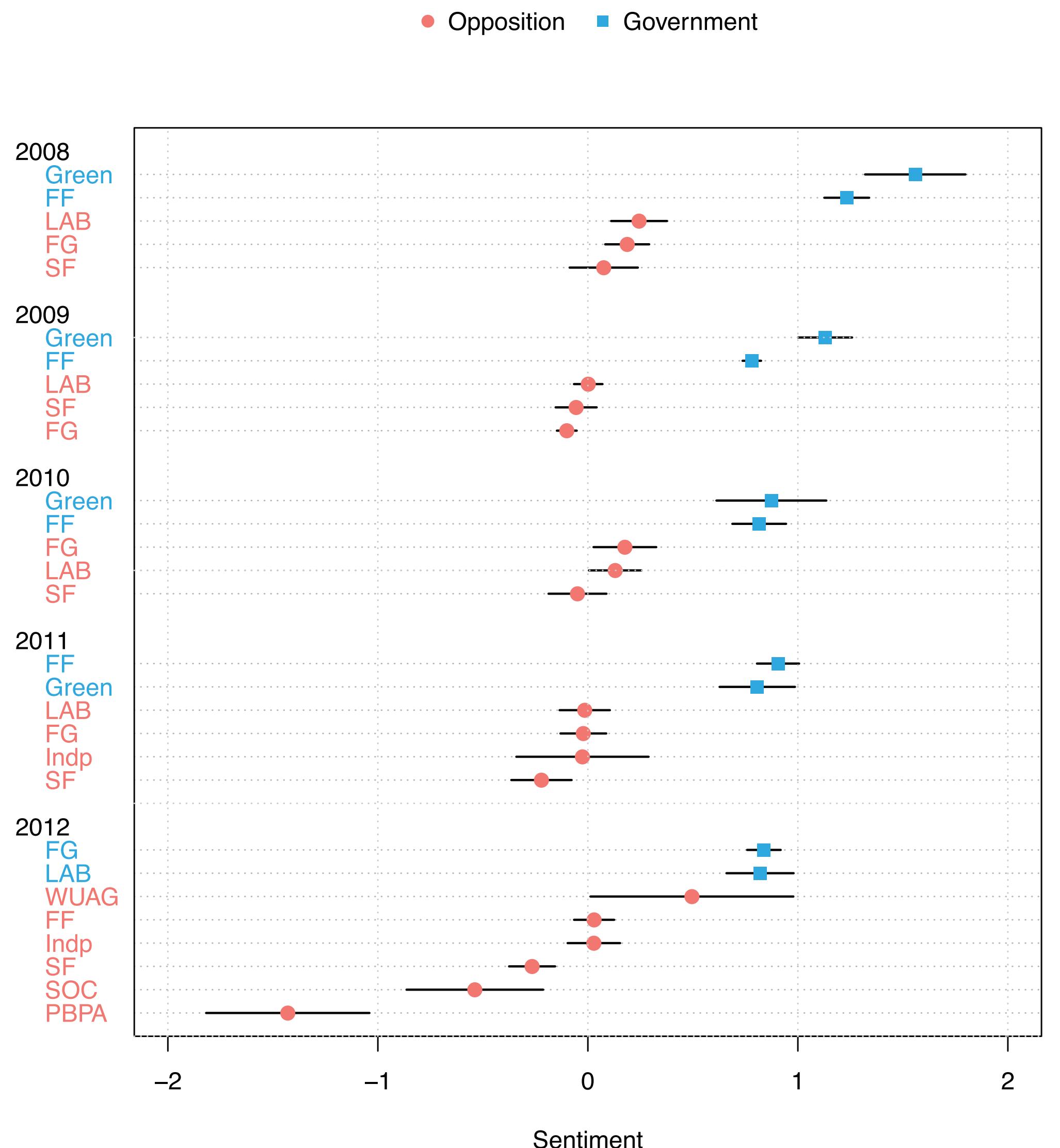
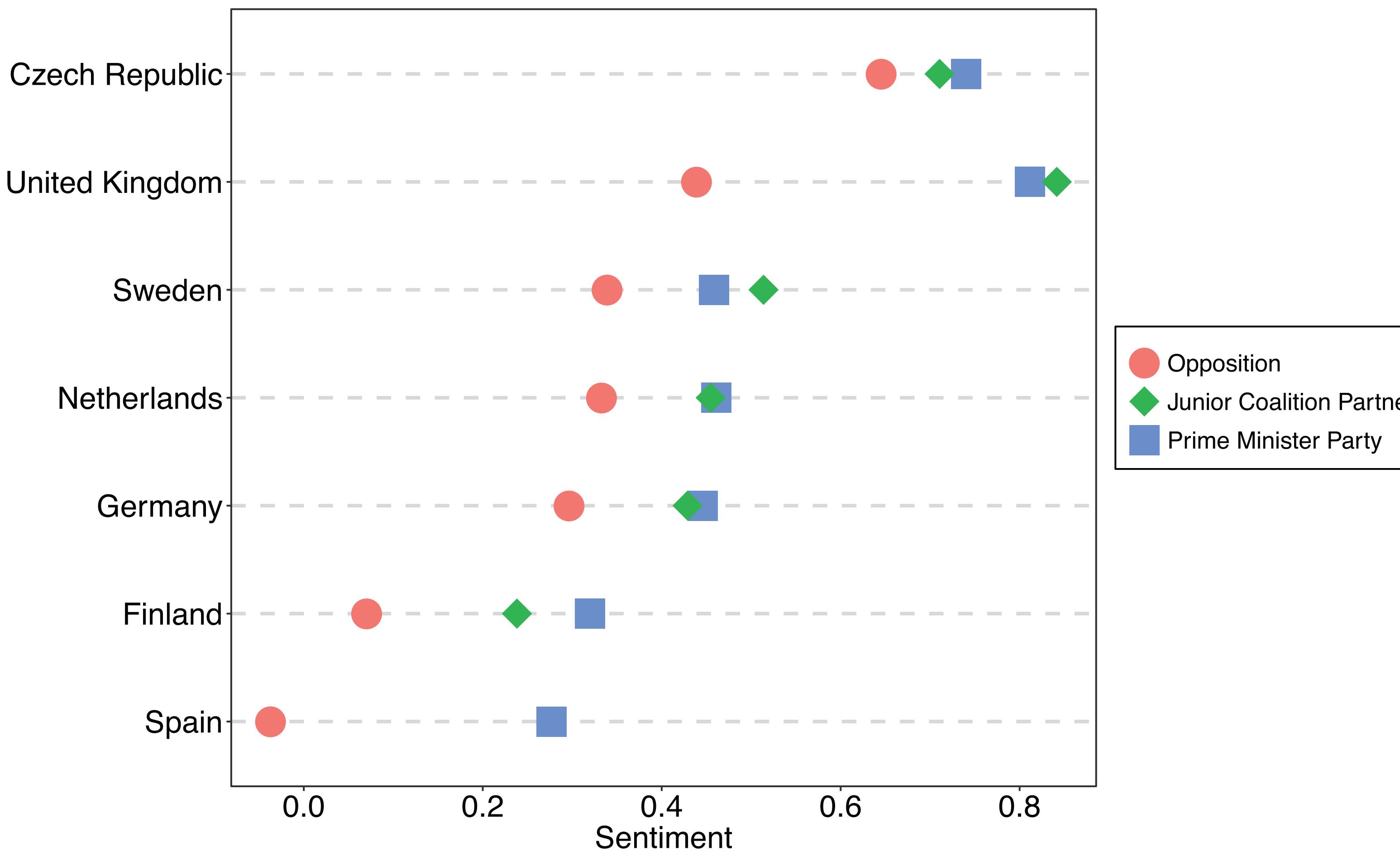


FIGURE 8
Prime Minister Party, Junior Coalition Partner, and Opposition
Sentiment in European Parliaments [Colour figure can be
viewed at wileyonlinelibrary.com]



Bag-of-words generally?

PUBLISH

ABOU

PLOS ONE



OPEN ACCESS



PEER-REVIEWED

RESEARCH ARTICLE

Automated content analysis across six languages

Leah Cathryn Windsor , James Grayson Cupit, Alistair James Windsor

Published: November 20, 2019 • <https://doi.org/10.1371/journal.pone.0224425>

Source English

The Council expresses concern regarding reports of involvement of some officers of the HNP in serious crimes and human rights violations and stresses the need to investigate fully any allegations of serious crimes and human rights violations

Language

Human Translation

Machine Translation

Russian

Совет выражает озабоченность в связи с сообщениями об участии некоторых служащих ГНП в совершении серьезных преступлений и нарушений прав человека и подчеркивает необходимость проведения всестороннего расследования всех обвинений в совершении серьезных преступлений и нарушений прав человека

The Council expresses its concern at reports about the involvement of some officers of the HNP in serious crimes and human rights violations, and stresses the need for a full investigation of all allegations of serious crimes and human rights violations

French

Il se déclare préoccupé par les informations selon lesquelles certains policiers auraient participé à des crimes graves et à de graves violations des droits de l'homme, et souligne qu'il est indispensable que toutes les allégations de crime grave et de violation grave des droits de l'homme fassent l'objet d'une enquête approfondie

It expresses concern at reports that some police officers allegedly involved in serious crimes and serious violations of human rights, and emphasizes that it is essential that all allegations of serious crime and gross violation of human rights are the subject of a thorough investigation

Arabic

ض عبد طرور تب ديفة يـتـلا رـيـراـقـتـلا عـازـاـ هـقـلـقـ نـعـ سـلـجـمـلا بـرـعـيـوـ
ـتـاـكـاهـتـذـاـوـ قـرـيـطـخـ مـهـارـجـ يـفـ ةـيـتـيـاهـذـاـ ةـيـنـطـوـلـاـ ةـطـرـشـذـاـ طـابـضـ
ـيـأـ يـفـ ةـلـمـاـكـ ةـرـوـصـبـ قـيـقـحـتـذـاـ ةـرـوـرـضـ دـكـؤـيـوـ ،ـ نـاسـذـلـاـ قـوـقـحـدـ
ـنـاسـذـلـاـ قـوـقـحـدـ تـاـكـاهـتـذـاـوـ قـرـيـطـخـ مـهـارـجـ بـاـكـتـرـابـ تـاءـاعـدـ.

The Council expresses its concern at reports of the involvement of some of the Haitian National Police officers in serious crimes and human rights violations, and stresses the need to investigate fully any allegations of serious crimes and human rights violations

Chinese

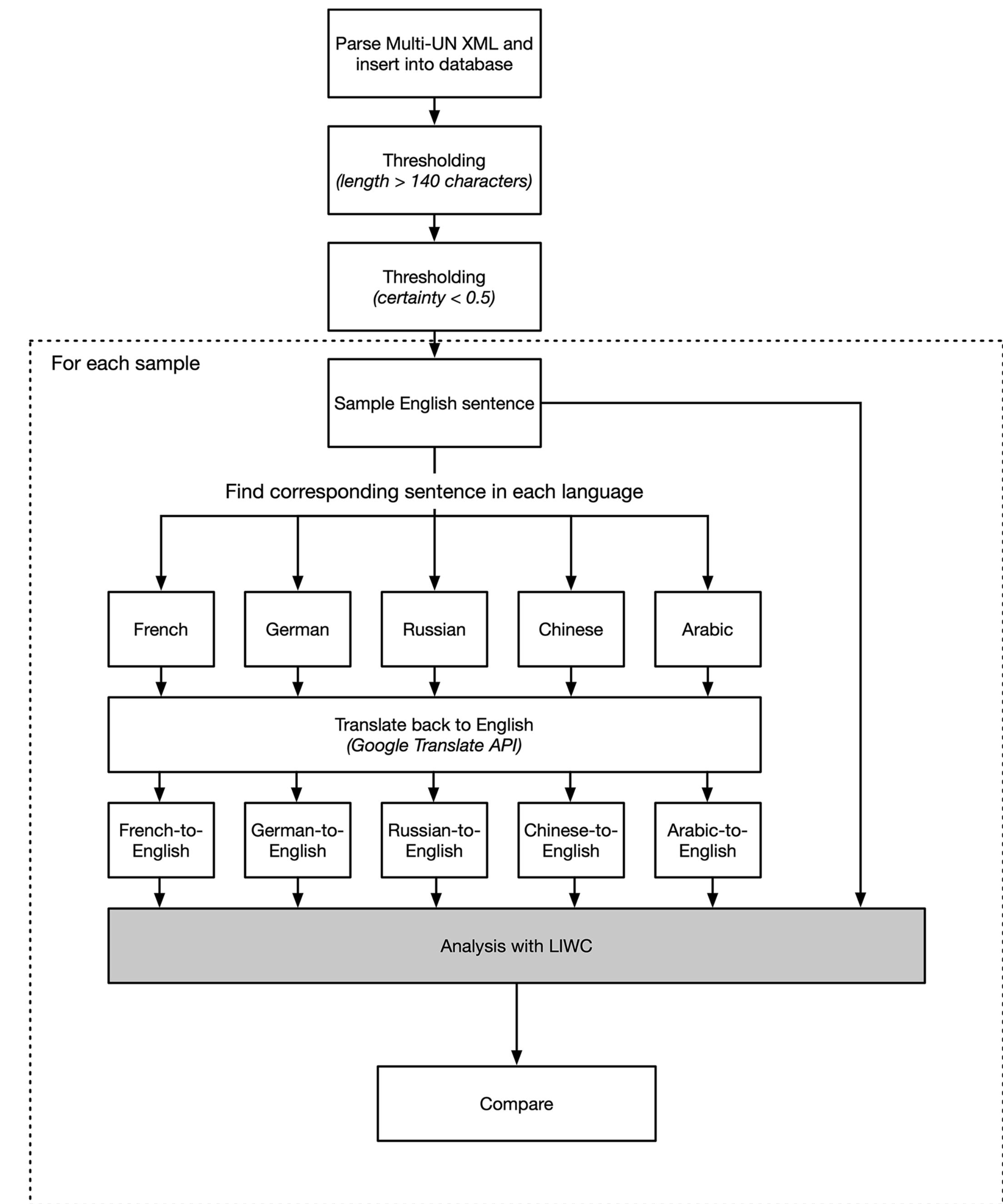
安理会对关于国家警察的一些成员参与严重犯罪和侵犯人权行为的报告 表示关切，并强调需要彻底调查任何关于严重犯罪和侵犯人权的指控。

Council report on the number of members of the National Police involved in serious crimes and human rights violations expressed concern and stressed the need to thoroughly investigate any allegations of serious crimes and human rights violations

German

Der Rat bekundet seine Besorgnis angesichts von Berichten, wonach Mitglieder der Haitianischen Nationalpolizei an schweren Verbrechen und Menschenrechtsverletzungen beteiligt gewesen sein sollen, und betont die Notwendigkeit, alle Anschuldigungen über schwere Verbrechen und Menschenrechtsverletzungen umfassend zu untersuchen

The Council expresses its concern at reports that members of the Haitian National Police to have been involved in serious crimes and human rights violations, and stresses the need to investigate all allegations of serious crimes and human rights violations comprehensively



LIWC Category	Language translated from					Mean
	Arabic	German	French	Russian	Mandarin	
All	0.831	0.814	0.822	0.843	0.783	0.820
Summary	0.863	0.833	0.856	0.906	0.761	0.844
Linguistic Dim.	0.729	0.728	0.769	0.788	0.651	0.733
Other Grammar	0.829	0.784	0.783	0.856	0.724	0.795
Psych. Proc.	0.862	0.838	0.832	0.887	0.836	0.851
Punctuation	0.787	0.813	0.837	0.614	0.728	0.771

<https://doi.org/10.1371/journal.pone.0224425.t001>

PA

No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications

Erik de Vries¹, Martijn Schoonvelde² and Gijs Schumacher³

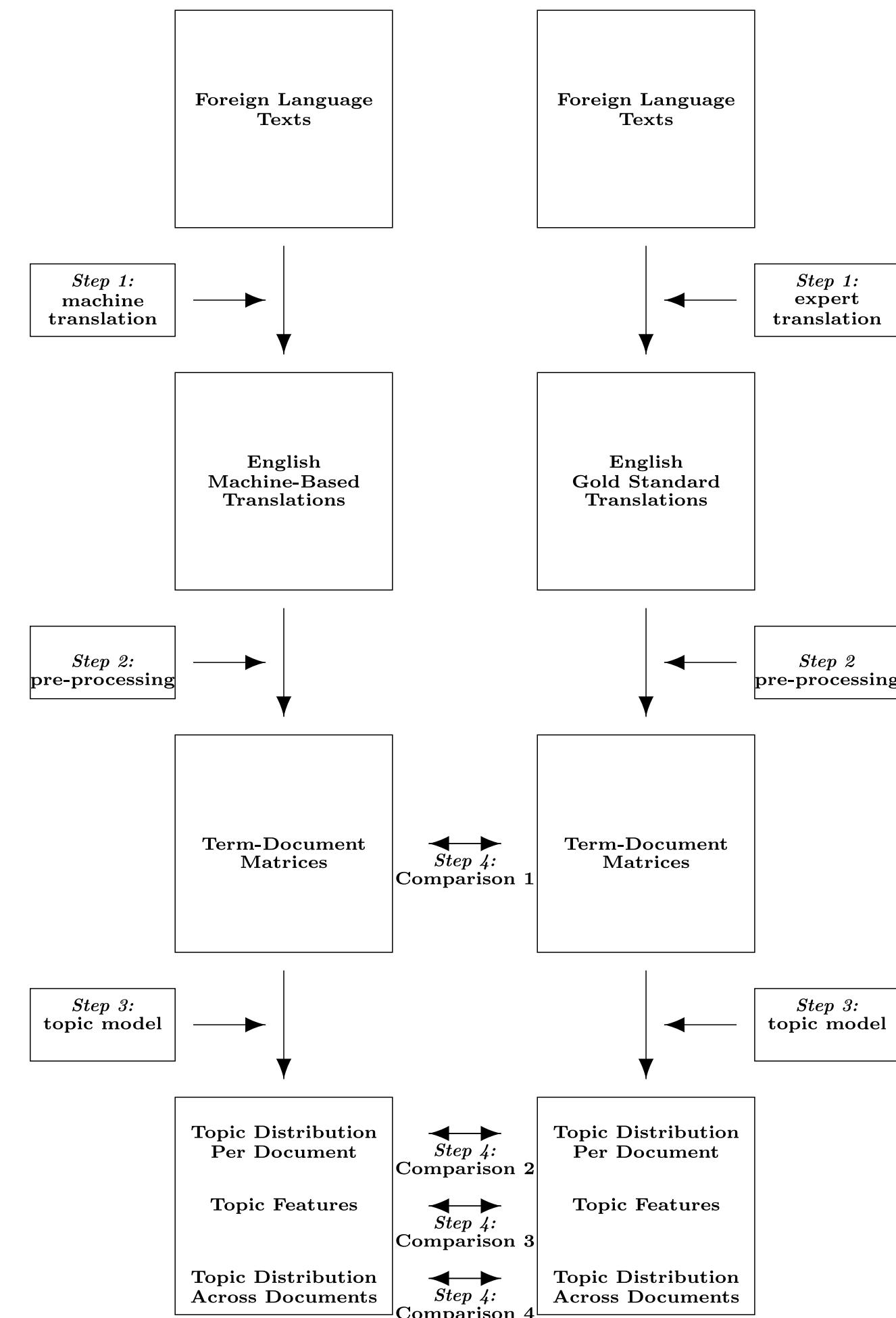


Figure 1. Research design.

Note: This figure shows the different steps of our research design. In both cases we start with non-English texts, which have been translated into English, either through Google Scholar or through EU-employed expert translators (Step 1). The English translations are then preprocessed and turned into TDMs (Step 2), on which we then estimate a topic model (Step 3). We then compare our four different outcome variables (Step 4). The comparisons are the following: *Comparison 1*: document-to-document comparison TDM similarity; *Comparison 2*: document-to-document comparison of topic distributions (topical prevalence); *Comparison 3*: topic-to-topic comparison of stem weights (topical content); *Comparison 4*: topic-to-topic comparison of topic distribution (topical prevalence).

Table 1. Comparisons between gold standard and machine-translated data.

	<i>Stems</i>	<i>Topics</i>
<i>Document level</i>	Stem counts per document pair	Topic distribution per document pair
<i>Corpus level</i>	Stem loadings per topic pair	Topic distribution per topic pair

Table 2. Cosine similarity distribution per language.

Language	N	Mean	St. Dev.	Min	Max
Danish	2,301	0.915	0.063	0.549	0.992
German	2,148	0.915	0.074	0.488	0.991
Spanish	2,335	0.929	0.059	0.483	0.991
French	2,347	0.925	0.064	0.564	0.989
Polish	2,338	0.913	0.073	0.475	0.989
Total:	11,469	0.919	0.066	0.475	0.992

Note: Statistically significant but substantively small difference between languages (ANOVA results:
 $F(4, 11464) = 27.855, \rho < 0.001, \eta^2 = 0.010$).

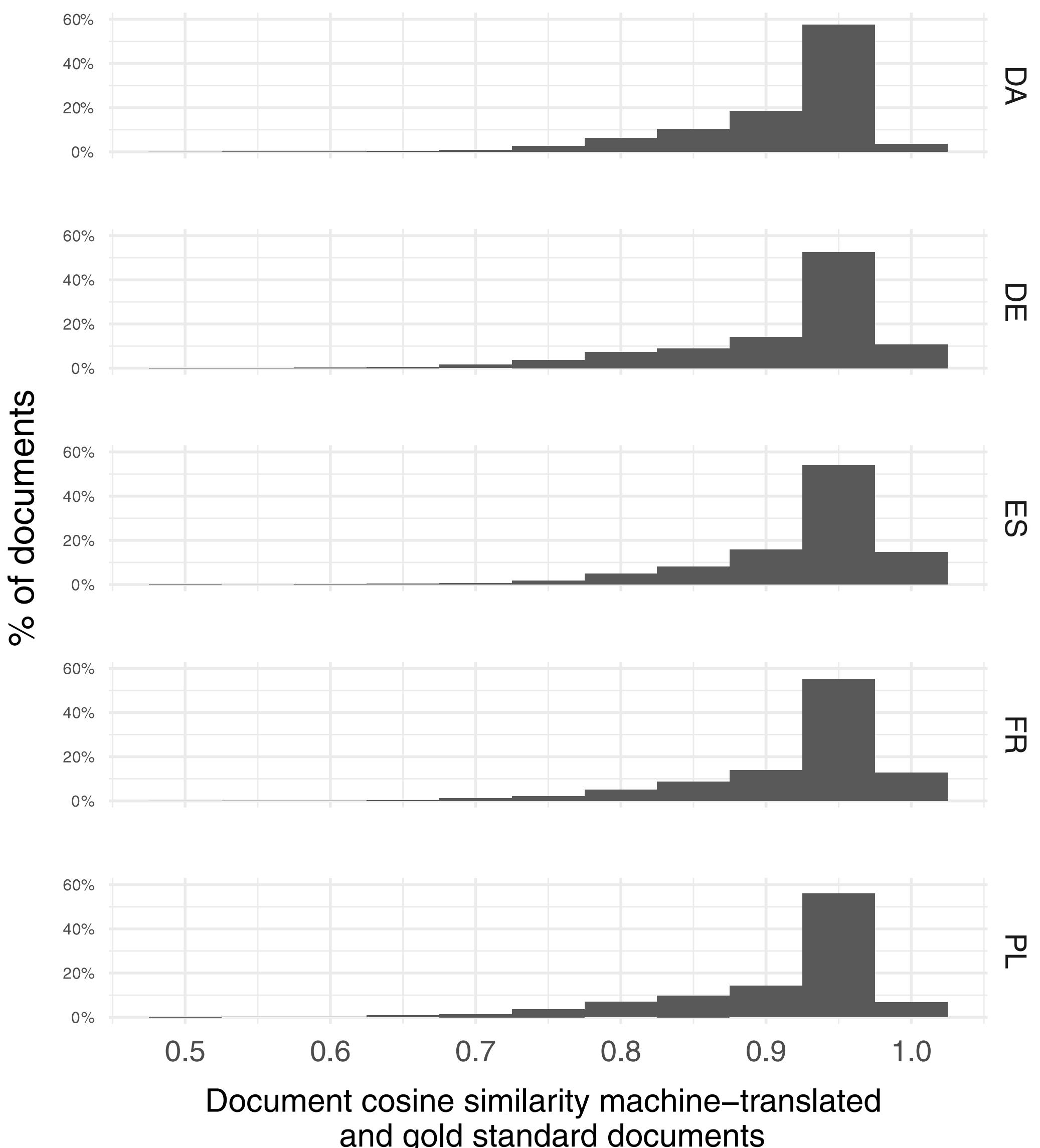


Figure 3. Distribution of cosine similarity per language pair.

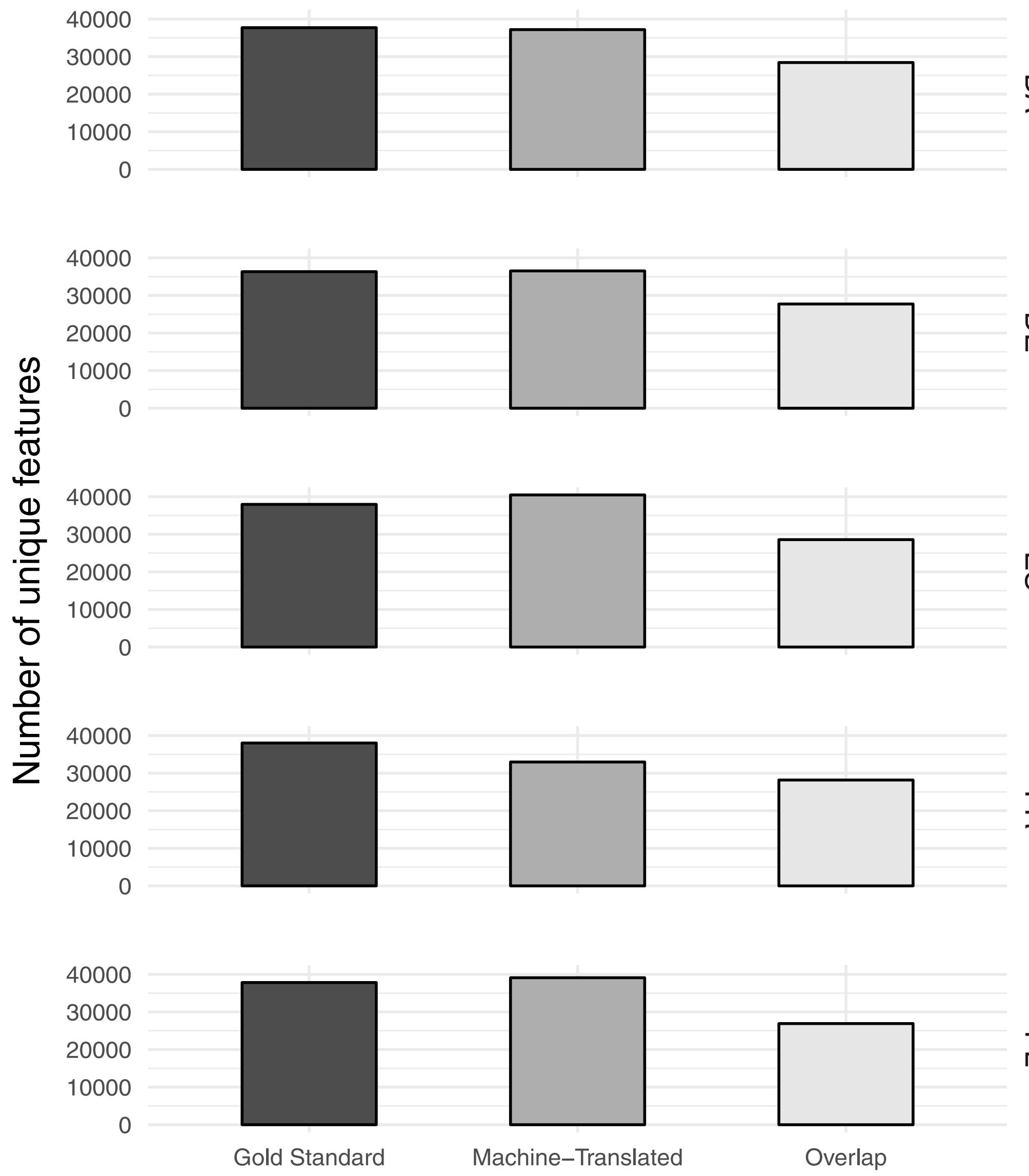


Figure 4. Unique TDM features for gold standard and machine-translated corpora. *Reading example:* for the French language, the amount of overlapping features is around 28,000, while the total number of features is around 33,000 for the machine-translated documents and around 38,000 for the gold standard documents.

Table 3. Similarity of document-level topical prevalence with equal number of topics.

Language	N	Mean	St. Dev.	Min	Max
Danish	2,301	0.783	0.202	-0.031	0.998
German	2,148	0.824	0.216	-0.031	0.999
Spanish	2,335	0.826	0.165	0.028	0.997
French	2,347	0.753	0.194	-0.043	0.996
Polish	2,338	0.809	0.206	-0.031	0.998
Total	11469	0.799	0.197	-0.043	0.999

Note: ANOVA results: $F(4, 11464) = 56.414, \rho < 0.001, \eta^2 = 0.019$.

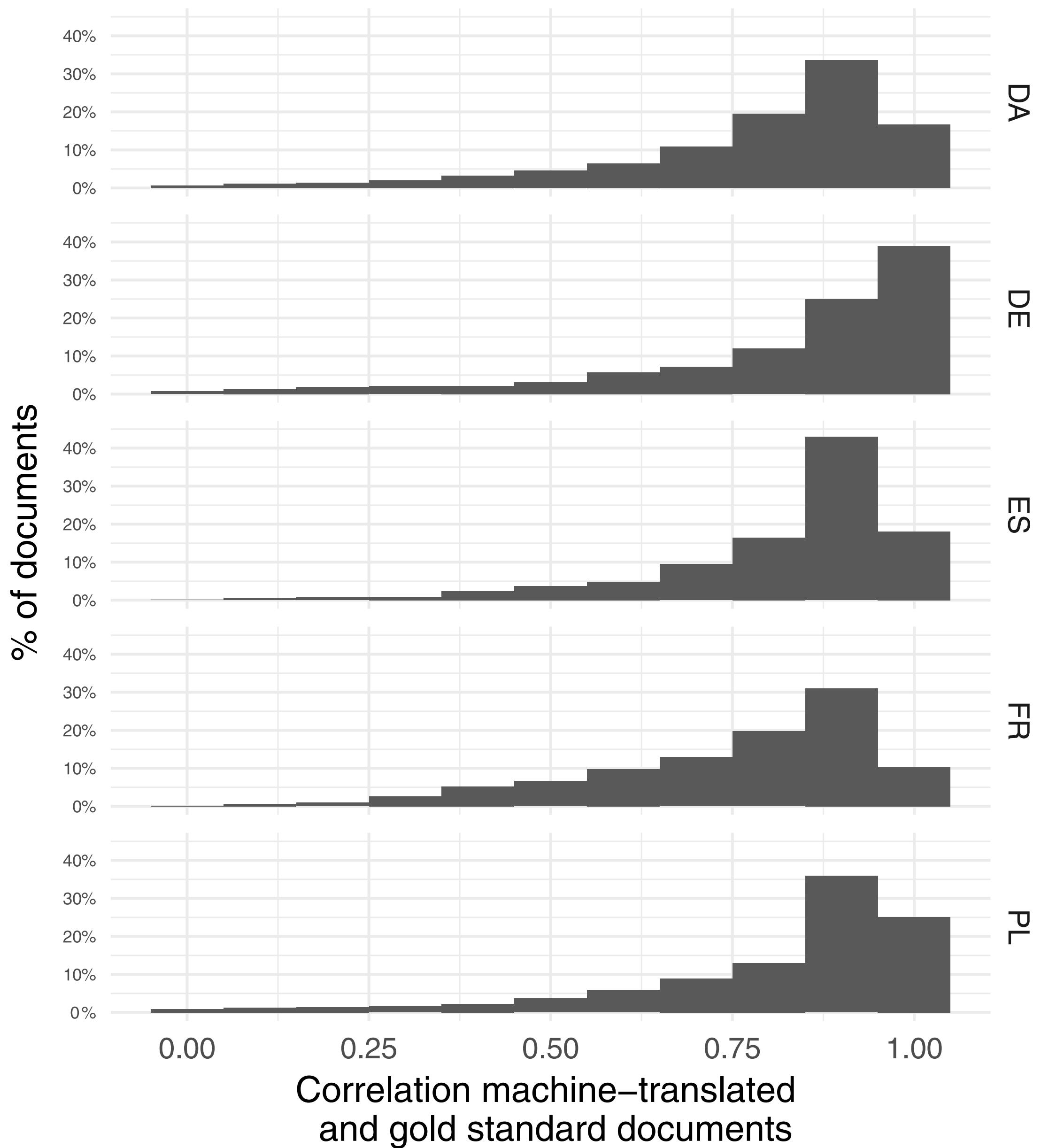


Figure 5. Similarity of document-level topical prevalence with equal number of topics.

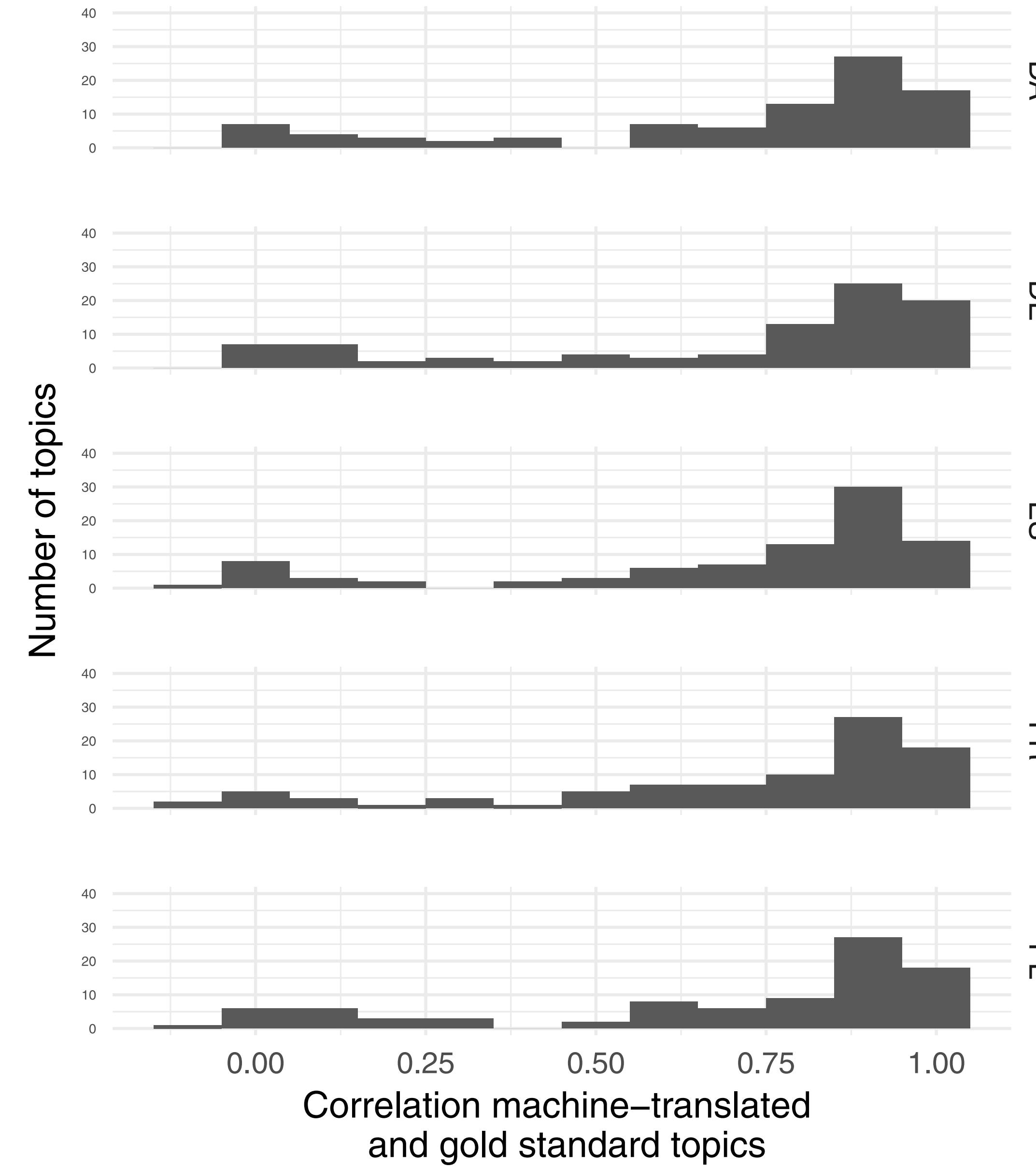


Figure 6. Similarity of corpus-level topical prevalence with equal number of topics. *Overall descriptives: N = 446, M = 0.699, SD = 0.321.*

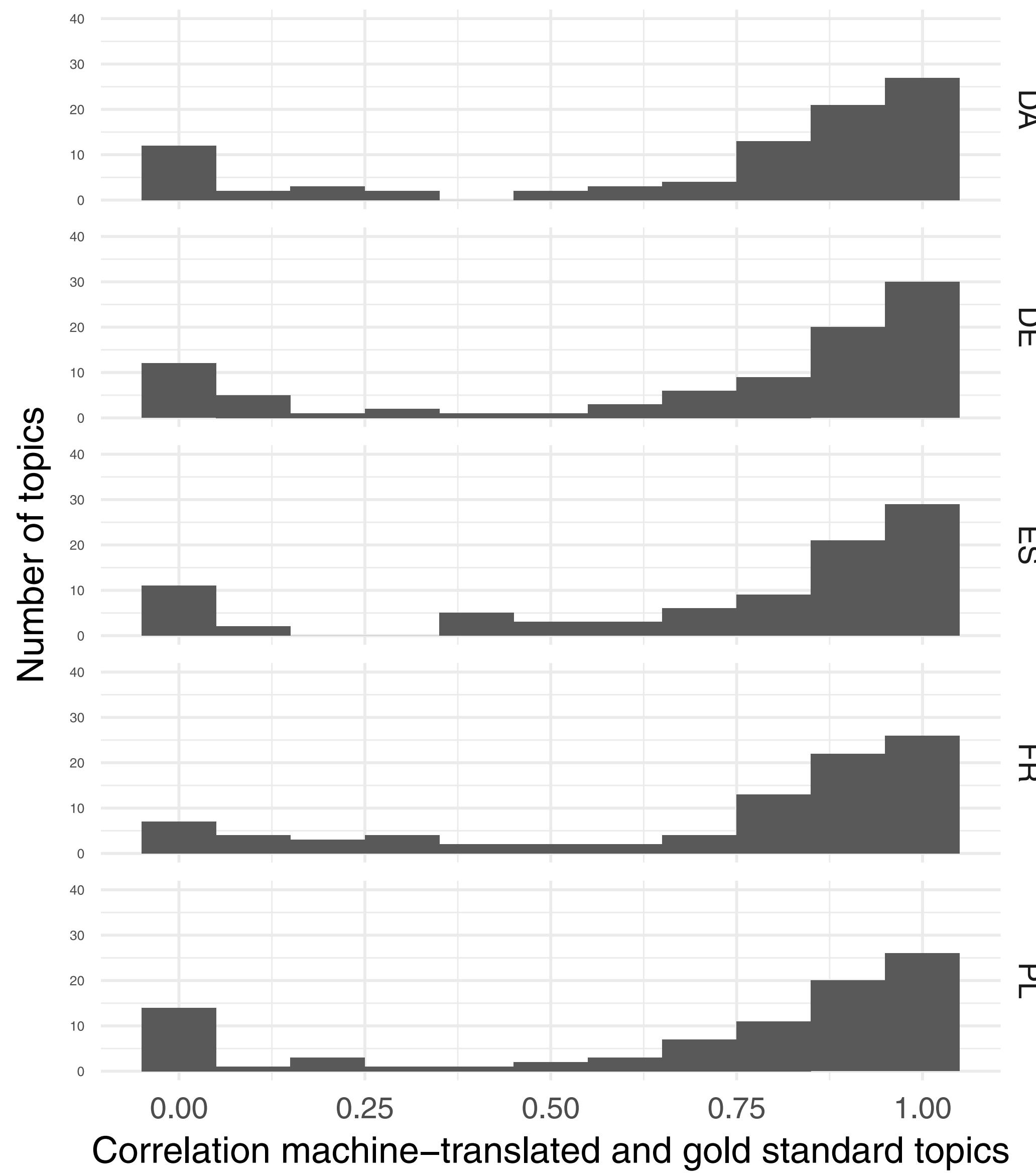


Figure 7. Similarity of topical content with equal number of topics. Overall descriptives: $N = 446$, $M = 0.708$, $SD = 0.345$.

A personal Google Translate through transfer learning - Sam