



**PennState**  
College of the  
Liberal Arts



## **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.





# Politicians talk a lot

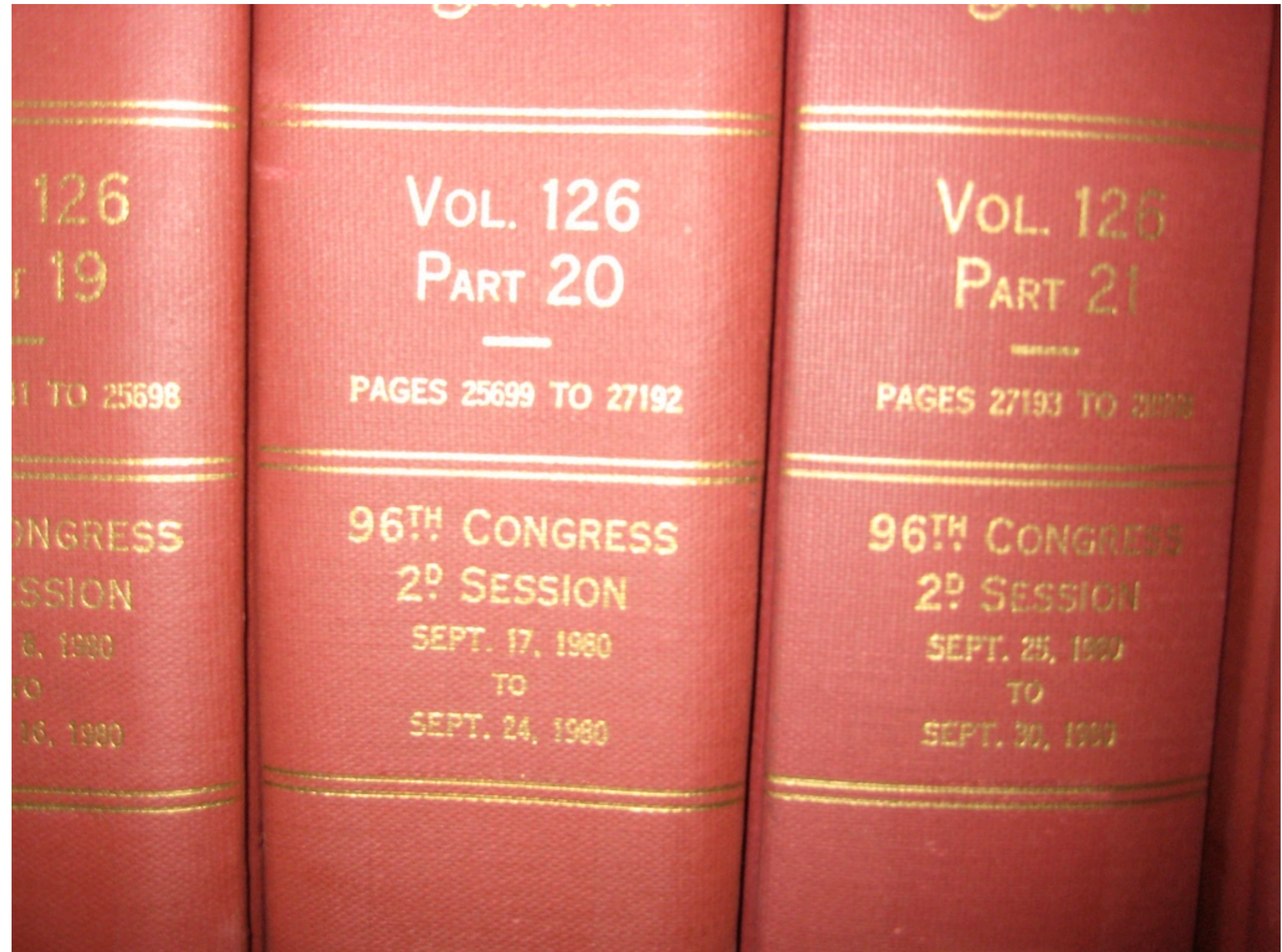
---

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

Political Map of the World, February 2021

- AUSTRALIA Independent state
- Bermuda Dependency or area of special sovereignty
- Sicily / AZORES Island / Island group
- ★ National capital
- ☆ Other capital

Scale 1:35,000,000  
Robinson Projection



# Politicians talk a lot - everywhere

## Uruguay

Political Map of the World, February 2021

AUSTRALIA Independent state  
Bermuda Dependency or area of special sovereignty  
Sicily / AZORES 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.



# Politicians talk a lot - everywhere

Political Map of the World, February 2021

- AUSTRALIA Independent state
- Bermuda Dependency or area of special sovereignty
- Sicily / AZORES Island / Island group
- ★ National capital
- ☆ Other capital

Scale 1:35,000,000  
Robinson Projection



# 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

Belgium  
(in two languages)

Political Map of the World, February 2021

AUSTRALIA Independent state  
Bermuda Dependency or area of special status  
Sicily / AZORES Island / Island group  
★ National capital  
☆ Other capital  
Scale 1:35,000,000  
Robinson Projection

**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)*

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

**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 voorzat, 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.



# Politicians talk a lot - everywhere

Political Map of the World, February 2021

- AUSTRALIA Independent state
- Bermuda Dependency or area of special sovereignty
- Sicily / AZORES Island / Island group
- ★ National capital
- ☆ Other capital

Scale 1:35,000,000  
Robinson Projection



# Politicians talk a lot - everywhere

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

ගරු සජිත් ප්‍රේමදාස මහතා  
(மாண்புமிகு சஜித் பிரேமதாஸ்)  
(The Hon. Sajith Premadasa)

ගරු කථානායකතුමනි, මට මේ අවස්ථාව ලබා දීම ගැන ඔබතුමාට ස්තූතියි. මා මේ ගැන ඔබතුමාට දැනුම් දුන්නා. දියවන්නා 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 දවසේත් ඒක අසත්‍යයක් බවට පත් වෙලා තිබෙනවා. ඔබතුමා මේ ගැන හොඳට සොයා බලන්න. මා හිතන විධියට, drainage systems, sewerage systems ගණනාවක් දියවන්නා ඔයට යොමු කර තිබෙනවා. එම නිසා අද දියවන්නා ඔයේ බරපතල පරිසර හානියක් සිදු වෙලා තිබෙනවා. මා මේ ගැන එදා සඳහන් කරන කොට අද ඔය කෑ ගහන ඇත්තෝ මට දොස් කීවා, ලකුණු දමා ගන්න හදනවා කියලා. ඒ වාගේම, අසහ්‍ය වචනත් පාවිච්චි කළා. නමුත් අද මම ඔබතුමාගෙන් බොහොම කාරුණික ඉල්ලීමක් කරනවා, ගරු කථානායකතුමනි. එදා සිට අද දවස දක්වා ඒ වෙනුවෙන් කිසිම දෙයක් කරලා නැහැ.

Sri Lanka



# It is easy to look like a fool using Google Translate

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

ගරු සජීත් ප්‍රේමදාස මහතා  
(மாண்புமிகு சஜித் பிரேமதாஸ)  
(The Hon. Sajith Premadasa)

ගරු කථානායකතුමනි, මට මේ අවස්ථාව ලබා දීම ගැන ඔබතුමාට ස්තූතියි. මා මේ ගැන ඔබතුමාට දැනුම් දුන්නා. දියවන්නා 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 දවසේත් ඒක අසත්‍යයක් බවට පත් වෙලා තිබෙනවා. ඔබතුමා මේ ගැන හොඳට සොයා බලන්න. මා හිතන විටියට, drainage systems, sewerage systems ගණනාවක් දියවන්නා ඔයට යොමු කර තිබෙනවා. එම නිසා අද දියවන්නා ඔයේ බරපතල පරිසර හානියක් සිදු වෙලා තිබෙනවා. මා මේ ගැන එදා සඳහන් කරන කොට අද ඔය කෑ ගහන ඇත්තෝ මට දොස් කිව්වා, ලකුණු දමා ගන්න හදනවා කියලා. ඒ වාගේම, අසත්‍ය වචනන් පාවිච්චි කළා. නමුත් අද මම ඔබතුමාගෙන් බොහොම කාරුණික ඉල්ලීමක් කරනවා, ගරු කථානායකතුමනි. එදා සිට අද දවස දක්වා ඒ වෙනුවෙන් කිසිම දෙයක් කරලා නැහැ.

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.

Personal communication: Kaushalya Perera

# Politicians talk a lot - everywhere

Political Map of the World, February 2021

- AUSTRALIA Independent state
- Bermuda Dependency or area of special sovereignty
- Sicily / AZORES Island / Island group
- ★ National capital
- ☆ Other capital

Scale 1:35,000,000  
Robinson Projection



# Politicians talk a lot - everywhere

Political Map of the World, February 20



محضر الجلسة (40) من الفصل التشريعي الاول السنة التشريعية الثالثة  
التاريخ: الأربعاء 10 ايلول 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 إنتخابات المحافظات نبدأ بالفقرات التالية من جدول الأعمال.

- السيد مثال 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 ان 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 بأسم حزب الأمة العراقي بل بأسم المجلس الإسلامي الأعلى والتيار الصدري والفضيلة والدعوة والتوافق والوفاق والعراقية وكل القوائم والأعضاء ذكرتها بالتسلسل لأنني أشعر أنا واحد منكم وكلنا لبعضنا من أجل هذا الوطن, لكن التهديدات مستمرة لأننا طالبنا بإعادة المهجرين الى منازلهم. ويجب أن أخبر هيئة الرئاسة الموقرة إن كلا الطرفين المهجر والذي لازال يسكن في حي الجامعة وبمفردات أخرى لا أحب أن أذكرها لكن للتوضيح أهاليينا من السنة الساكنين في حي الجامعة وحي العدل يقولون لي وبرسائل موثقة يقولون عودوا إلينا بأهلنا الشيعة والشيعة المهجرين يقولون دعونا نذهب الى أهلنا السنة ونعود الى الحياة الطبيعية. وحصلت على دعم رئيس الوزراء الأخ نوري المالكي في هذا الشأن لكن الدنيا قامت ولم تقعد وأصابع الإتهام واضحة لمن لا يريد الخير للعراق ولمن يشعر ان هذا المشروع يهدد



# Politicians talk a lot - everywhere

Political Map of the World, February 2021

- AUSTRALIA Independent state
- Bermuda Dependency or area of special sovereignty
- Sicily / AZORES Island / Island group
- ★ National capital
- ☆ Other capital

Scale 1:35,000,000  
Robinson Projection



# 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

Freitag 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.

(Geschicht.)

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. Oberfohren u. Gen. eingebrachten Entwurfs eines Gesetzes zur **Behebung 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 Deutschnationalen Volkspartei Ihnen durch einen Initiativantrag ein Gesetz zur Behebung 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-



## Weimar Republic

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

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

בָּלַל (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>...</i>	<i>turmoil</i>	<i>markets</i>
	0	0	<i>...</i>	1	1
	1	1	<i>...</i>	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)



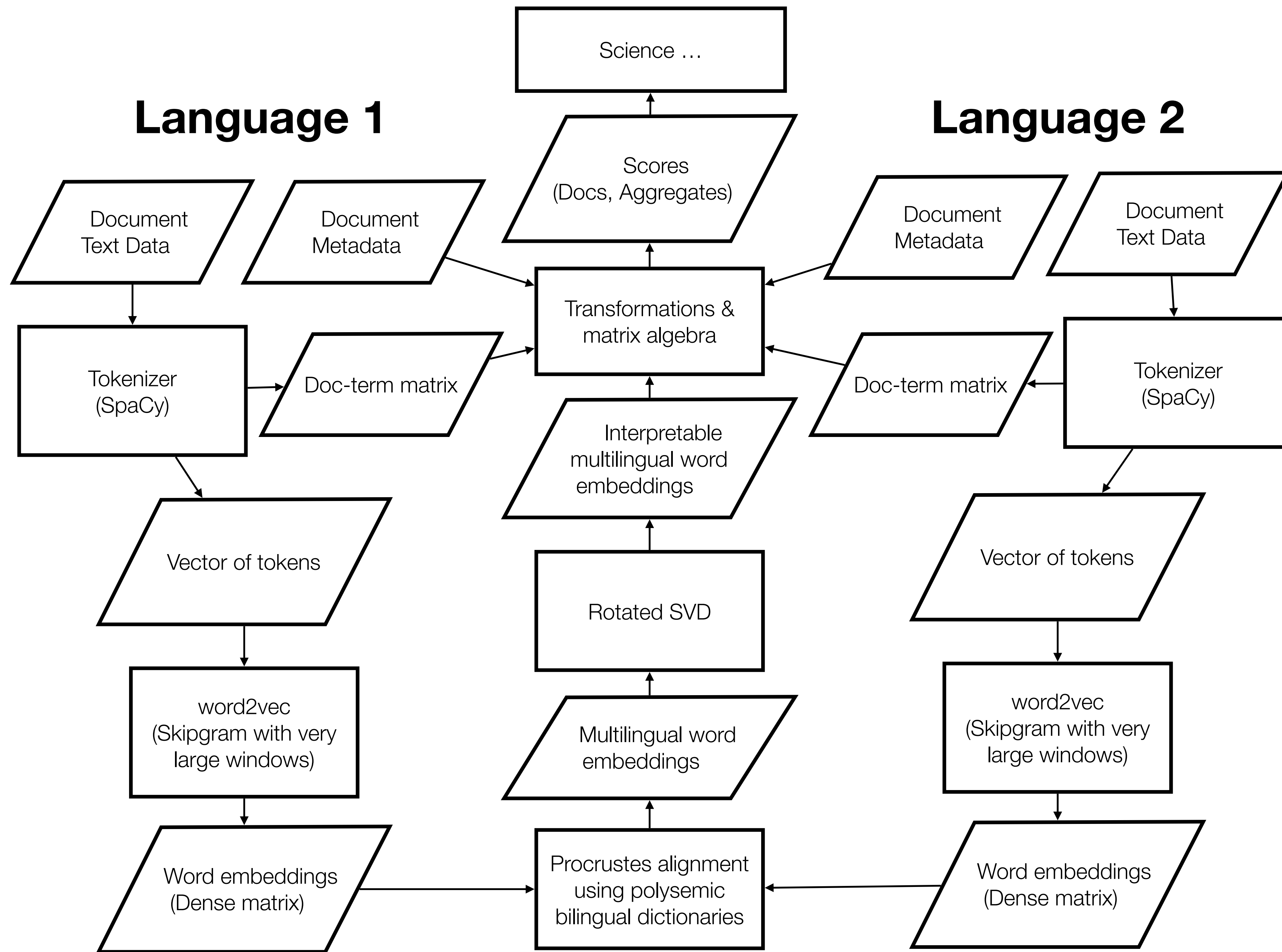
<b>Chamber</b>	<i>Riksdag</i>	<i>Eduskunta</i>	<i>Parlament</i>	<i>Tweede Kamer</i>	<i>Bundestag</i>	<i>Congreso</i>
<b>Years</b>	1990-2015	1991-2015	1993-2016	1994-2015	1991-2013	1989-2015
<b>Speeches</b>	317,132	245,852	329,135	900,796	299,844	290,680
<b>Vocabulary</b>	638,825	1,422,492	337,127	409,323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

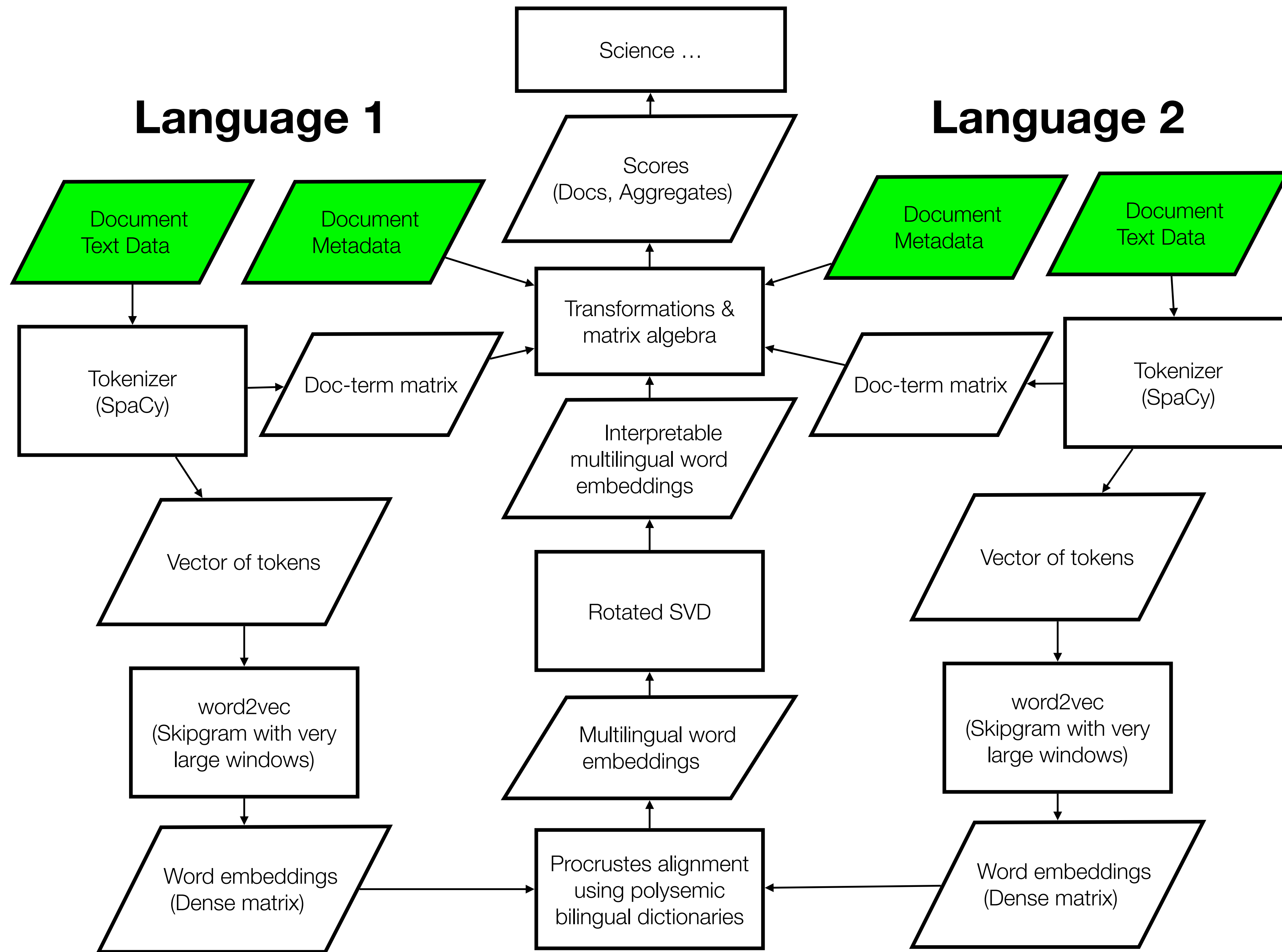
## Custom

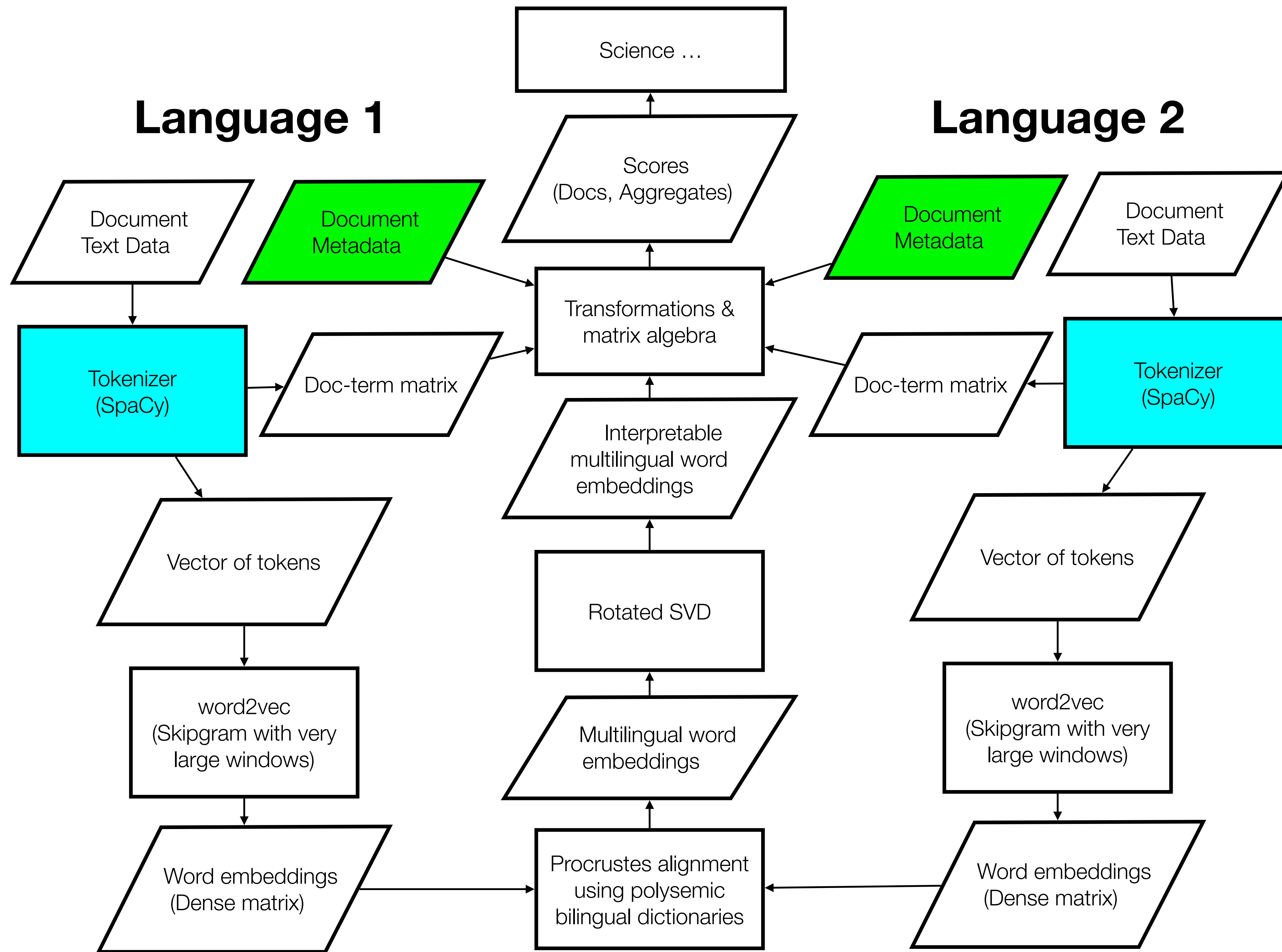


<b>Chamber</b>	<i>House of Commons</i>	<i>Camera dei Deputati</i>
<b>Years</b>	2006-2017	2006-2017
<b>Speeches</b>	602,763	215,022
<b>Vocabulary</b>	140,717	284,959

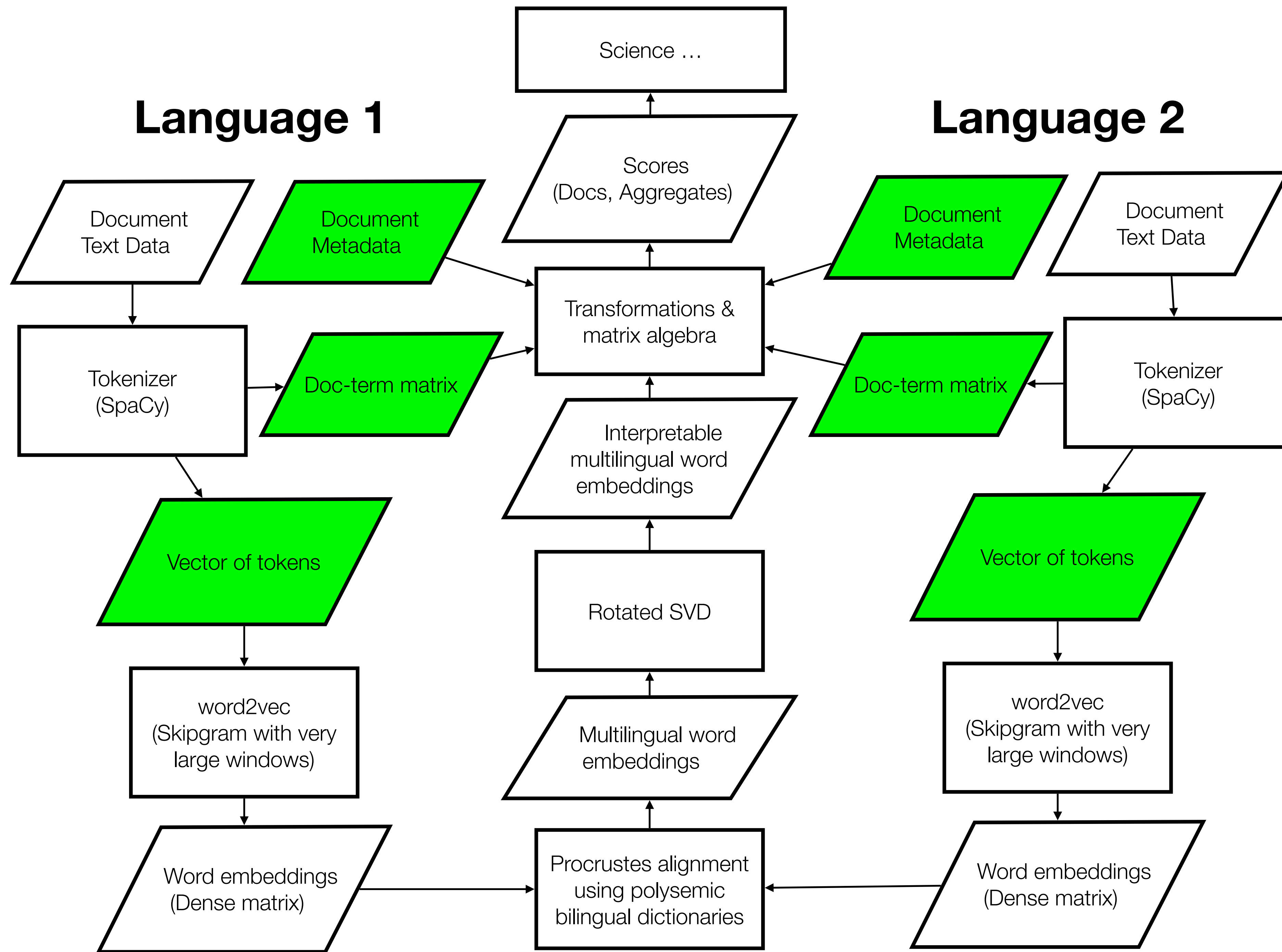
The pipeline ...

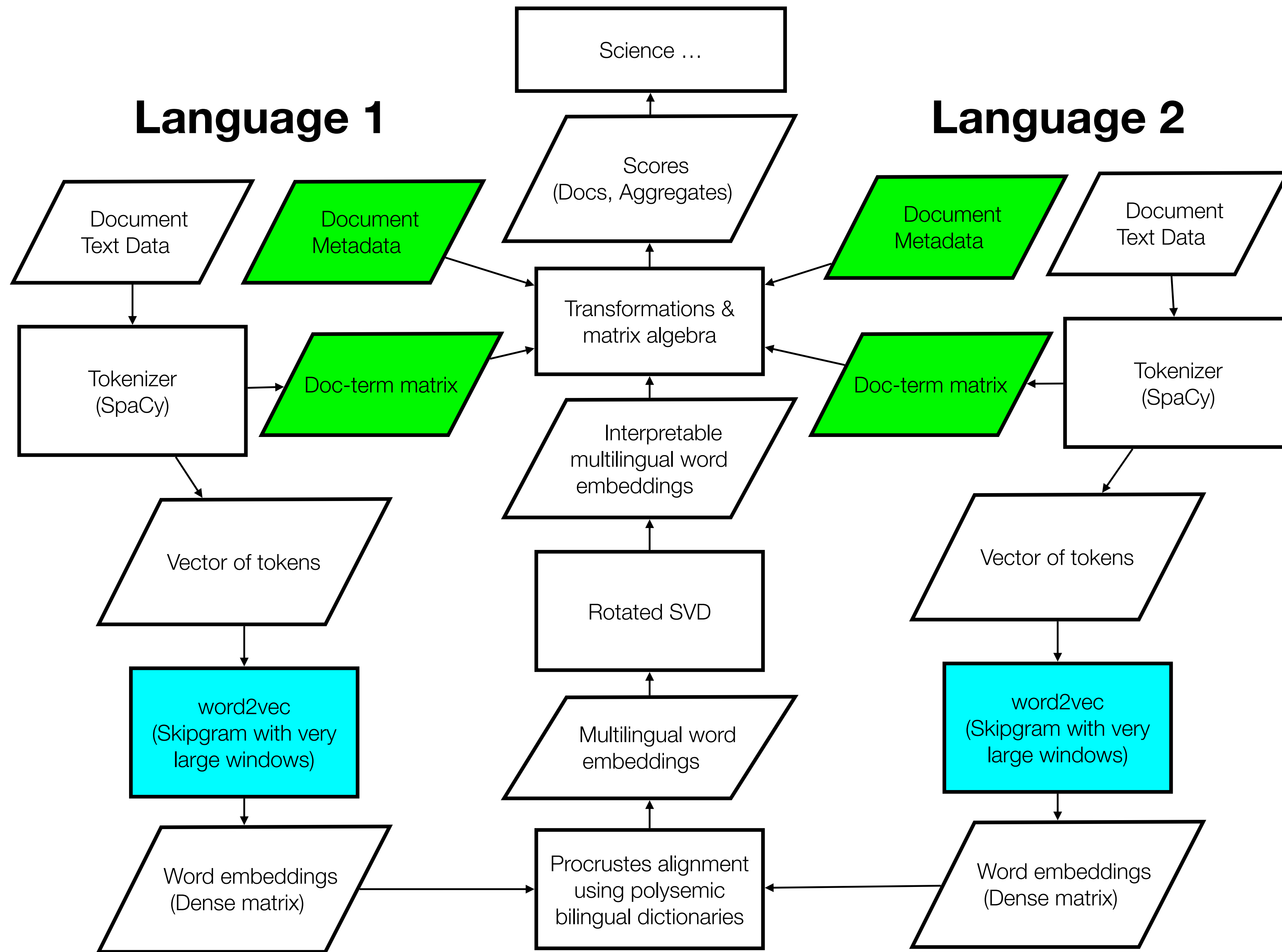




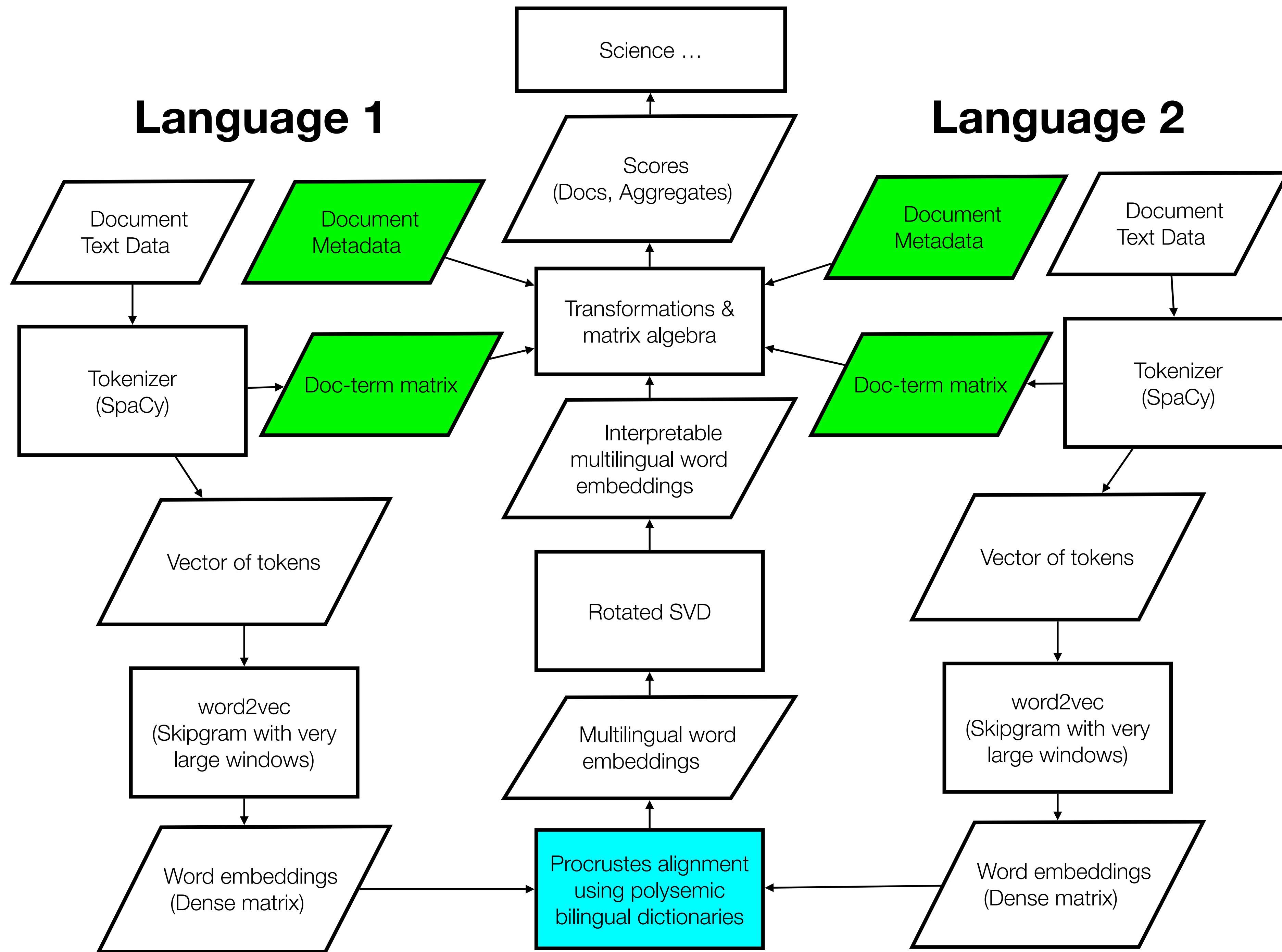


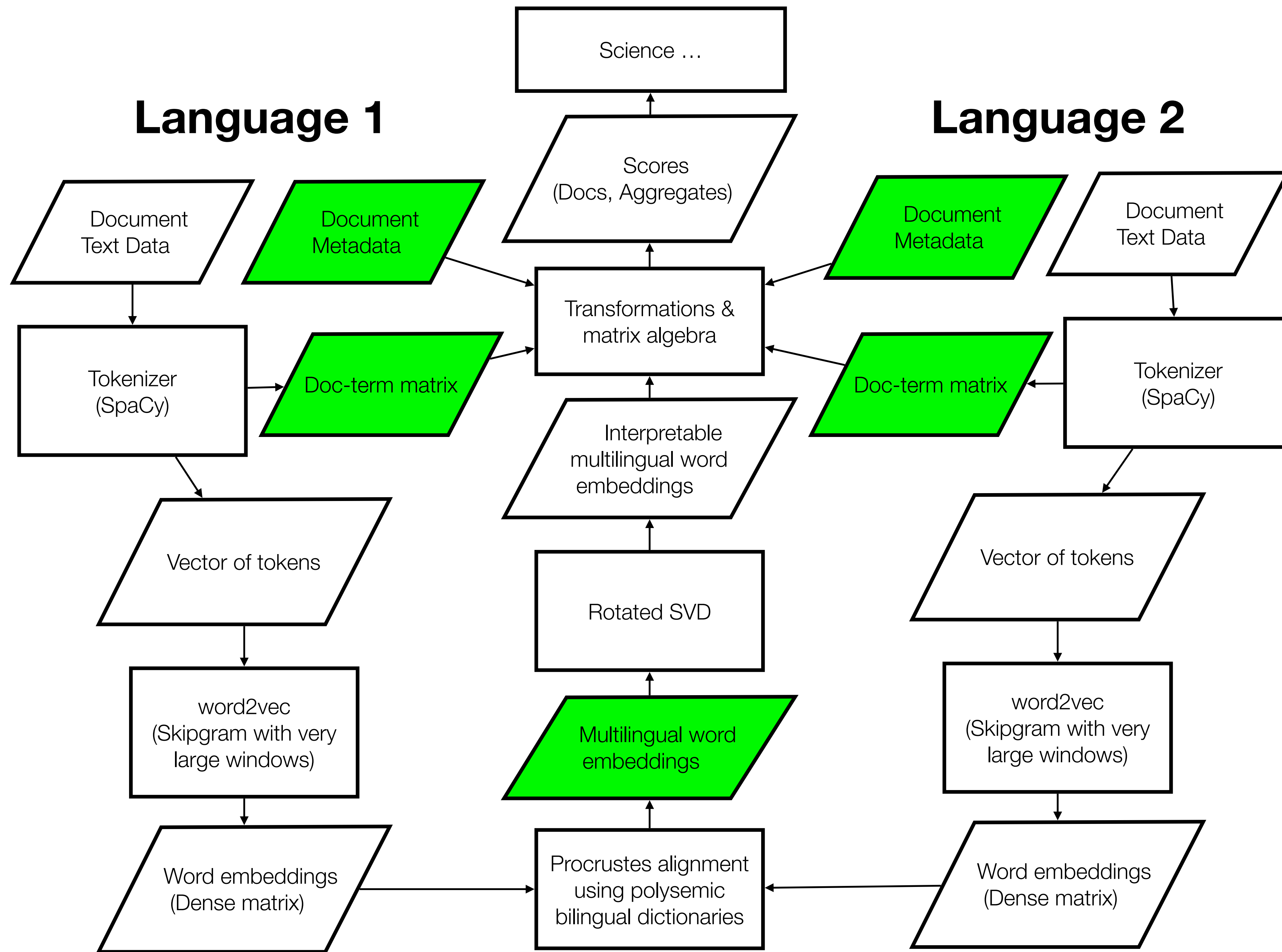


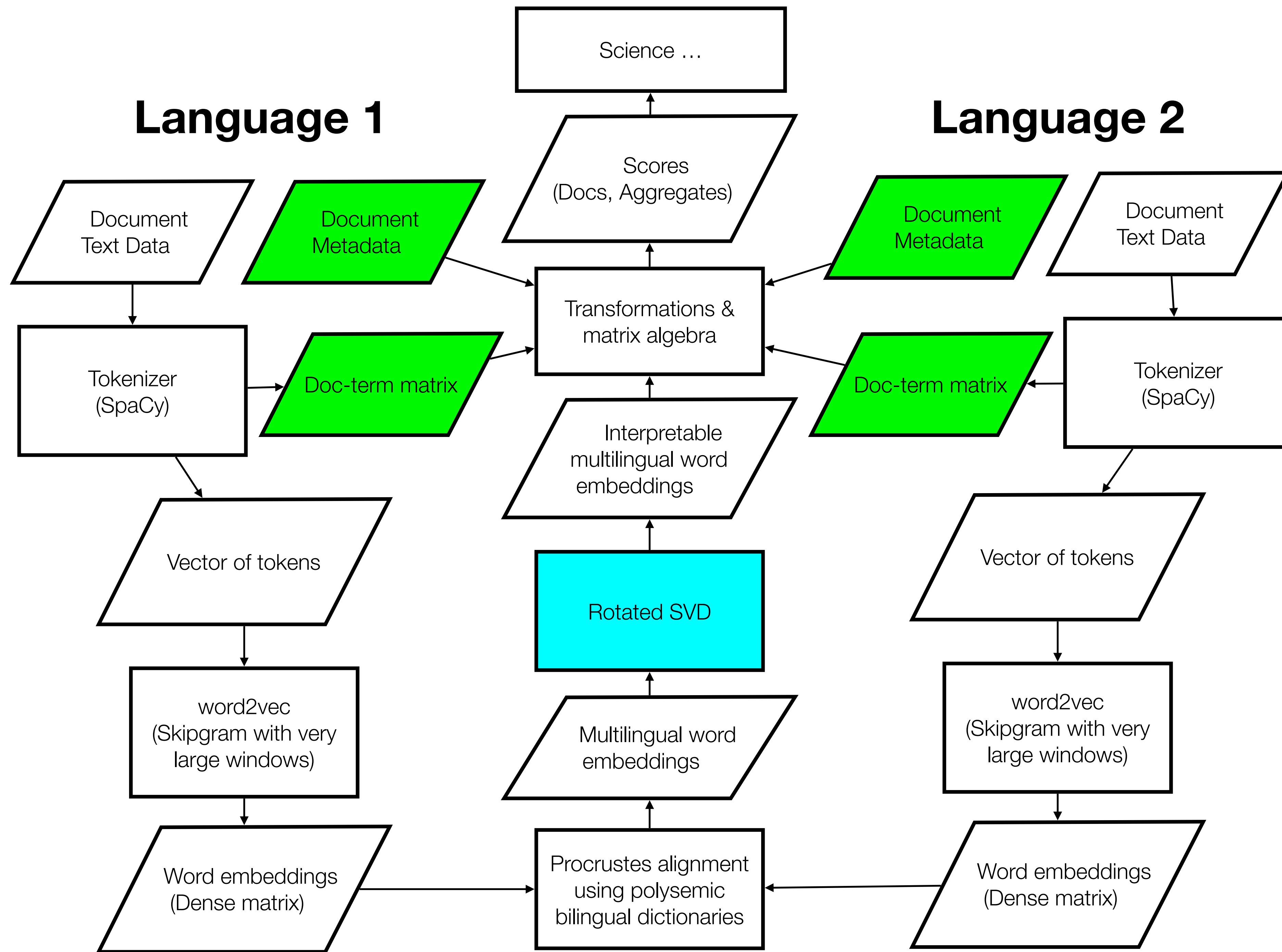


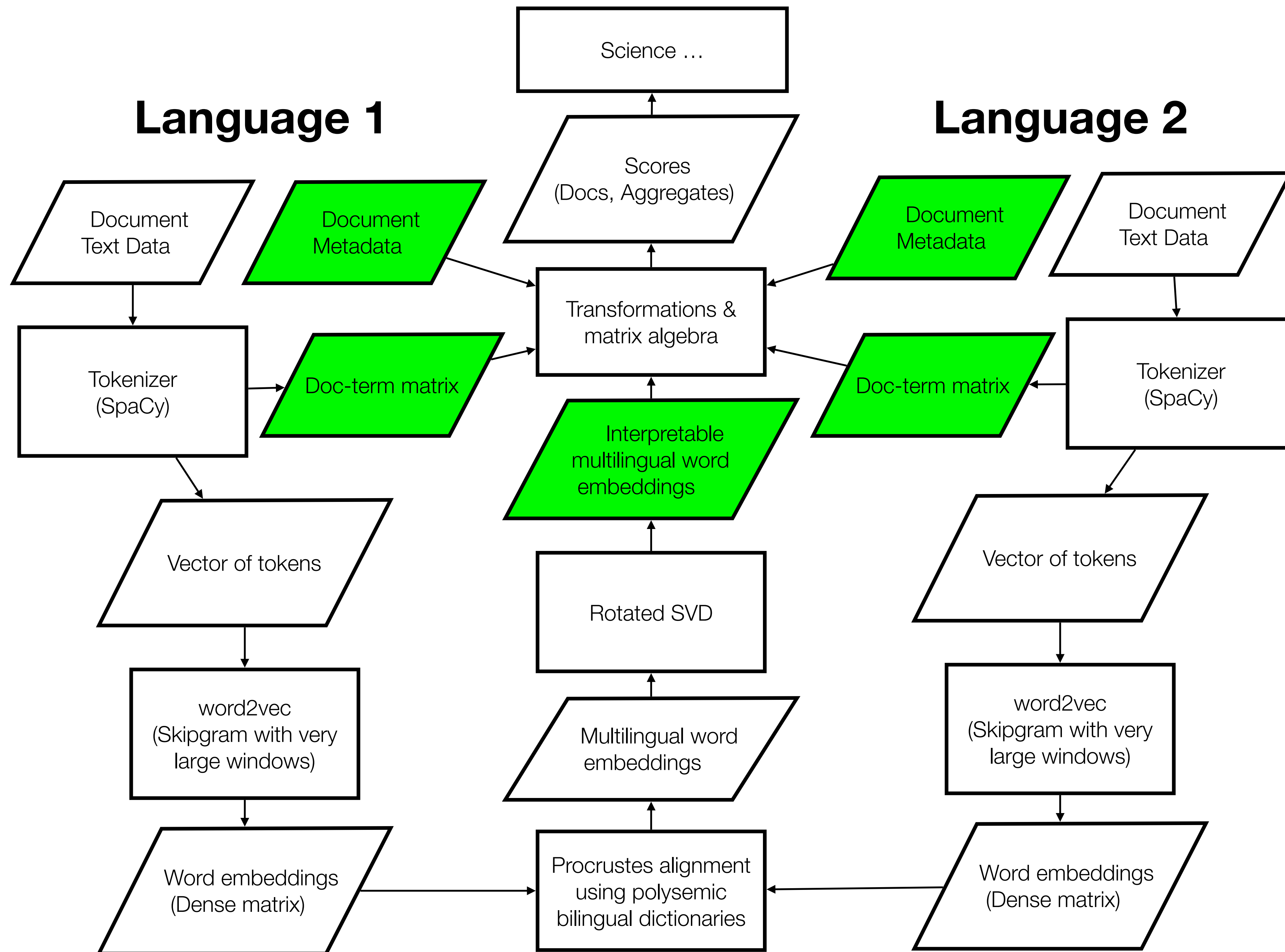


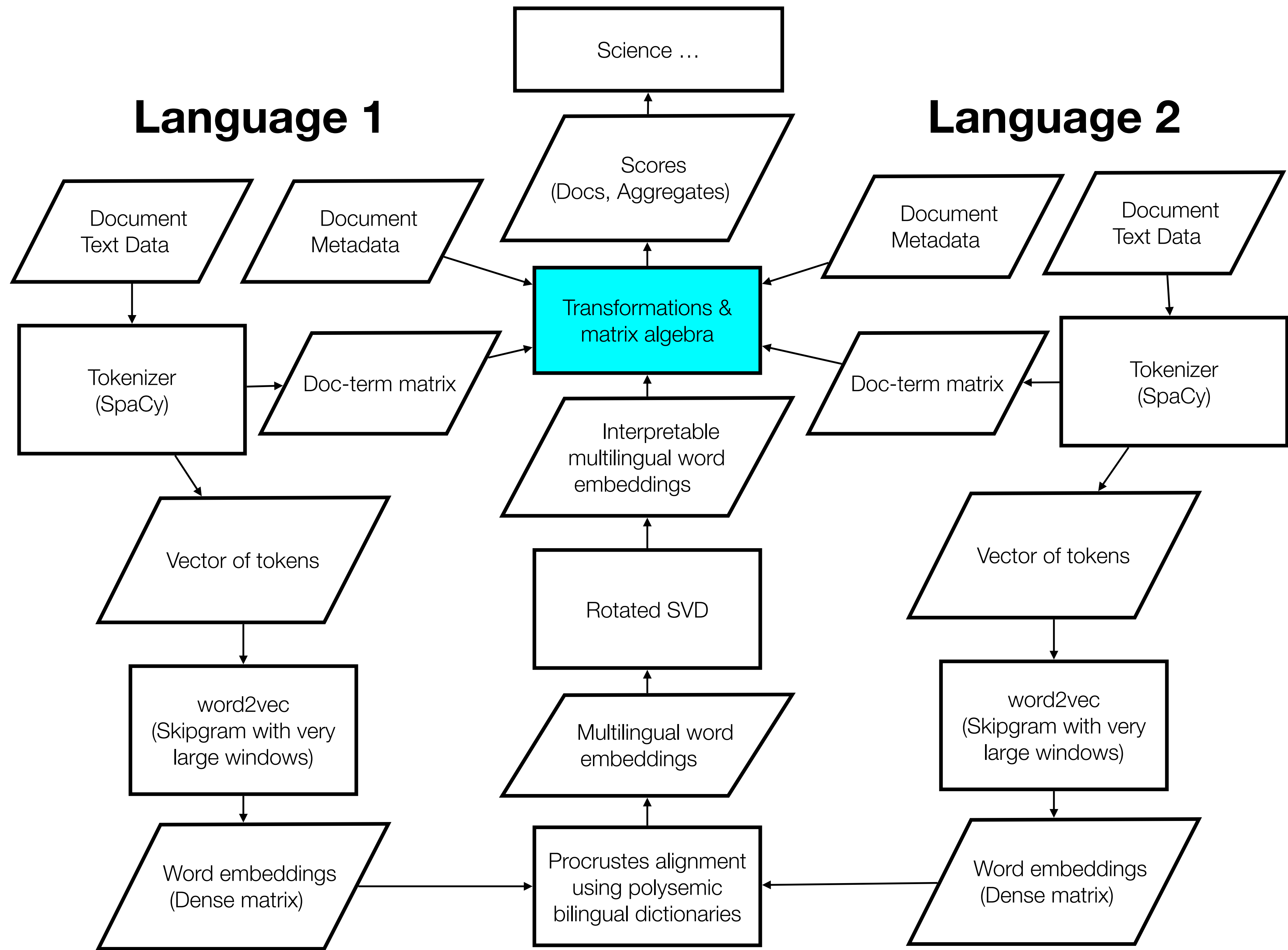




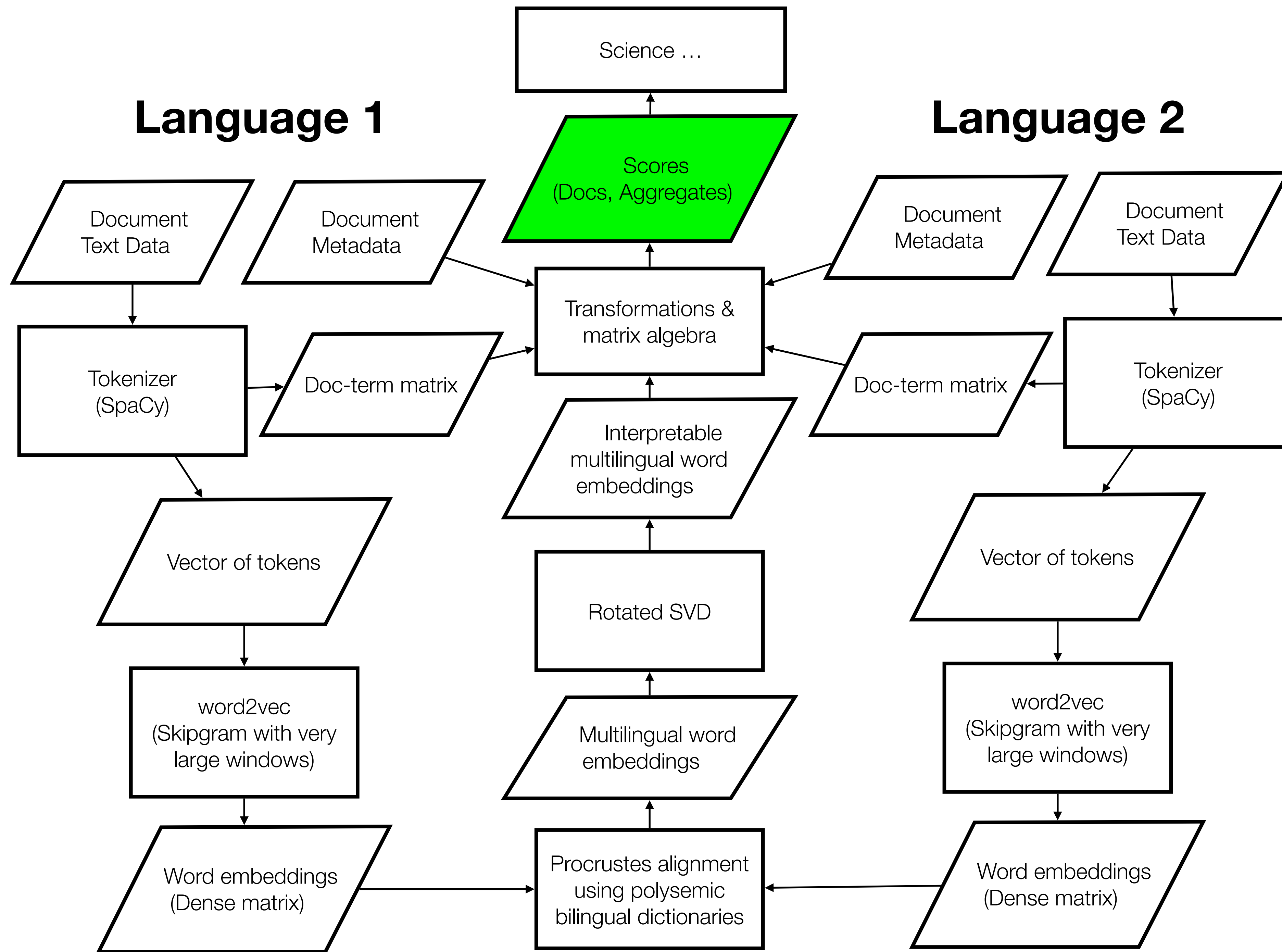


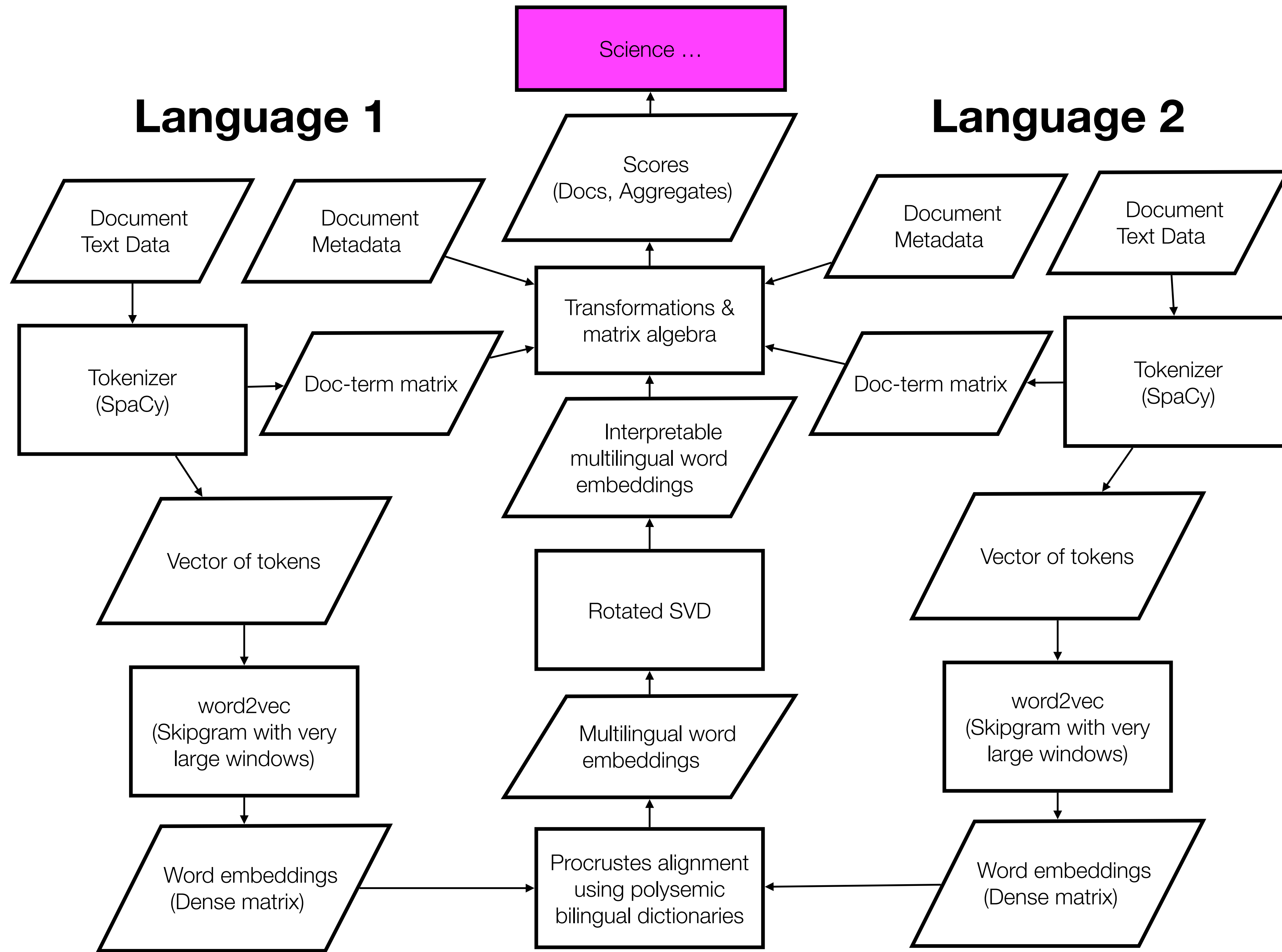


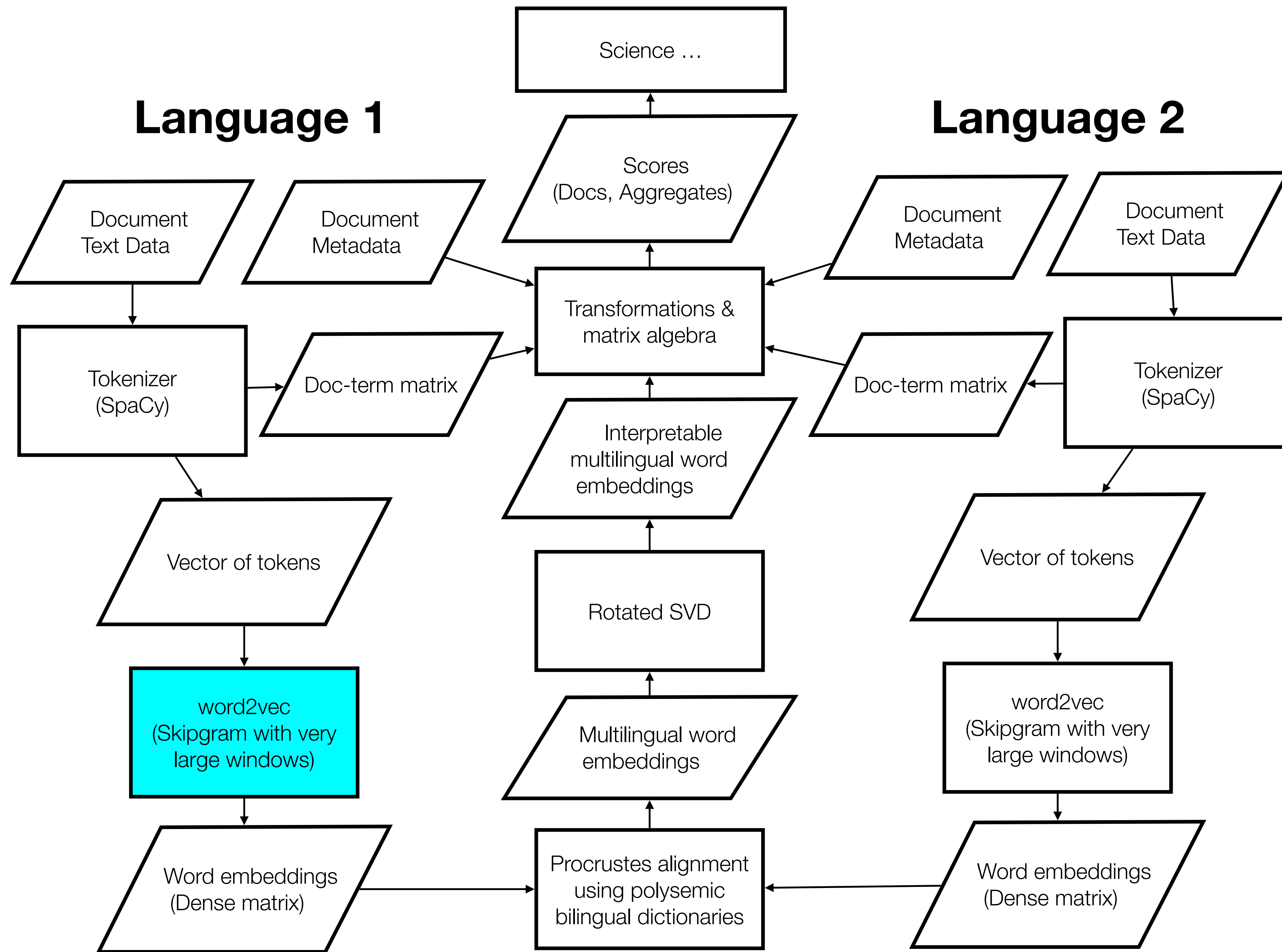












# Word embeddings

---

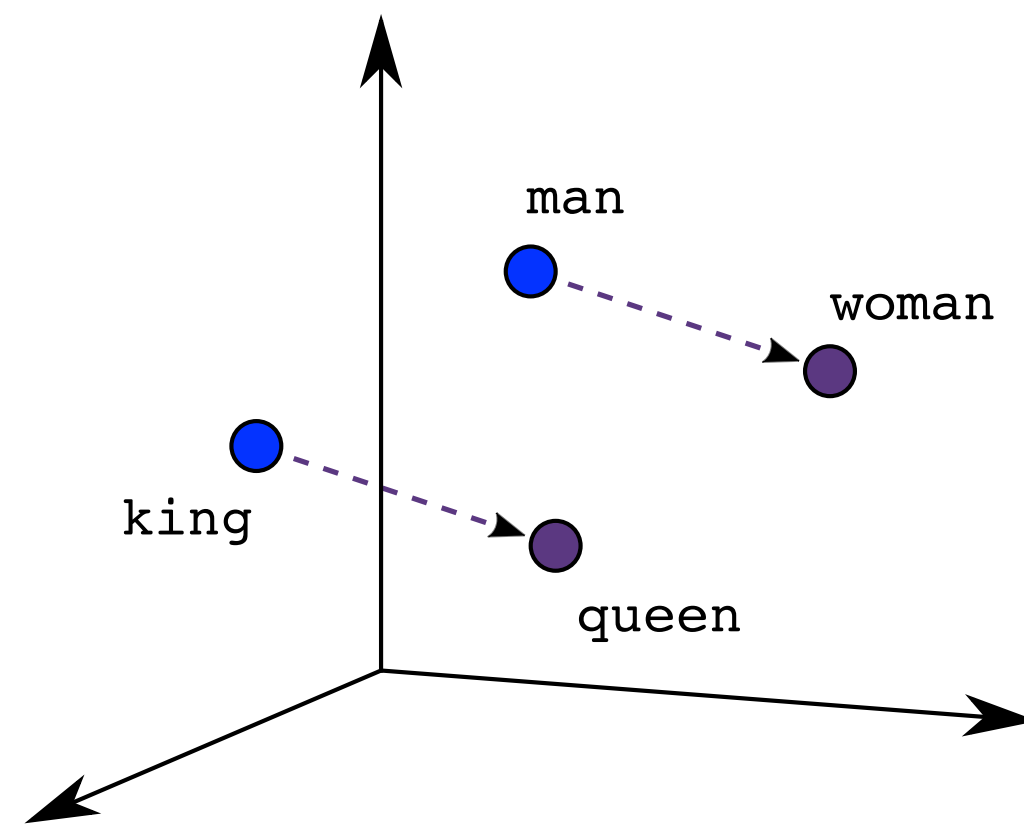
- The old (Mikolov, et al. 2013 — word2vec) new hotness in NLP / text-as-data.
- Represent each word as a vector of real numbers, a location in  $n$ -dimensional space.
- Premise: “You shall know a word by the company it keeps.” (Firth 1957)
- Goal: words that appear in the same “contexts” should be 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

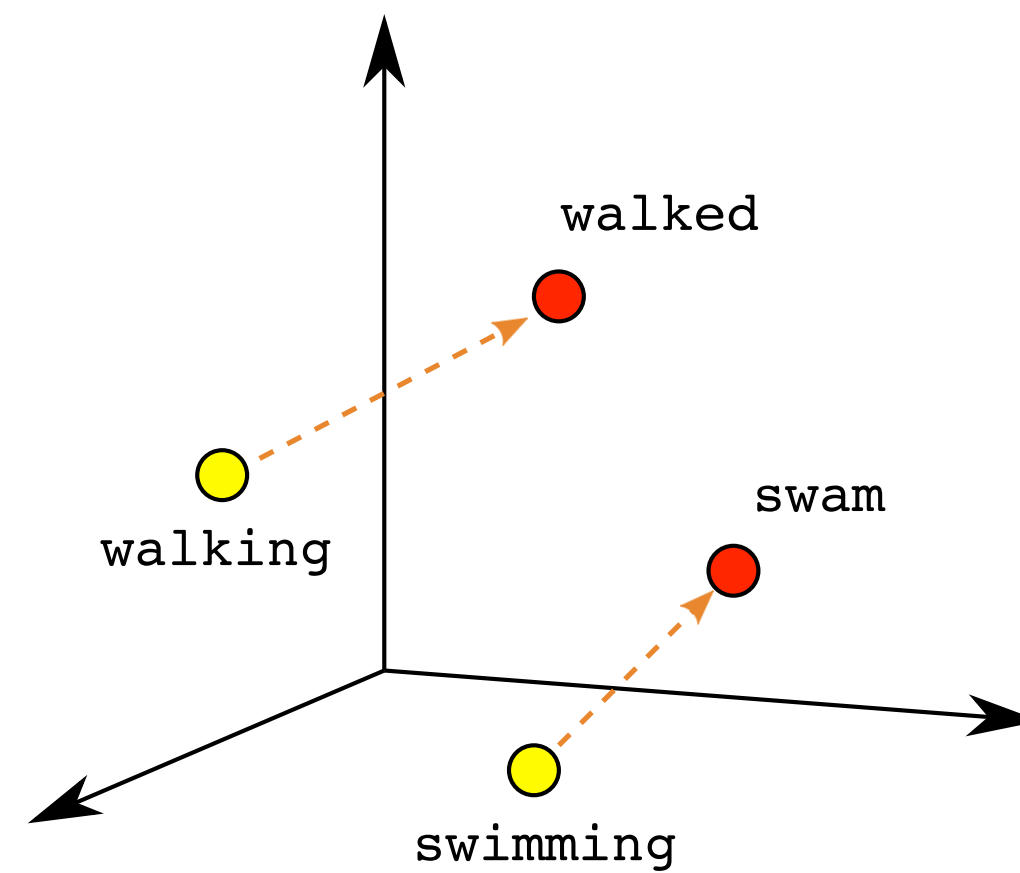
**([Context words], focus word) with window size of 1:**

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

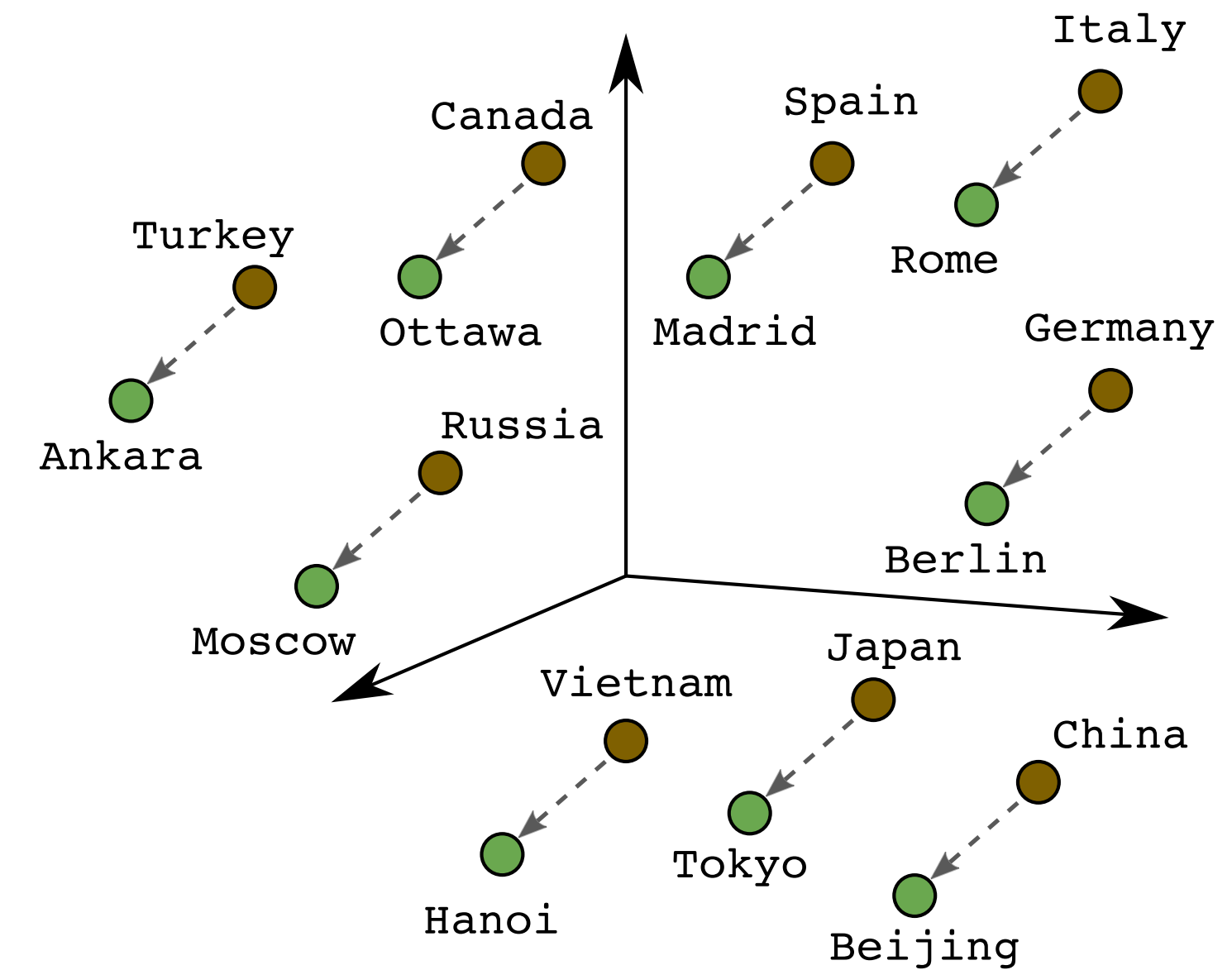
# If you know one thing about word embeddings — vector semantics



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

---

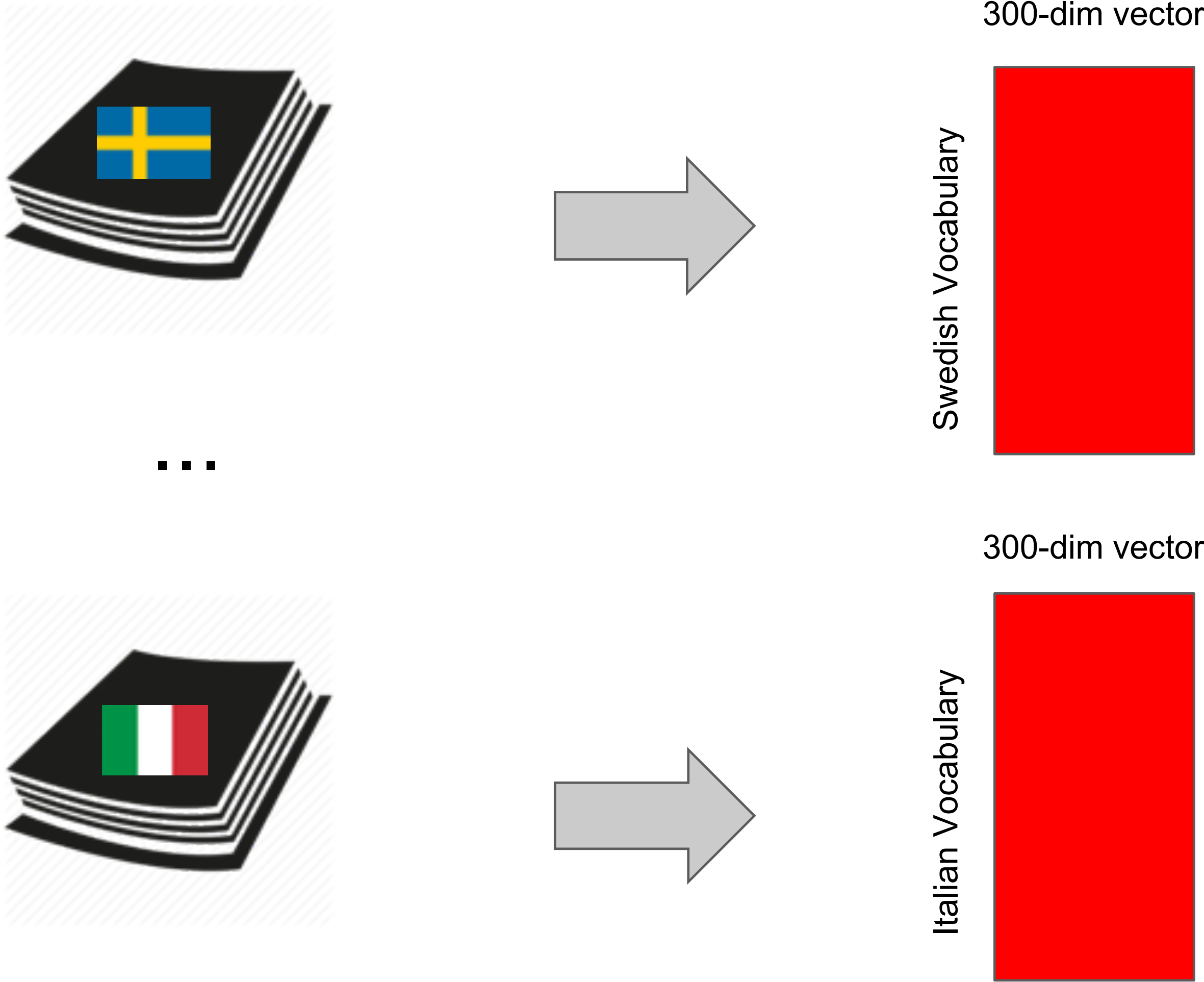
window size: 3		window size: 30		window size: 300	
putting	able	effective	crimes	diseases	tourism
bringing	can	deliver	offences	disease	visitors
taking	unable	improve	offence	vaccine	museums
giving	trying	strengthen	prosecutions	obesity	tourist
introducing	prepared	improving	murder	screening	citizenship
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.

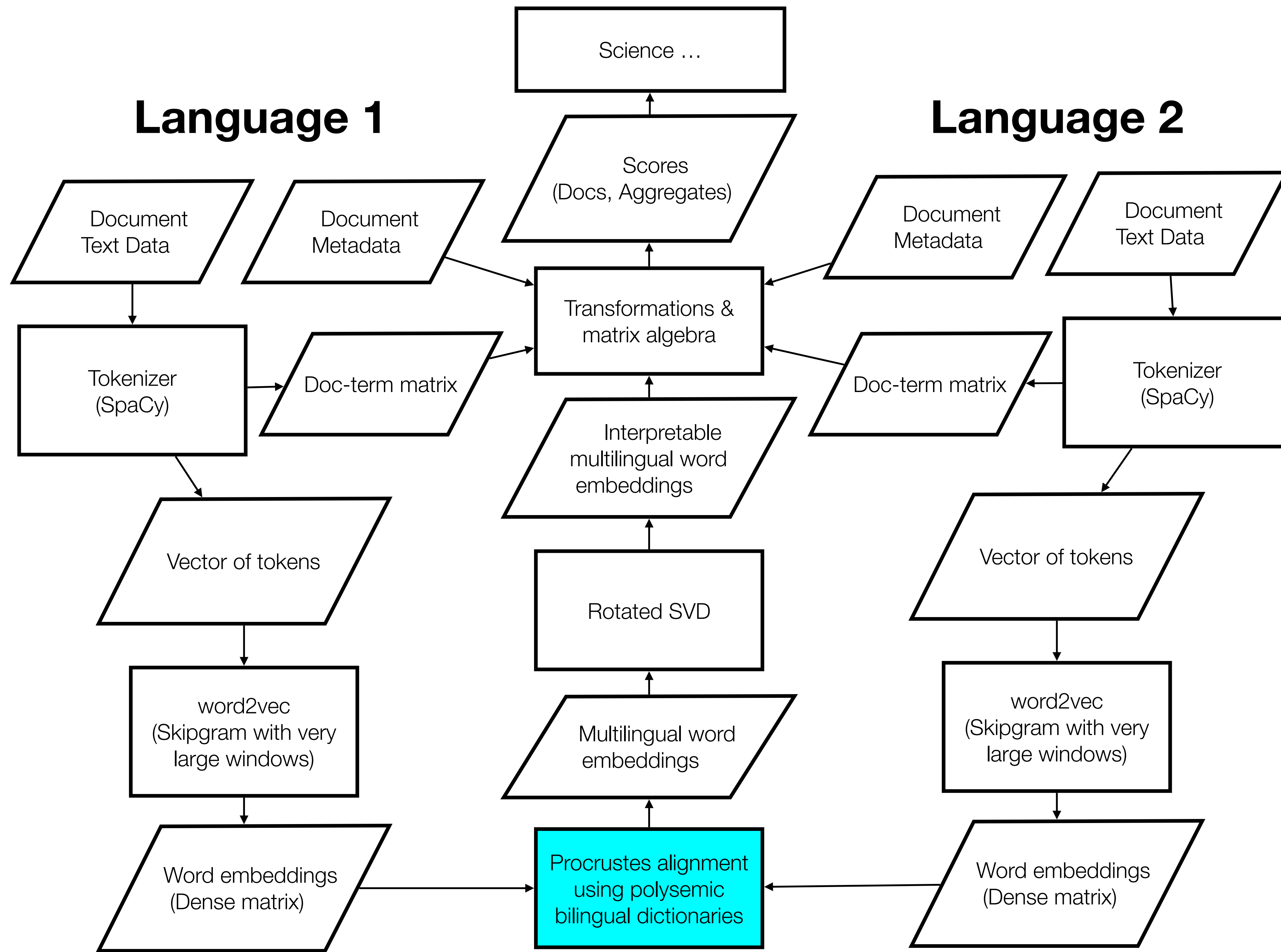


# Monolingual word embeddings, 300 dimensions

---



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



# Prokrustes.



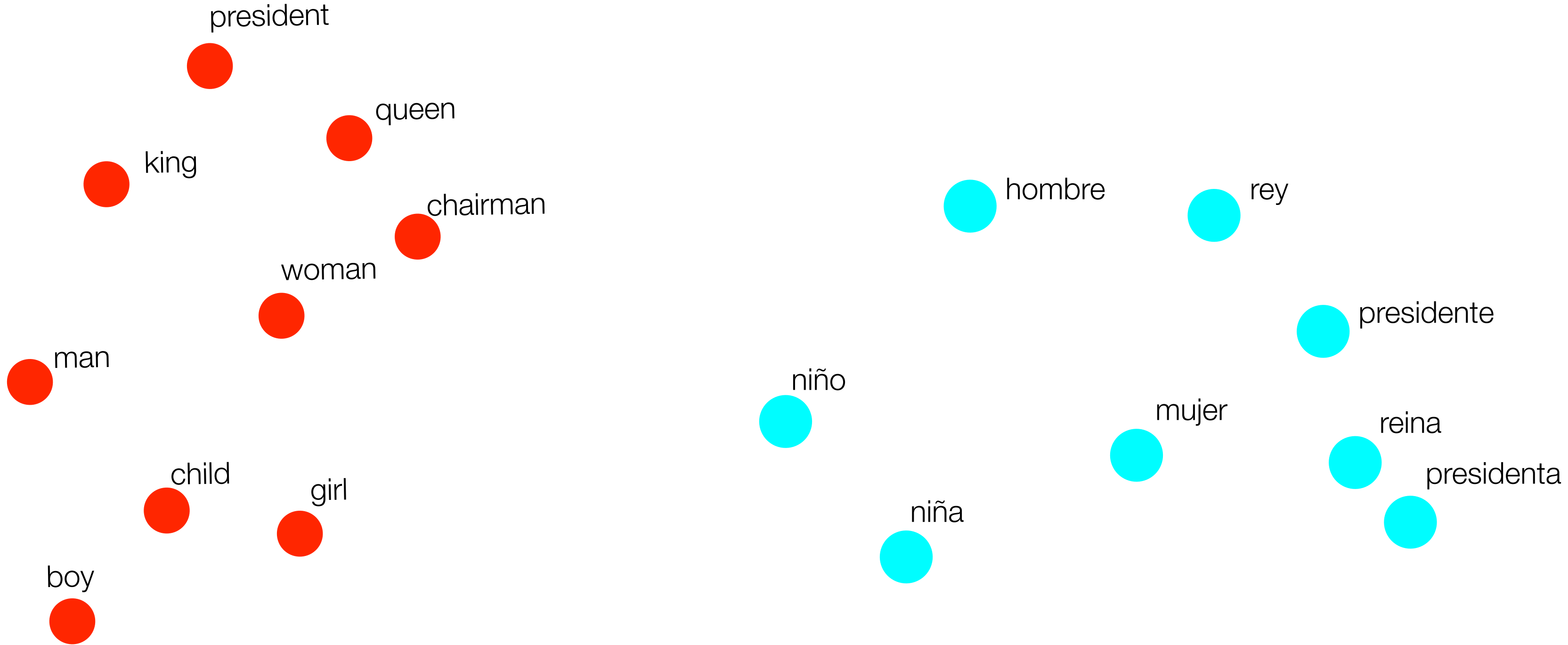
„Wie ich sehe, ist die Freiheit etwas zu groß, — das wollen wir gleich zu ihrer Zufriedenheit abändern!“ (Er haßt ihr die Beine ab.)

# Aligning embeddings through Procrustes analysis

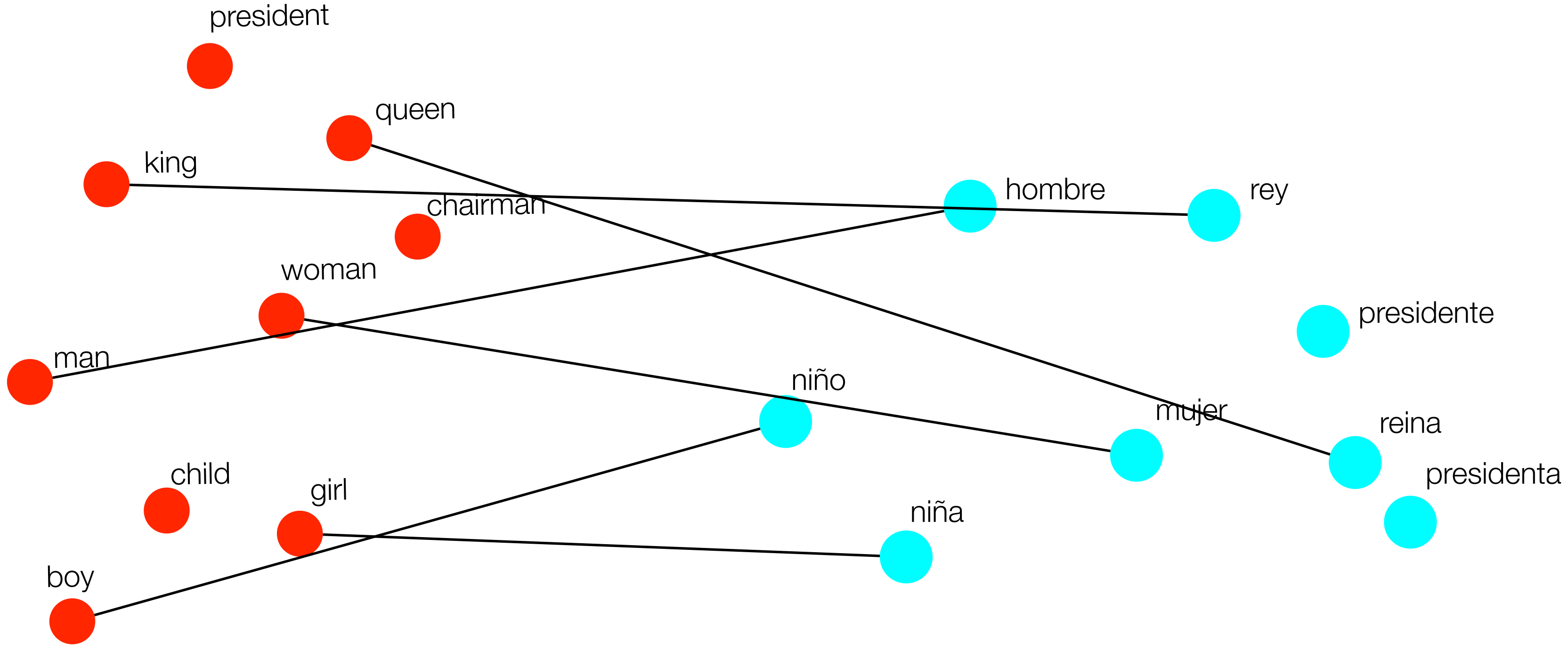
---

- 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.

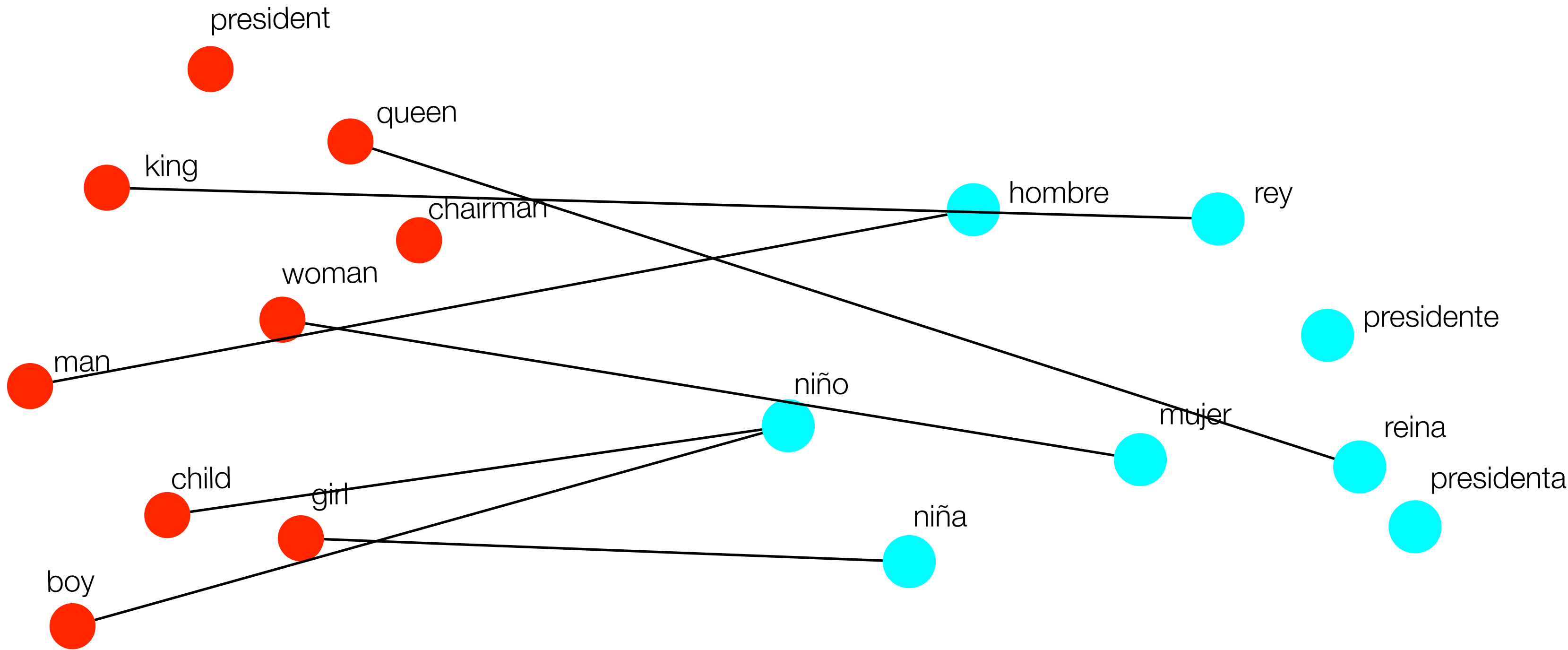
# Procrustes through polysemic bilingual dictionaries



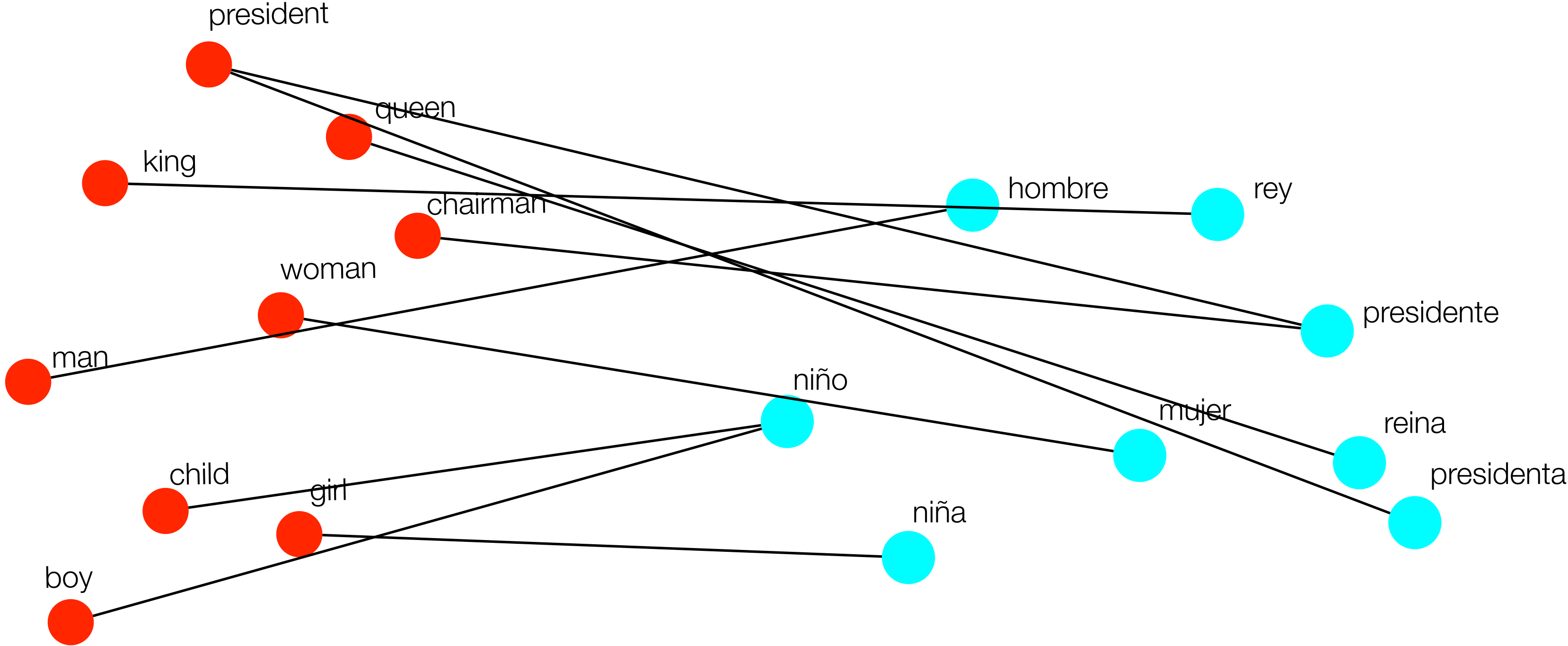
# Procrustes through polysemic bilingual dictionaries



# Procrustes through polysemic bilingual dictionaries

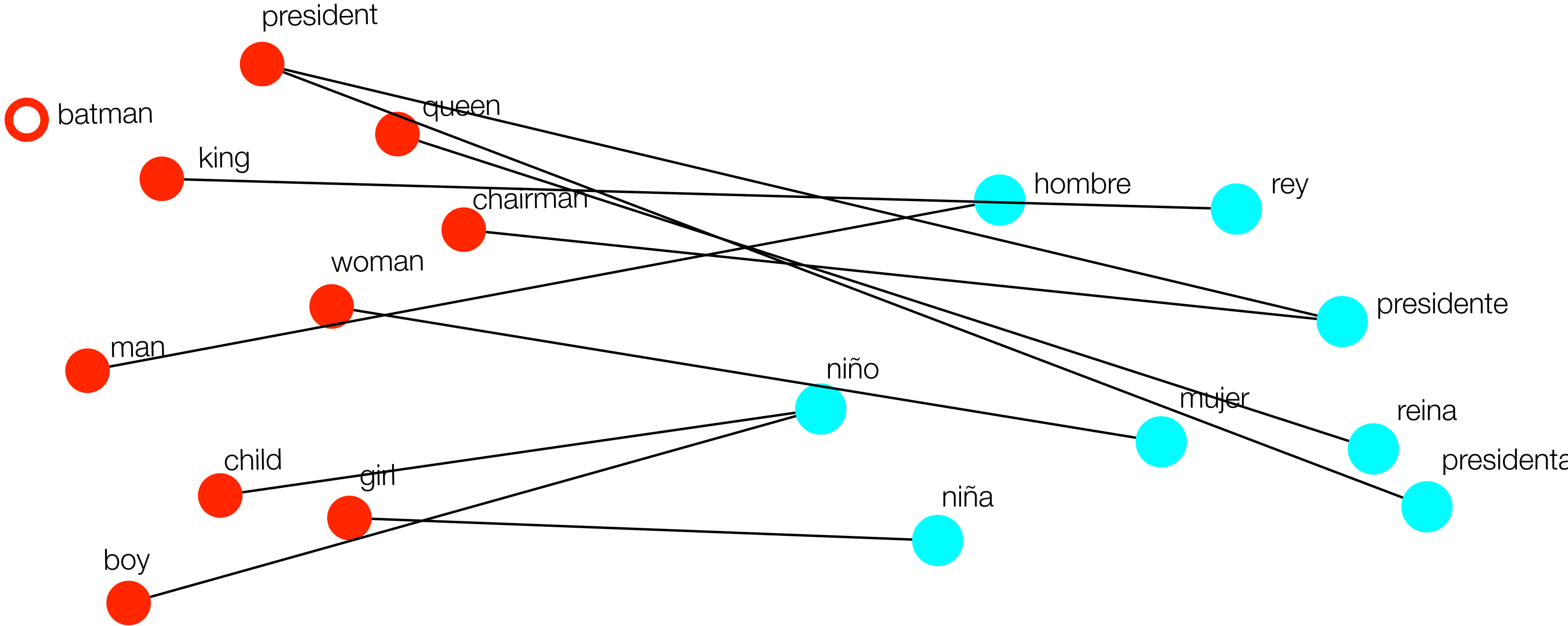


# Procrustes through polysemic bilingual dictionaries

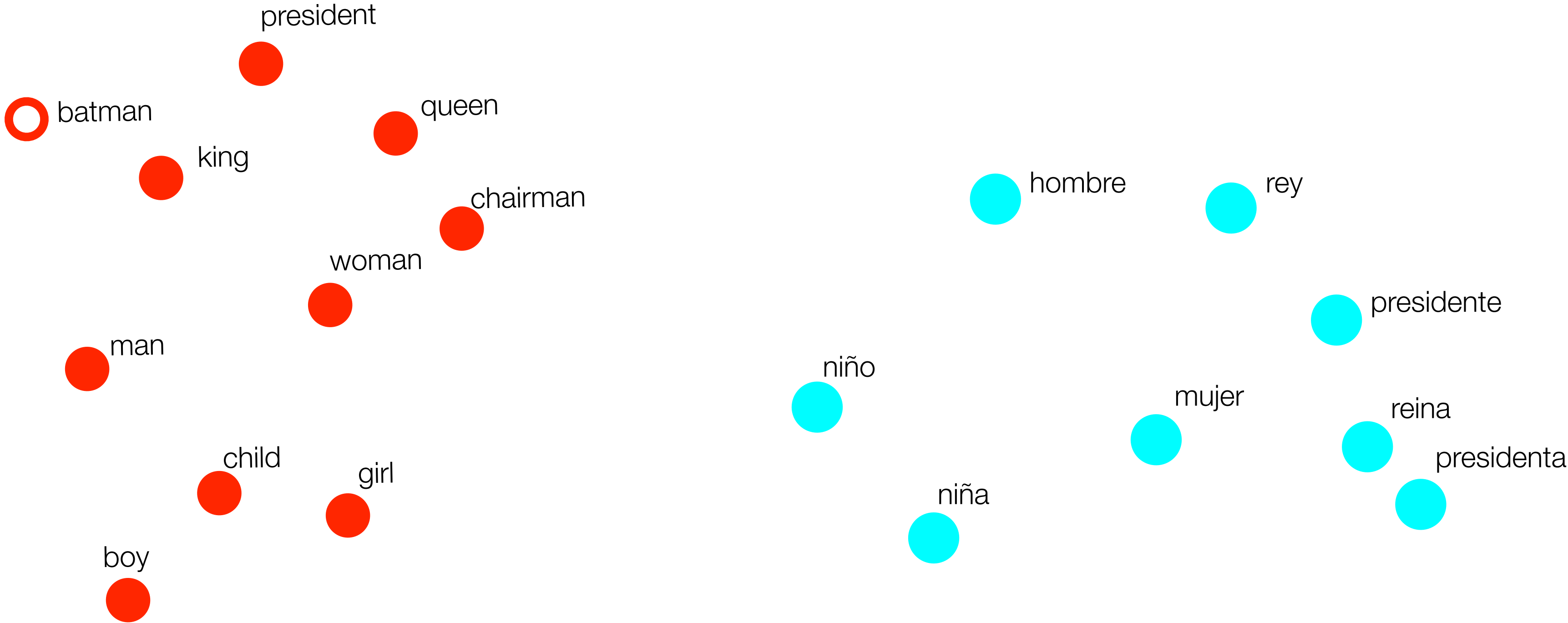




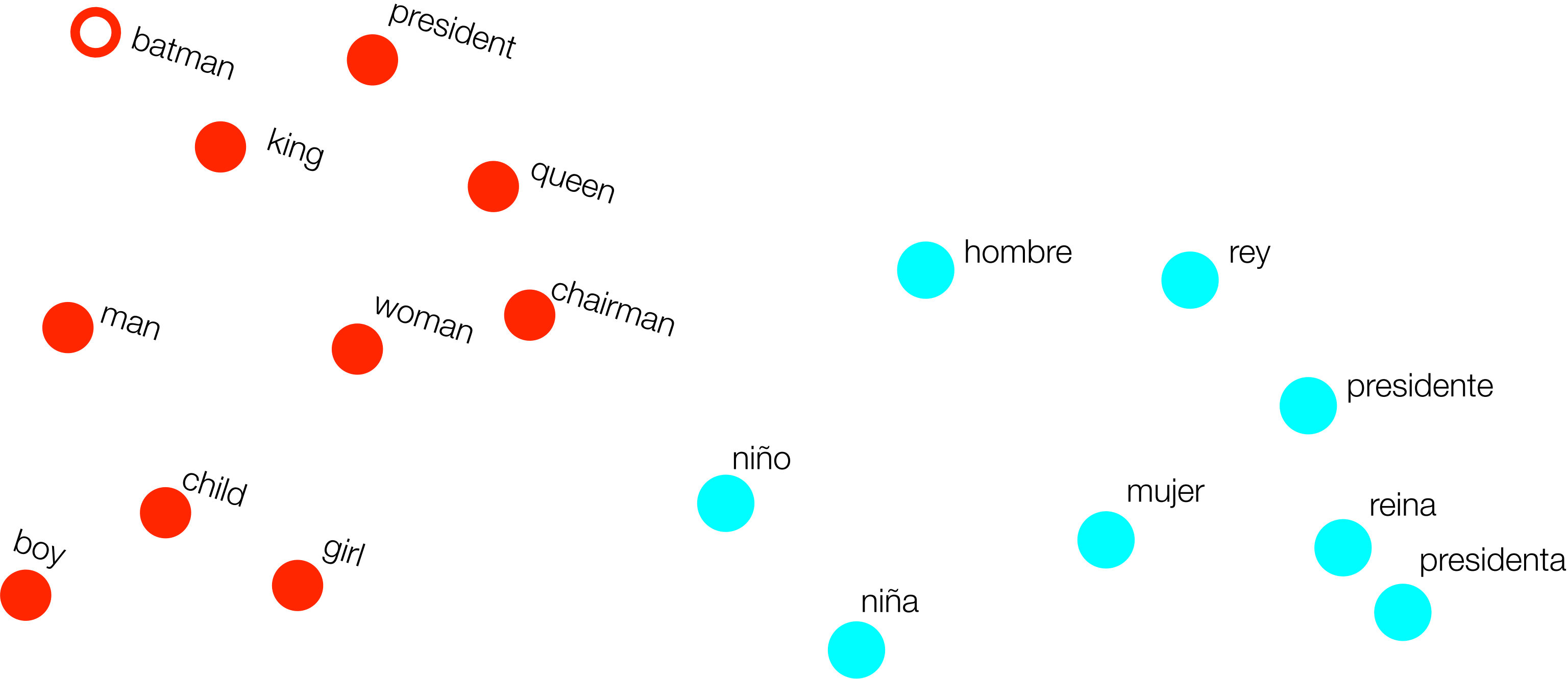
# Procrustes through polysemic bilingual dictionaries



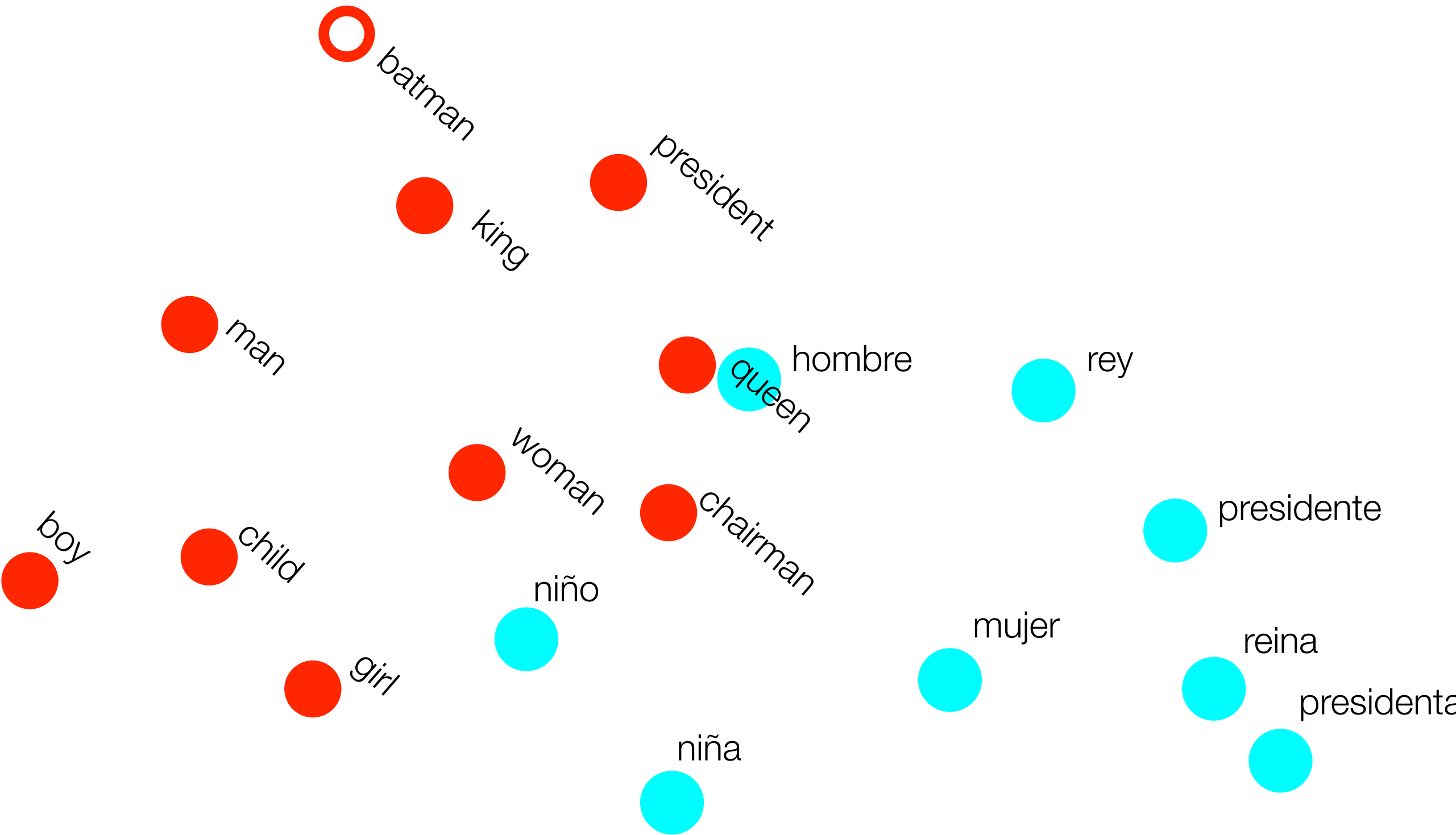
# Procrustes through polysemic bilingual dictionaries



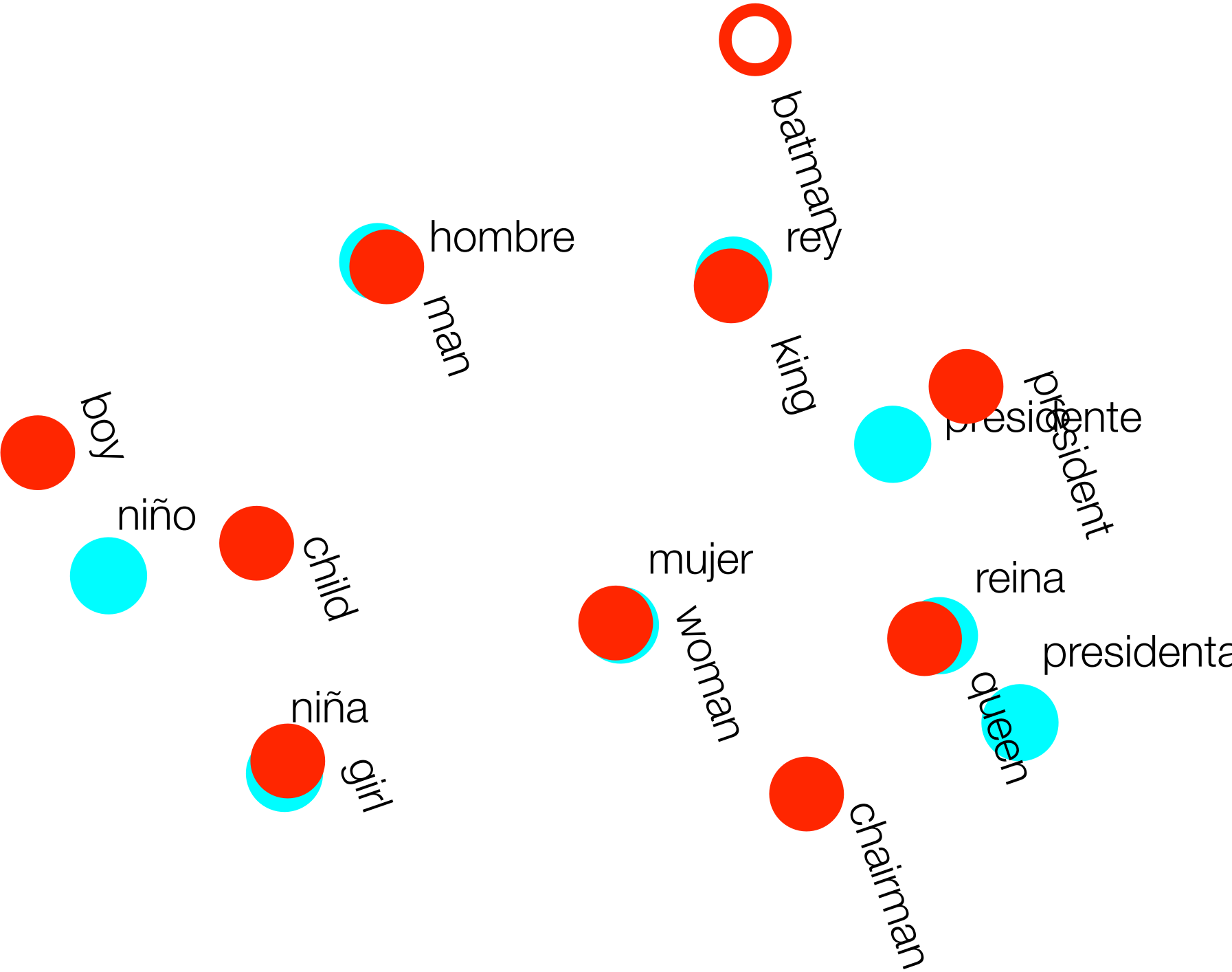
# Procrustes through polysemic bilingual dictionaries



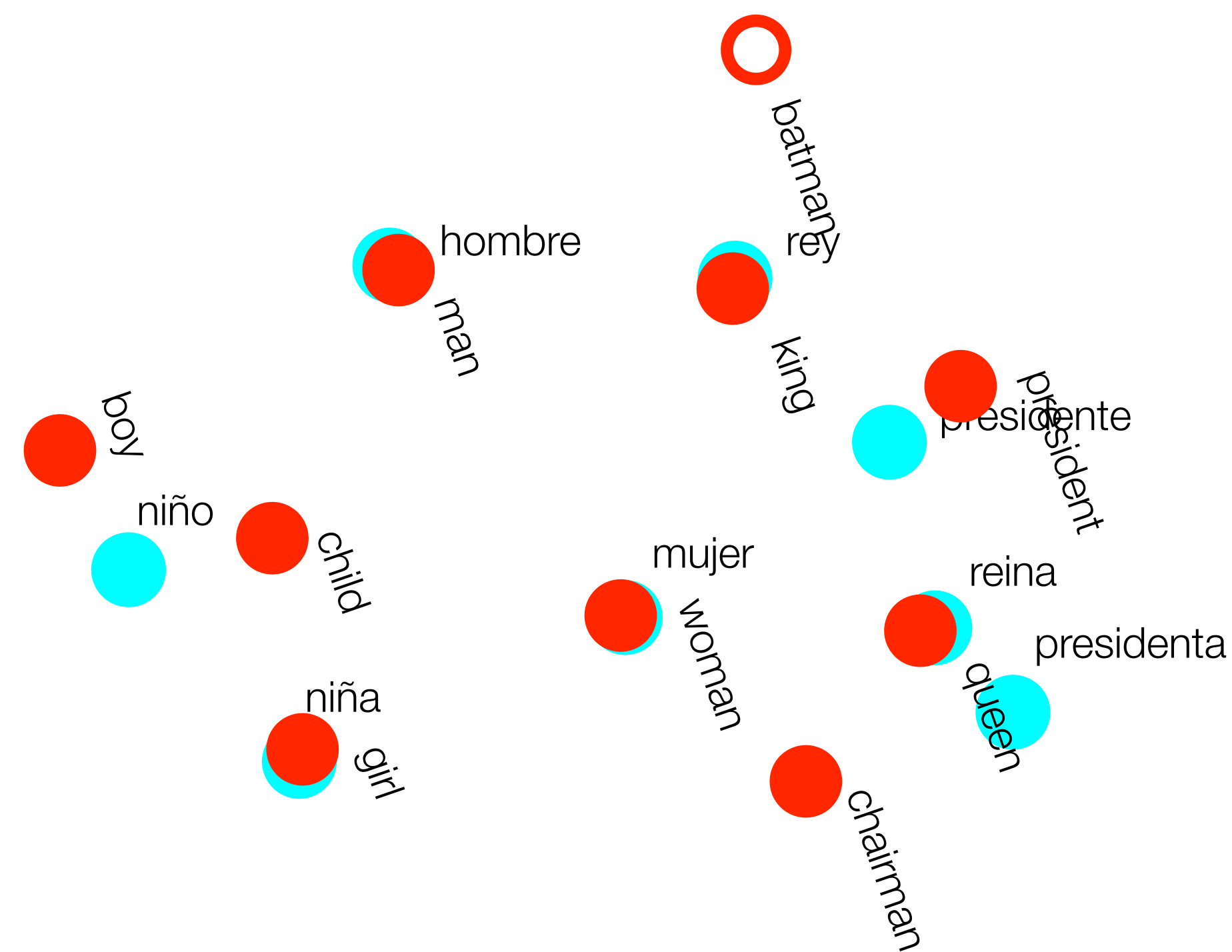
# Procrustes through polysemic bilingual dictionaries



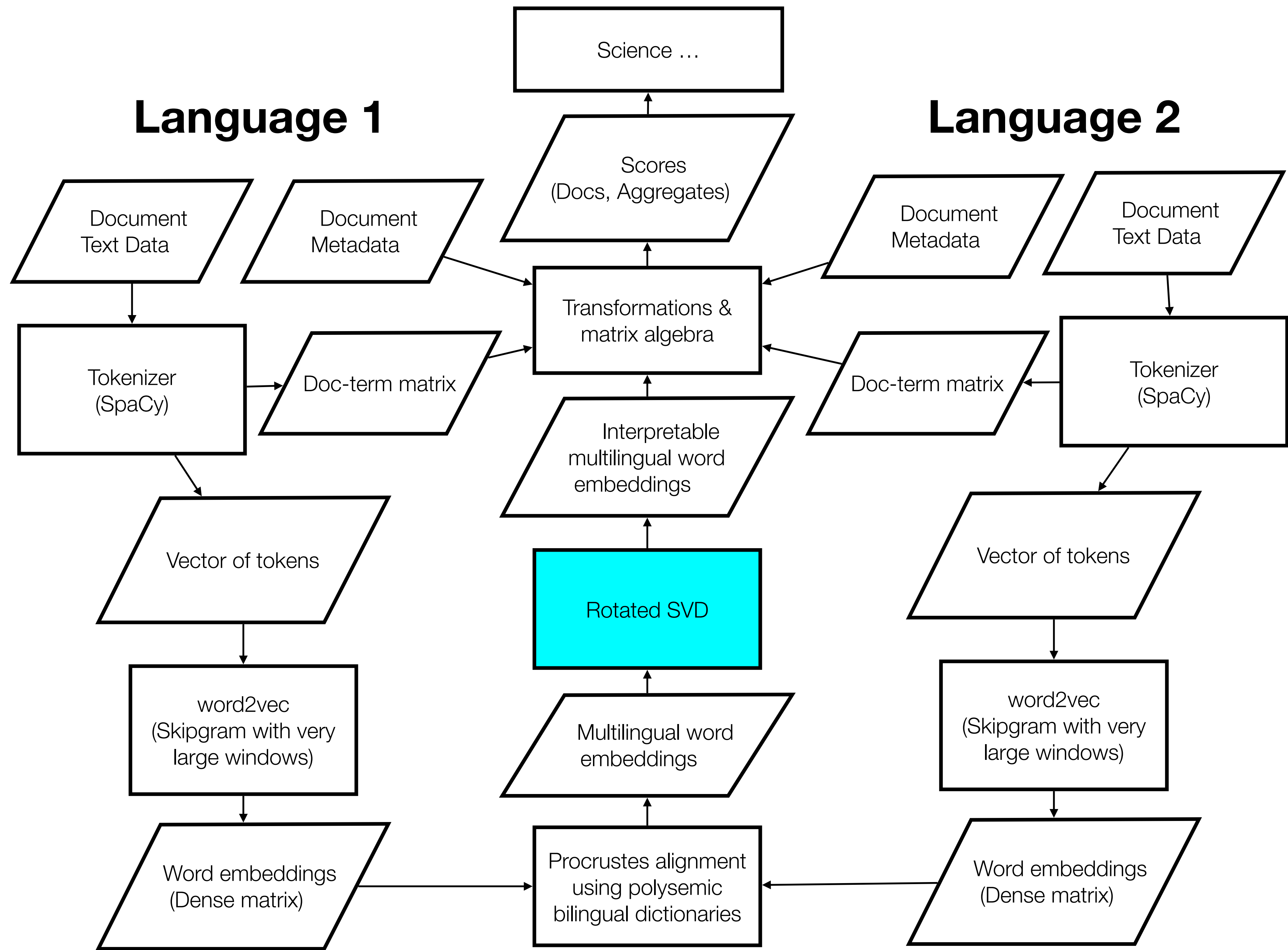
# Procrustes through polysemic bilingual dictionaries



# 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.)



# 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 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 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.

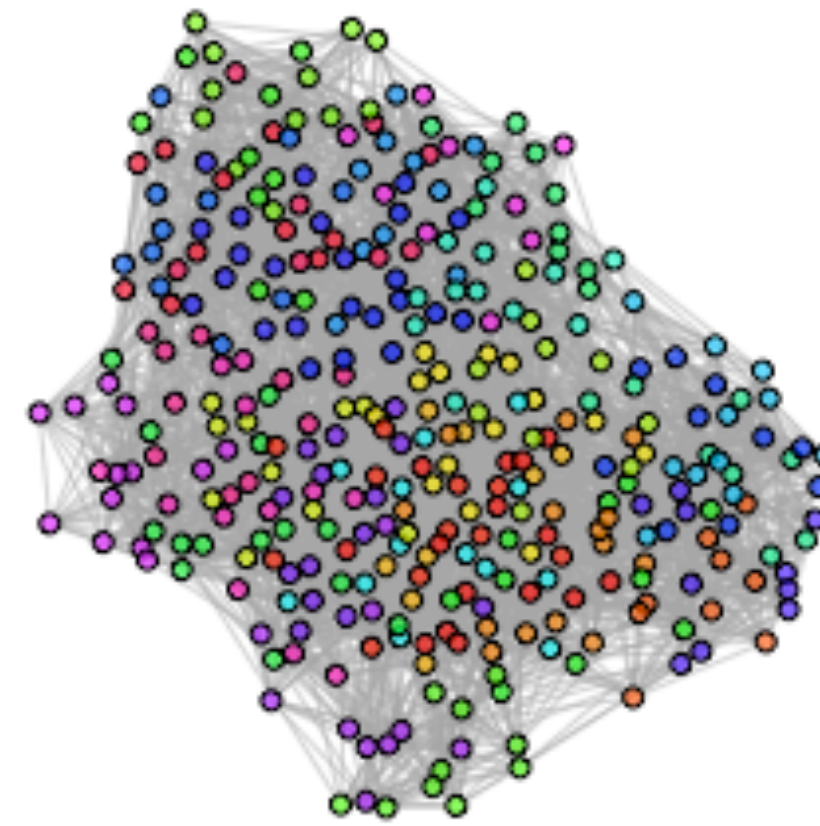


Sidebar:

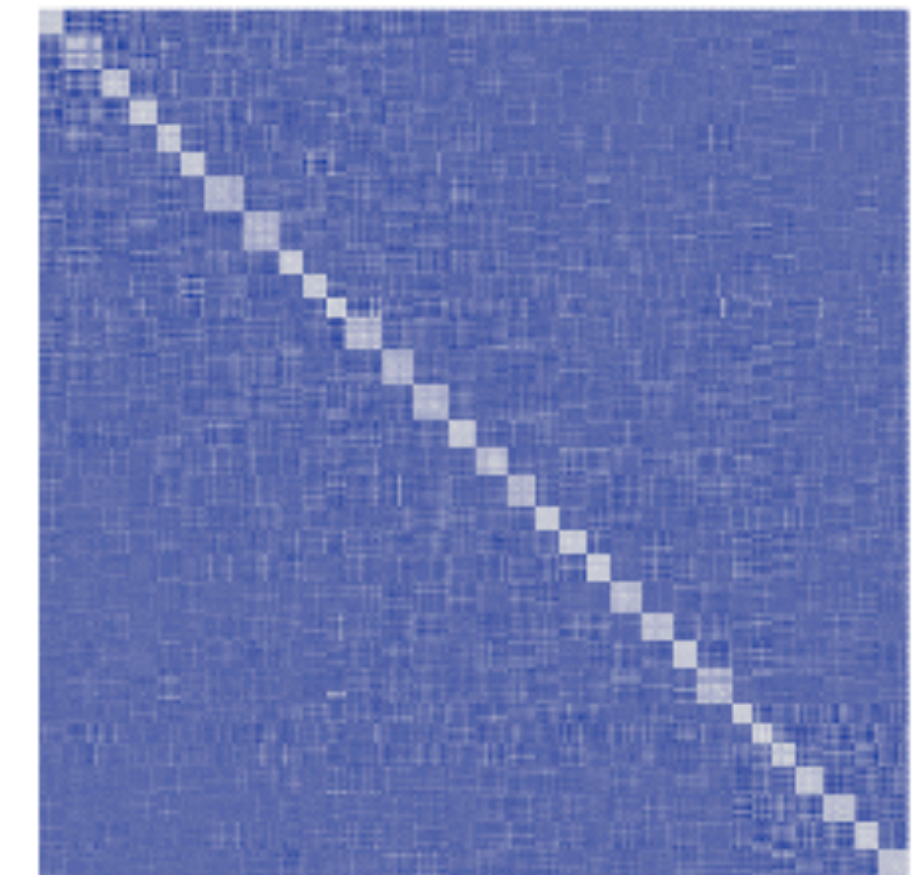
The Little Jiffy is dope.

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.

NCAA Basketball Games Played, Midseason



Jiffy Communities



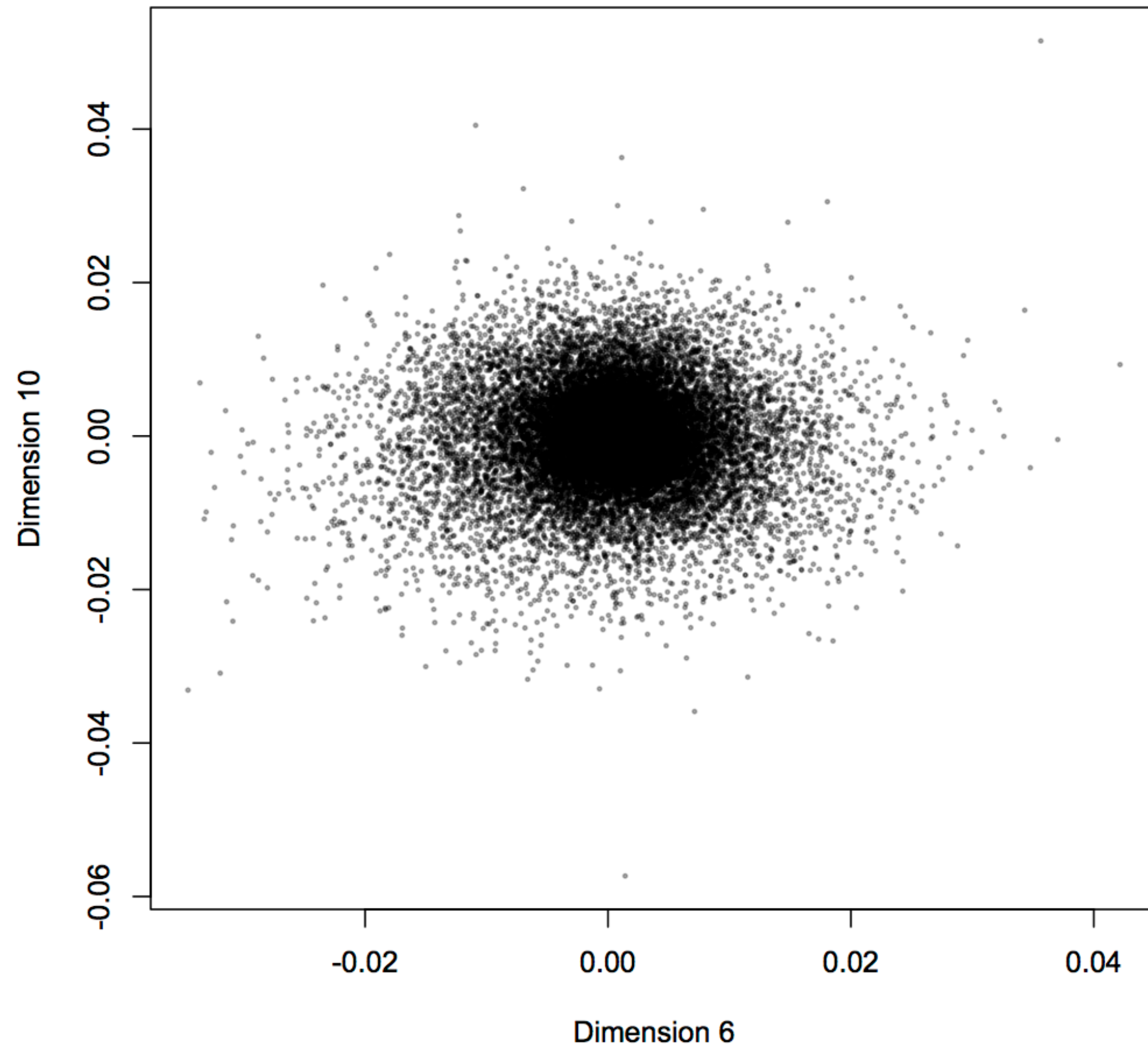
Performance on Basketball Conference Detection

	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.	—

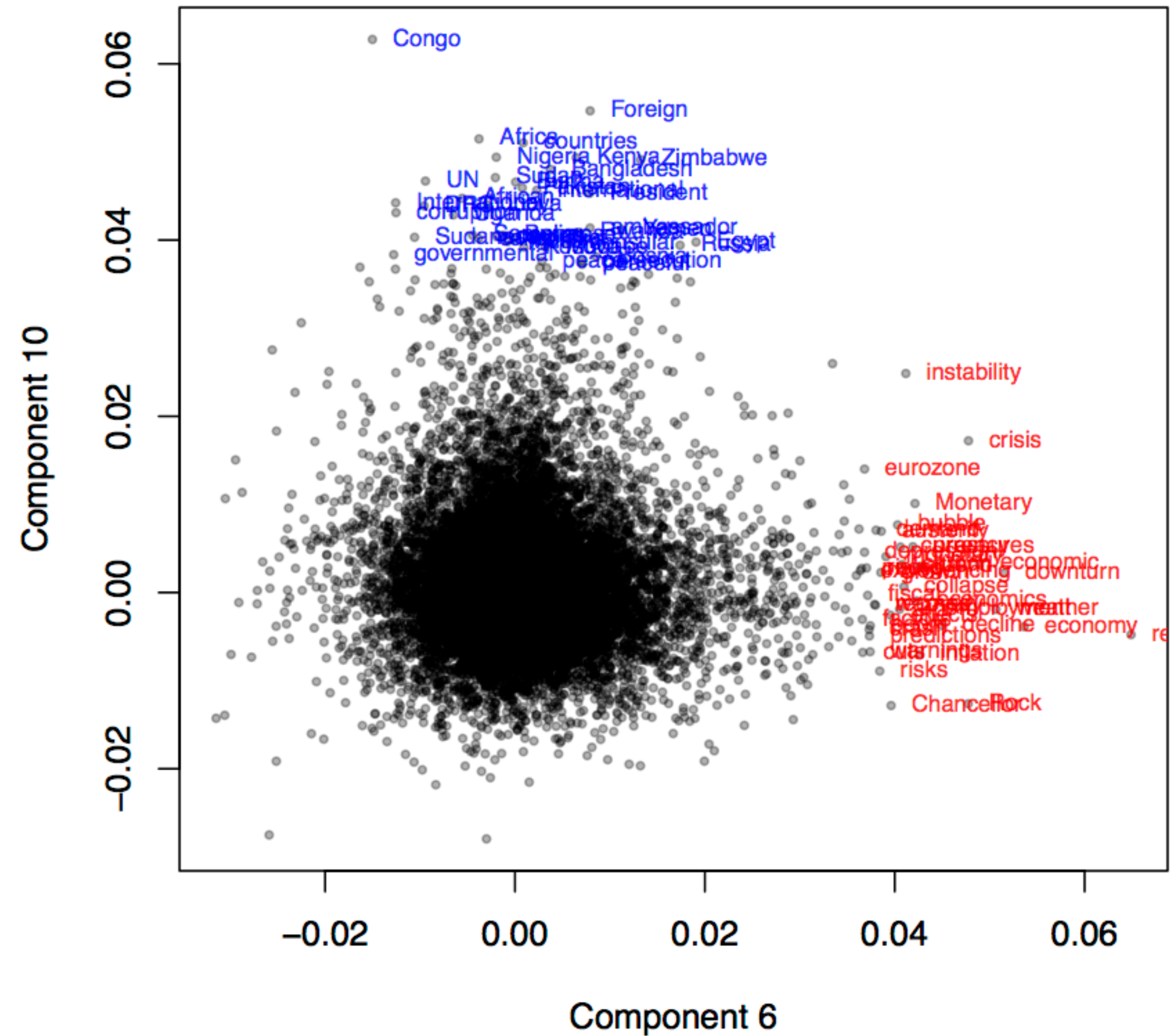
\* Coded by Monroe in R. Same eigengap method. All others in igraph in R.

# Effect of rotation

Unrotated SVD of embeddings



Rotated Loadings, English – 50 Dimensions



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	
Top terms	Top English terms
de_Terrorismus	en_terrorism
it_po	en_terrorist
it_terrorismo	en_threat
en_terrorism	en_intelligence
sv_terrorism	en_terrorists
en_terrorist	en_terror
nl_AIVD	en_Terrorism
nl_terroristische	en_intercept
it_ndrangheta	en_Intelligence
en_threat	en_extremism
de-Anschläge	en_counter
en_intelligence	en_proscription
de_Sicherheitsbehörden	en_TPIMs
sv_terrorismen	en_ISC
en_terrorists	en_orders
de_Verfassungsschutz	en_Carlile
sv_terrorister	en_MI5
nl_dreiging	en_al
sv_Svar	en_7/7
nl_aanslagen	en_9/11

Environment	
Top terms	Top English terms
de_Naturschutz	en_environmental
es_ambiental	en_Natural
sv_miljöbalken	en_wildlife
de_Umwelt	en_biodiversity
de_Natur	en_beauty
en_environmental	en_natural
en_Natural	en_environment
fi_luonnon	en_Environment
cs_prostředí	en_conservation
en_wildlife	en_Protection
de_Flächen	en_fracking
es_ambiente	en_rivers
es_medioambiental	en_pollution
es_Ambiente	en_trees
en_biodiversity	en_countryside
cs_přírody	en_reservoir
sv_biologisk	en_habitats
sv_naturen	en_Antarctic
en_beauty	en_species
es_Costas	en_planning

Macroeconomy	
Top terms	Top English terms
de_Konjunktur	en_deficit
de_Wachstum	en_fiscal
it_debito	en_economy
es_déficit	en_borrowing
en_deficit	en_finances
es_crecimiento	en_recession
sv_finanser	en_economic
es_economía	en_recovery
es_inflación	en_euro
sv_räntan	en_growth
de_Finanzpolitik	en_Monetary
it_deficit	en_monetary
en_fiscal	en_eurozone
sv_ekonomi	en_inflation
de_Stabilitäts-	en_Chancellor
sv_finanserna	en_debt
es_Monetario	en_currency
de_Neuverschuldung	en_Budget
de_konjunkturellen	en_borrow
es_Estabilidad	en_billion

Immigration	
Top terms	Top English terms
es_inmigrantes	en_immigration
en_immigration	en_asylum
es_extranjeros	en_migrants
de_Zuwanderung	en_Immigration
es_inmigración	en_migration
en_asylum	en_immigrants
en_migrants	en_visa
de_A Asylbewerber	en_citizenship
en_Immigration	en_refugees
en_migration	en_Border
sv_flyktingpolitik	en_visas
it_soggiorno	en_migrant
es_asilo	en_refugee
de_A Asylrecht	en_nationals
es_extranjería	en_passport
sv_invandring	en_illegally
de_Aufenthaltsrecht	en_deportation
en_immigrants	en_deported
de_A Asyl	en_Borders
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	finer; 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_iniziativa	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_communicatie	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	en_fails
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_belofte	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

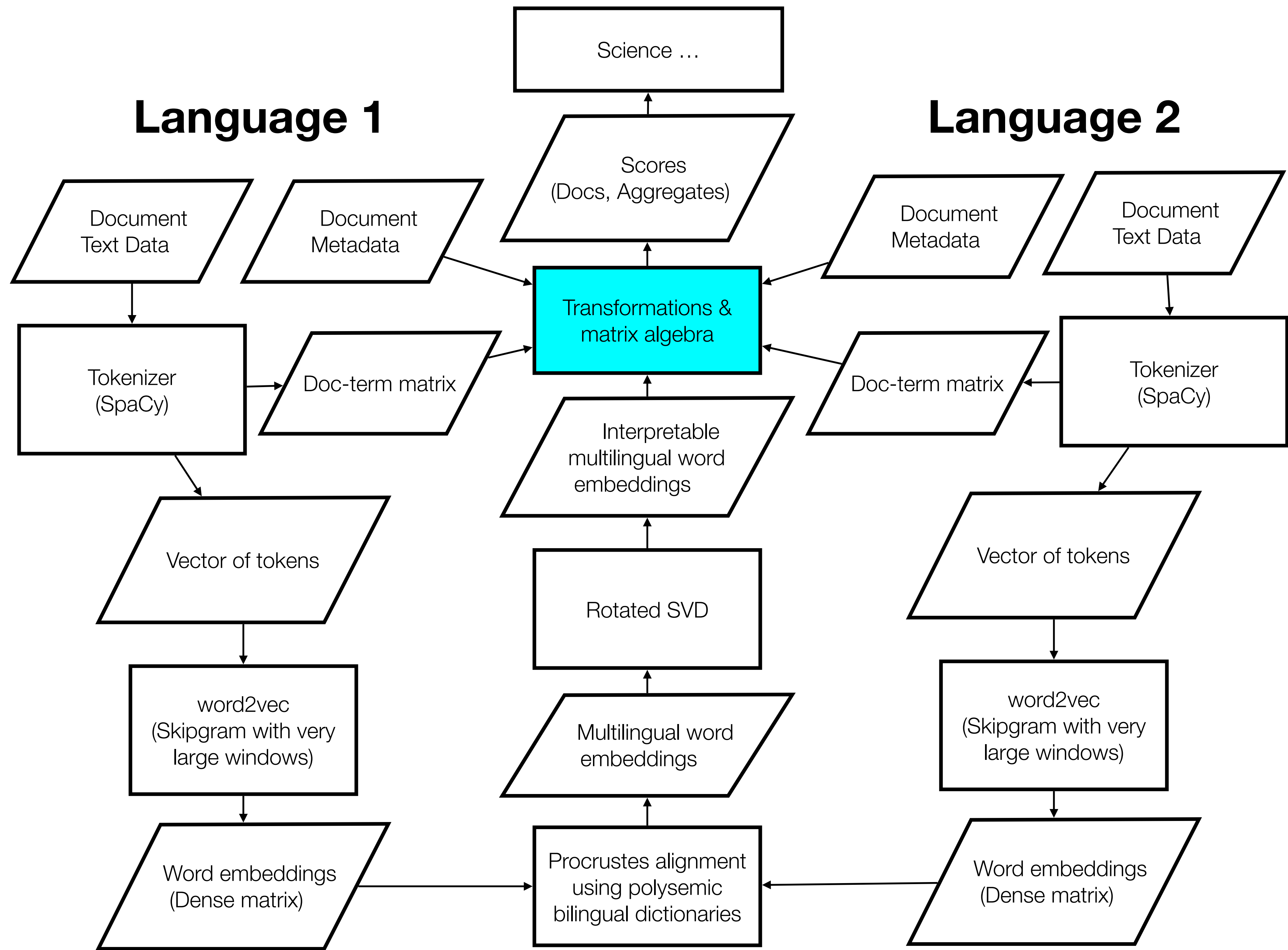
nl\_Sipkes

nl\_Kalsbeek

nl\_Ouwehand

nl\_Leijten

nl\_Sterk

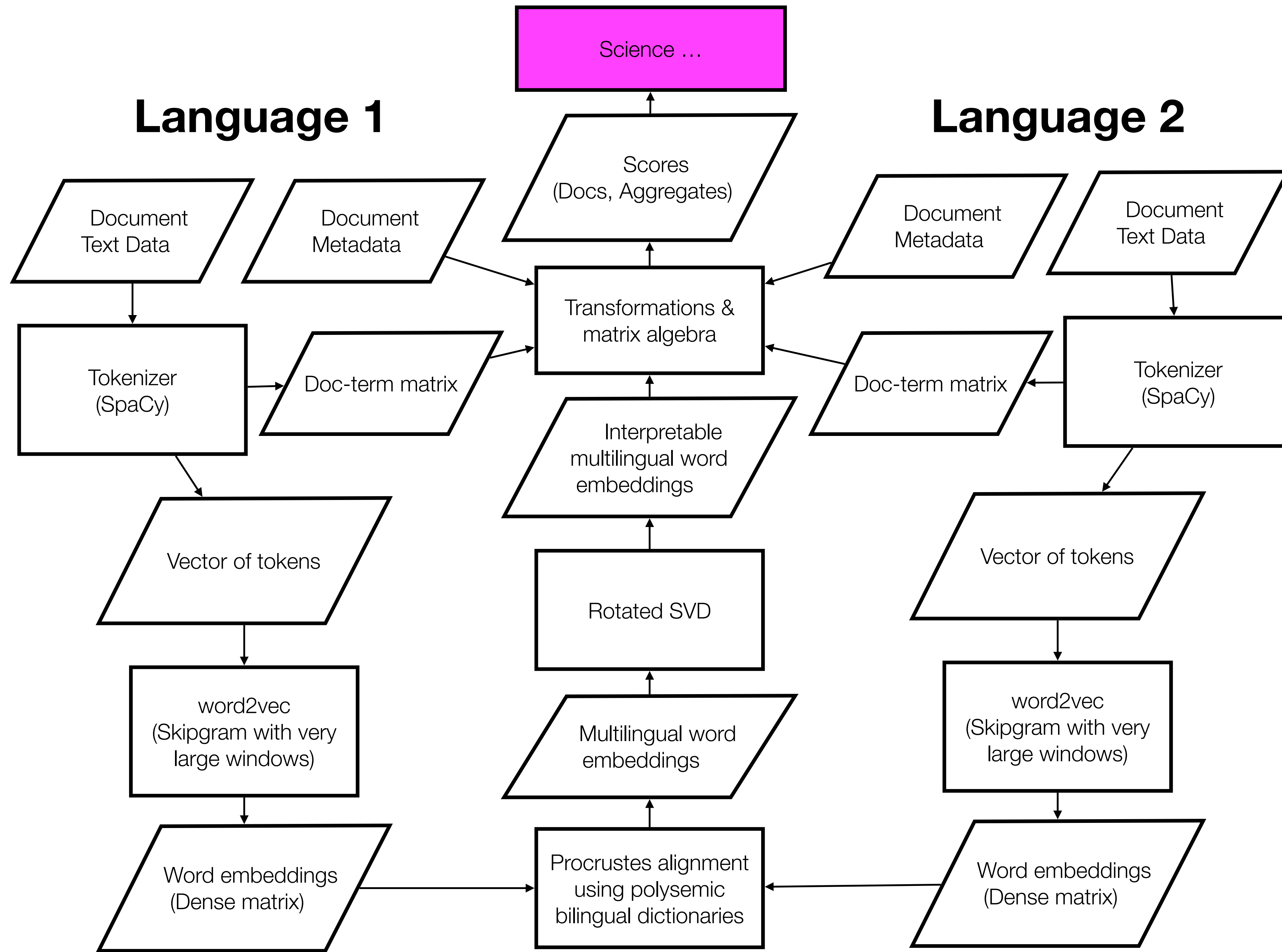


# Scoring documents and aggregates of documents (like parties)

---

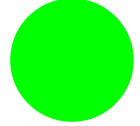
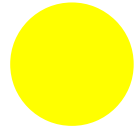
- Represent documents with regularized and weighted doc-token PPMI. This is in essence, “Fightin’ Words” (Monroe, et al. 2008).
- Raw topic score = cosine (doc-vector, topic-vector).
- Norm (L2) scores across topics/dimensions of interest (e.g., substantive topics).
- For groupings like party or party-year,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



# All graphics color parties by Comparative Manifesto Project family

---

CMP Party Family	
	Ecological
	Left
	Social democratic
	Liberal
	Christian
	Conservative
	Nationalist
	Ethnic/Regional

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.

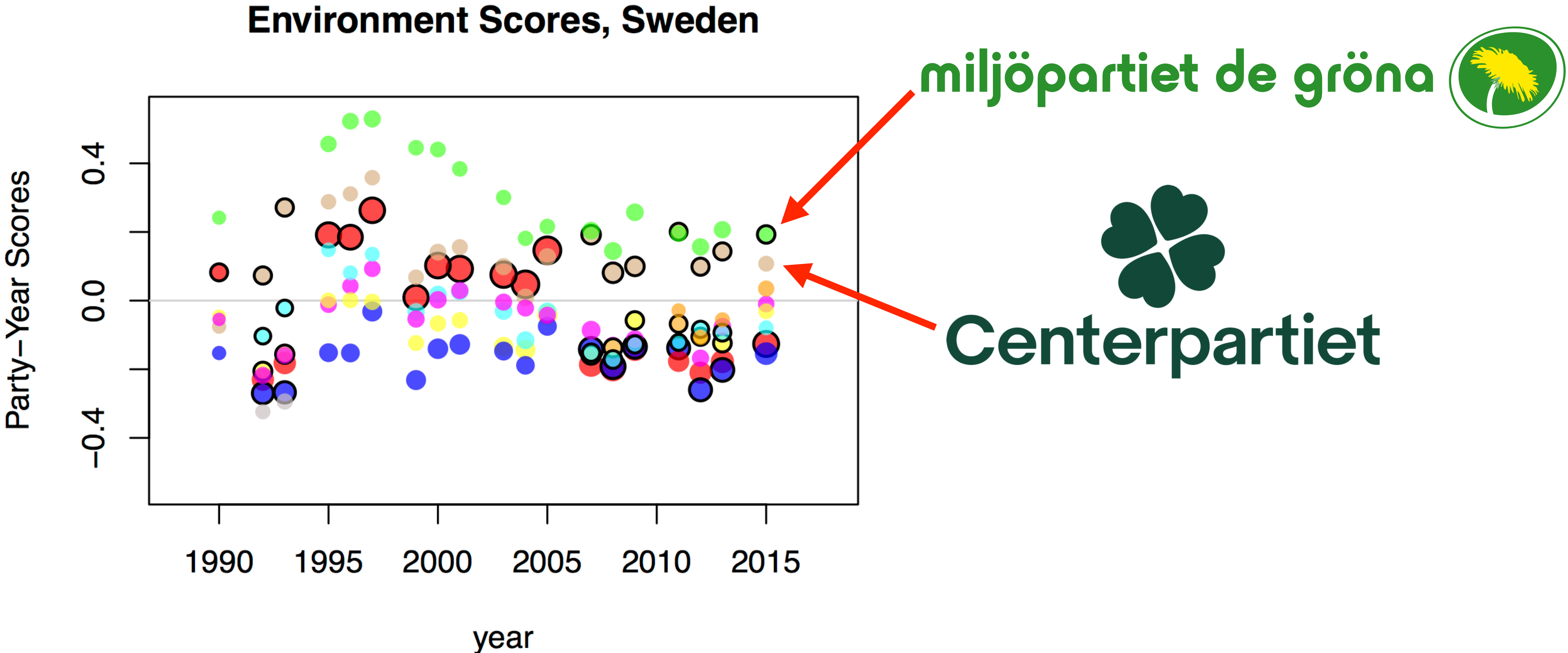
Sanity checks —

Do parties dominate the issues they are expected to?

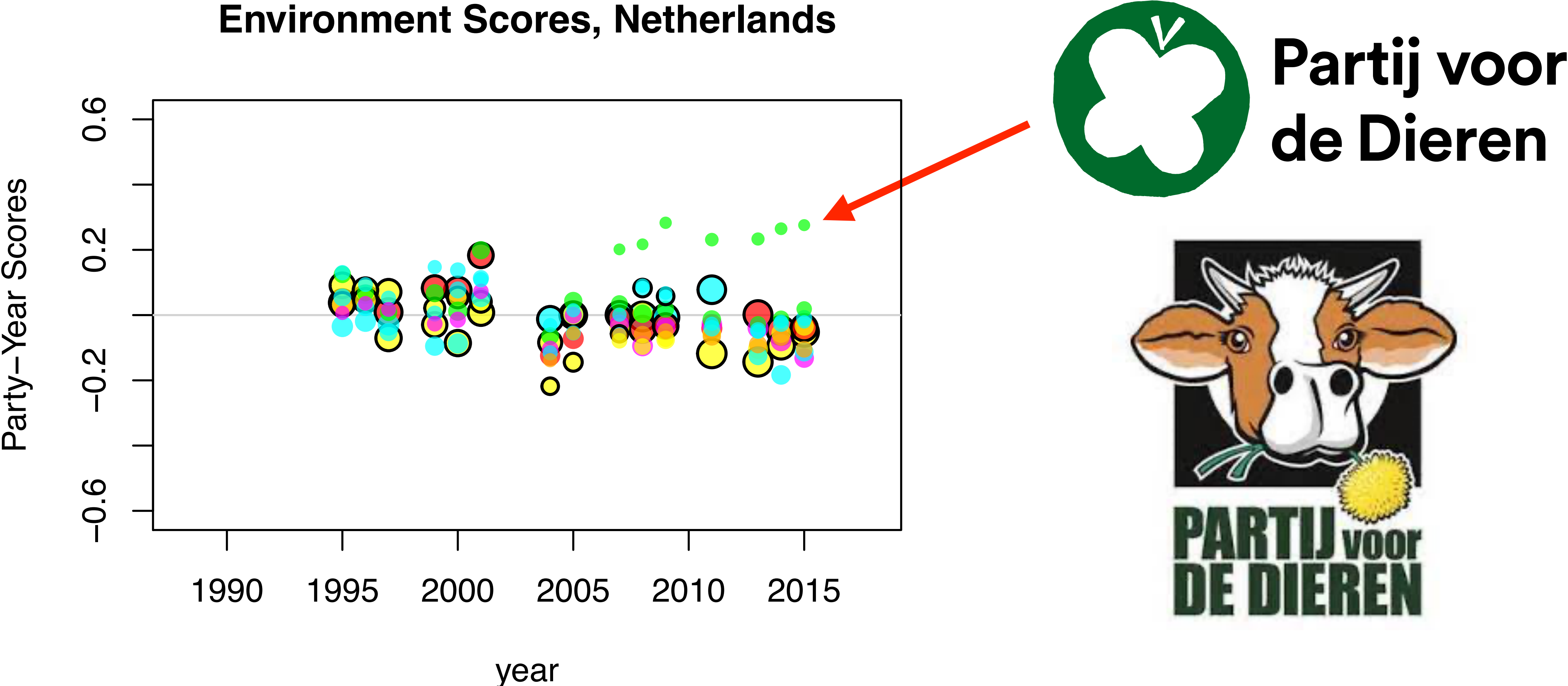
Do movements over time correspond to events?

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

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



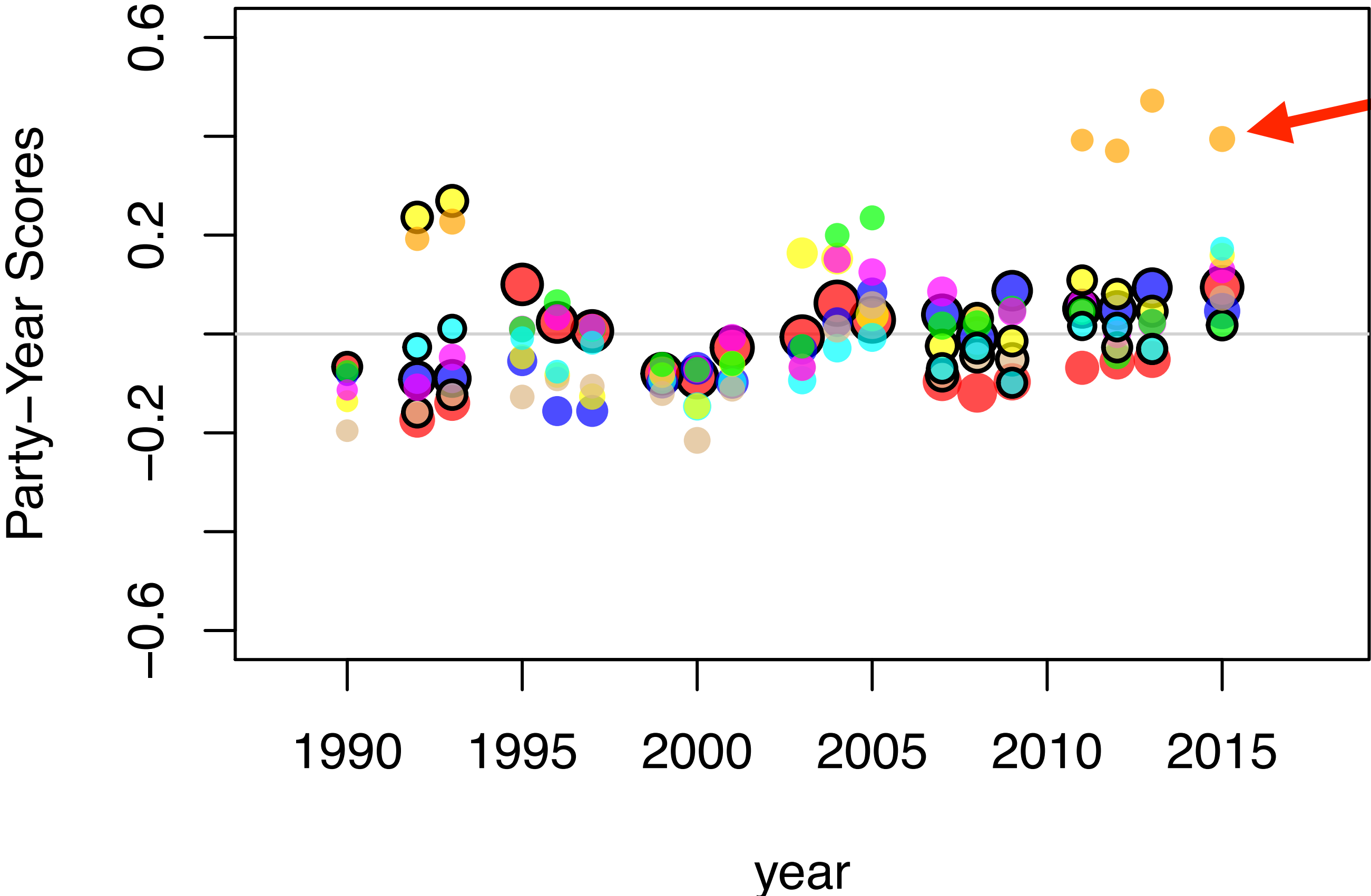
# Scores of Dutch parties on the “Environment” topic over time





# Scores in Sweden on the “Immigration” topic over time

## Immigration Scores, Sweden



“Swedes first!”

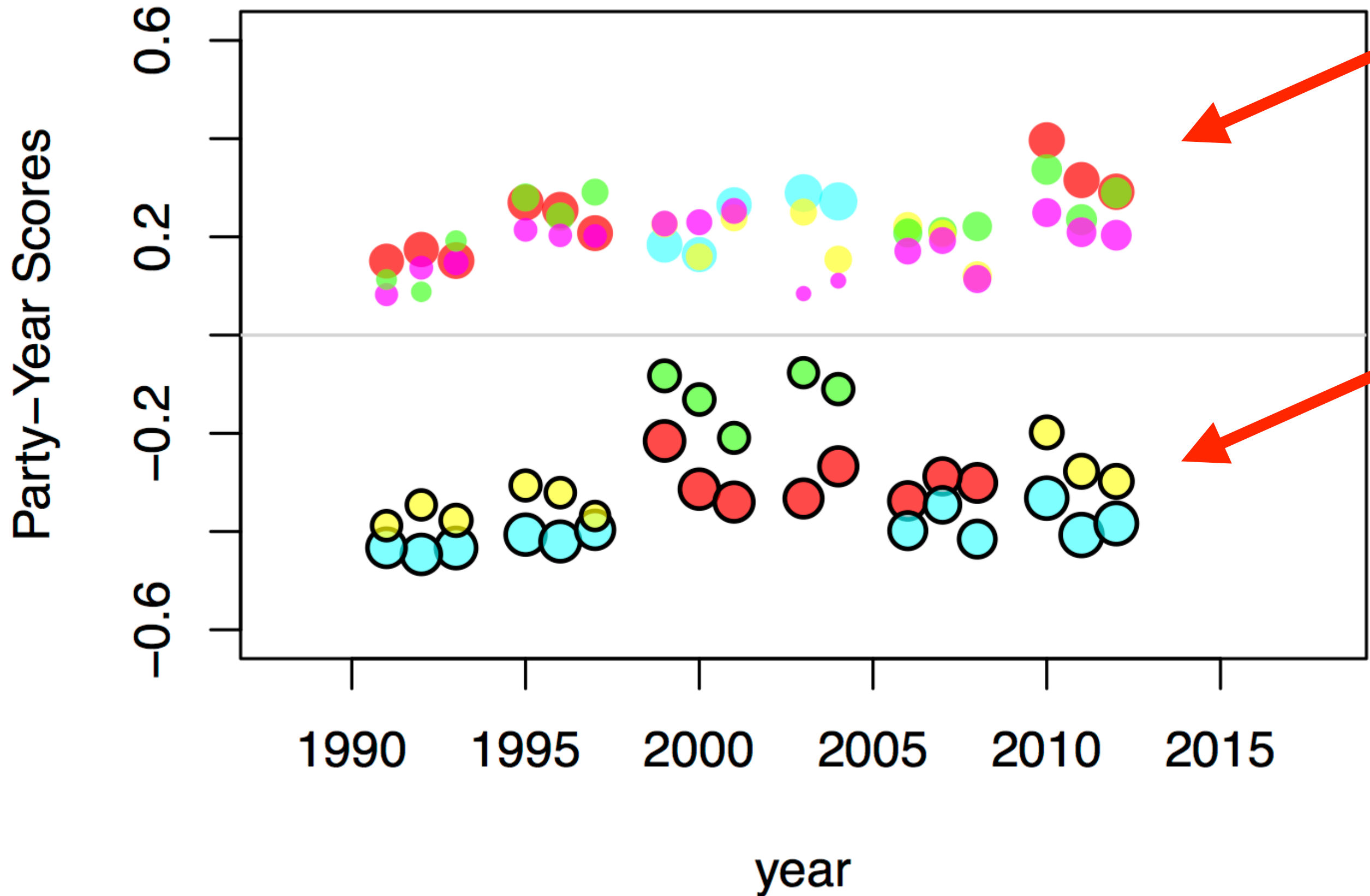


“Keep Sweden Swedish”

We can characterize how oppositions and governments behave

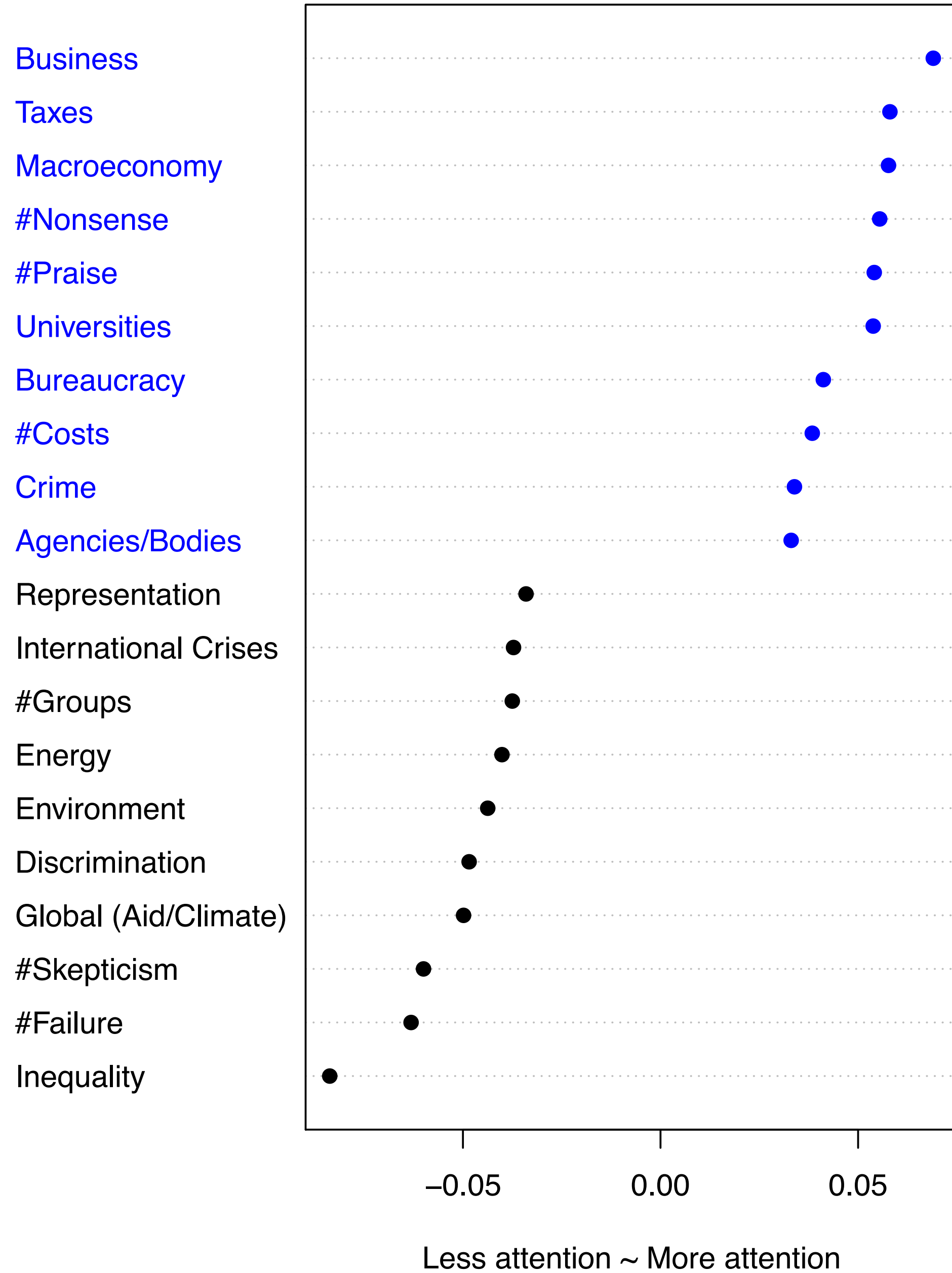
# Where you stand is where you sit ...

## #Failure Scores, Germany

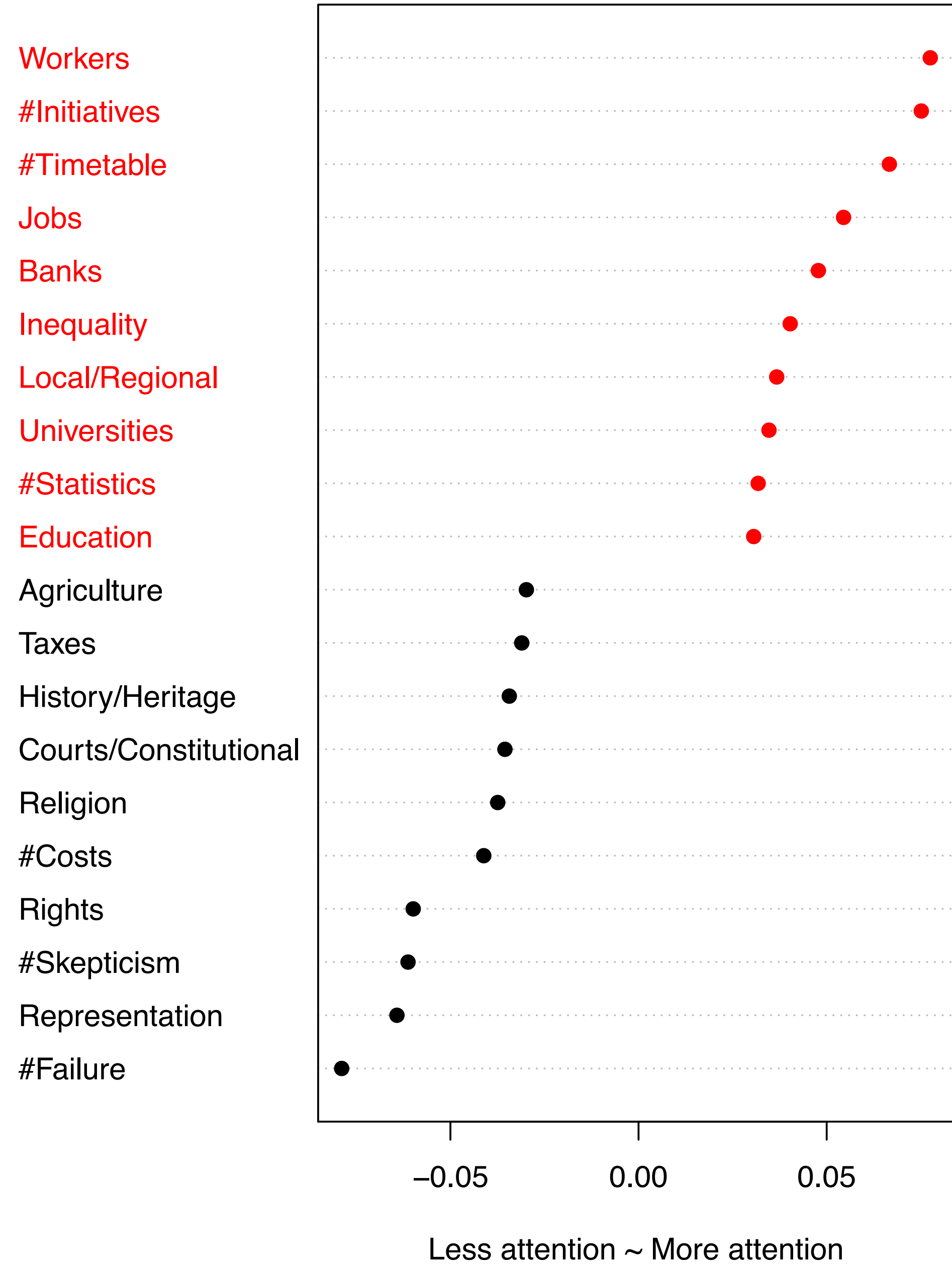


What topics do parties of particular families give most emphasis, relative to other parties?

### Conservative parties vs. others

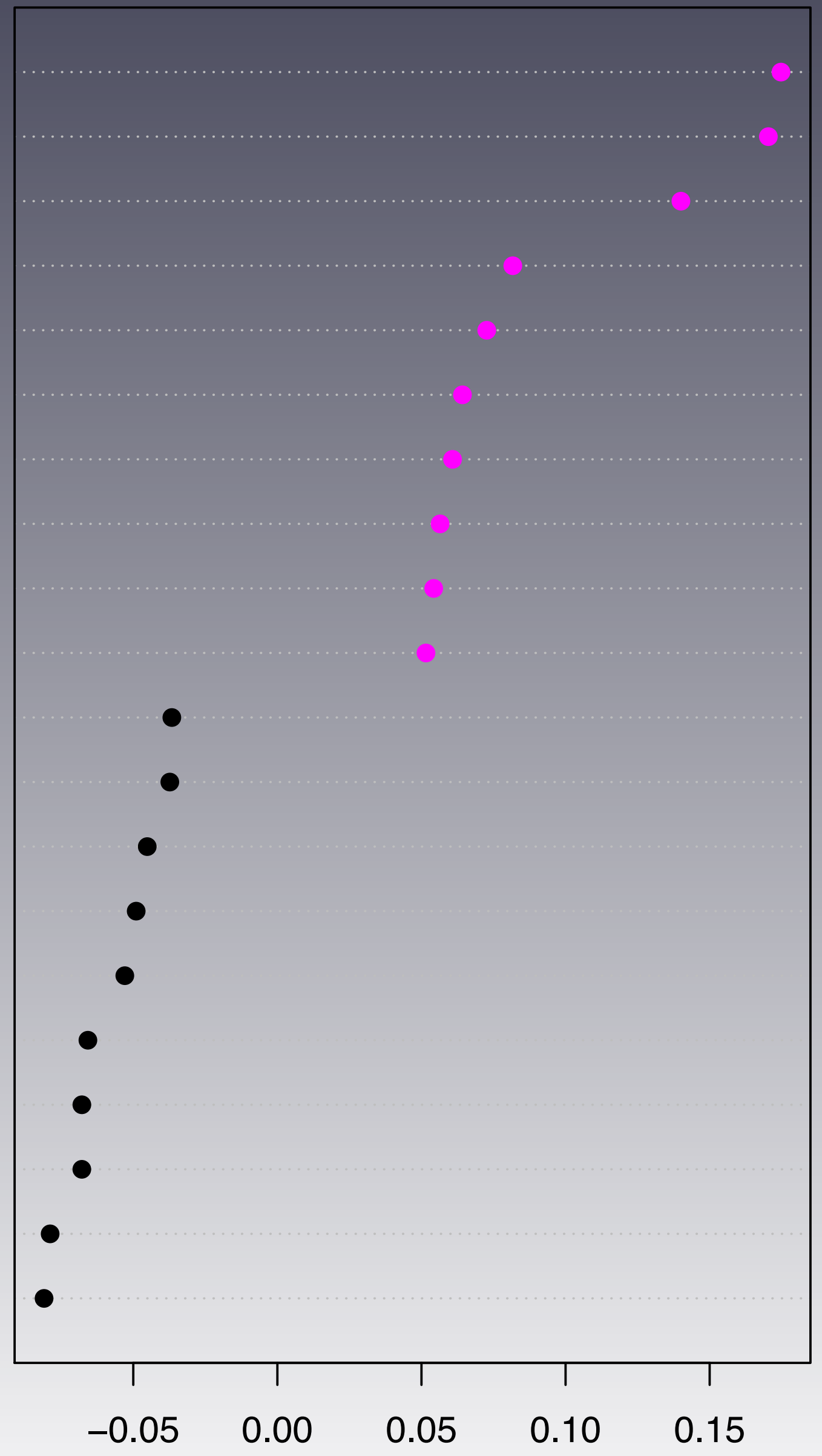


### Social Democratic parties vs. others



Left parties vs. others

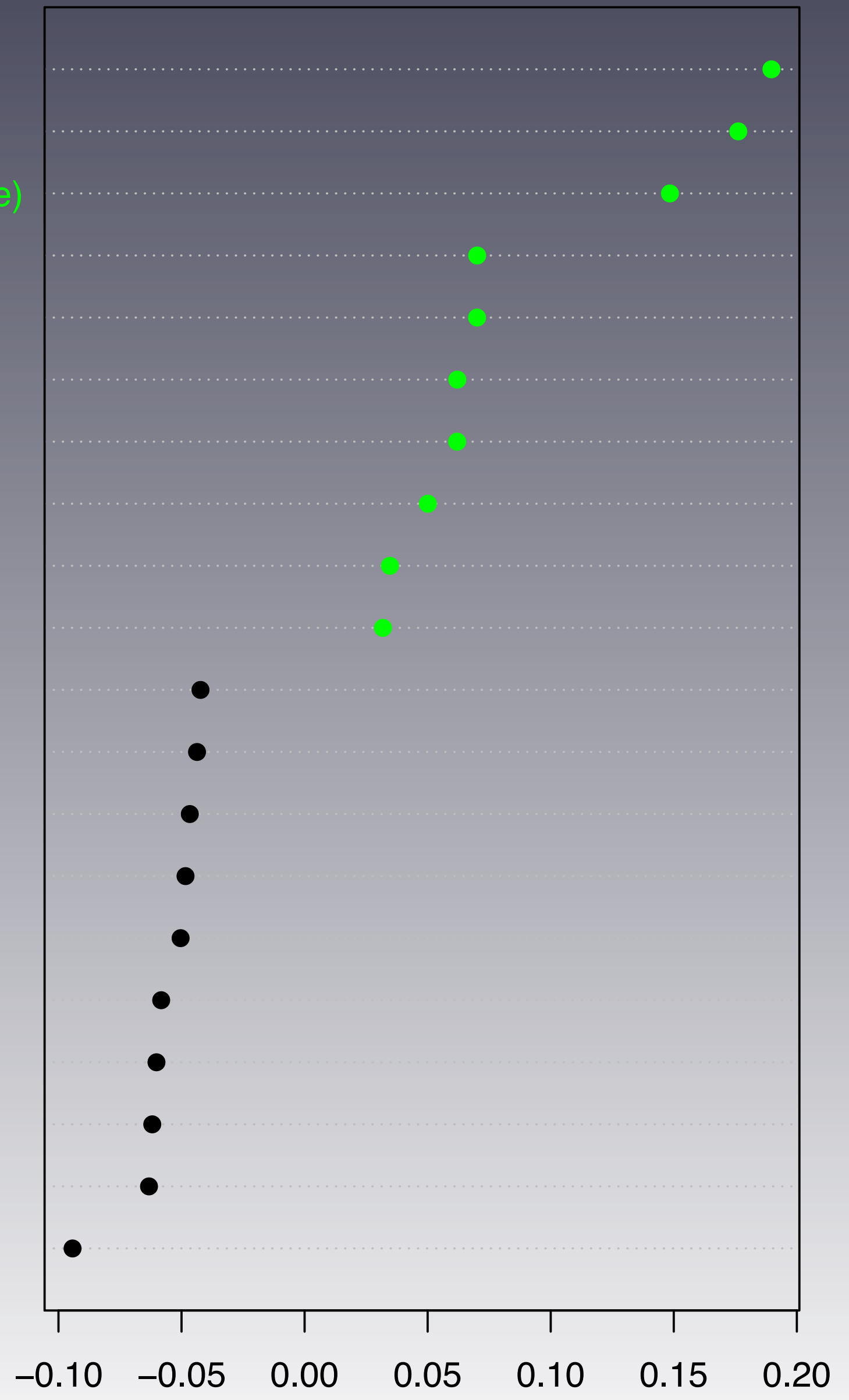
- Workers
- Inequality
- #Failure
- International Crises
- Discrimination
- Housing
- #Standards
- Banks
- Jobs
- Private/Public
- Bureaucracy
- Agreements
- Sport
- Agencies/Bodies
- Agriculture
- #Rules
- #Timetable
- #Initiatives
- #Praise
- Business



Less attention ~ More attention

Ecological parties vs. others

- Energy
- Environment
- Global (Aid/Climate)
- #Failure
- Agriculture
- Science/R&D
- #Skepticism
- #Studies
- #Nonsense
- Representation
- #Procurement
- #Standards
- #Issues
- Taxes
- Professions
- Workers
- Agencies/Bodies
- Business
- Health
- Local/Regional

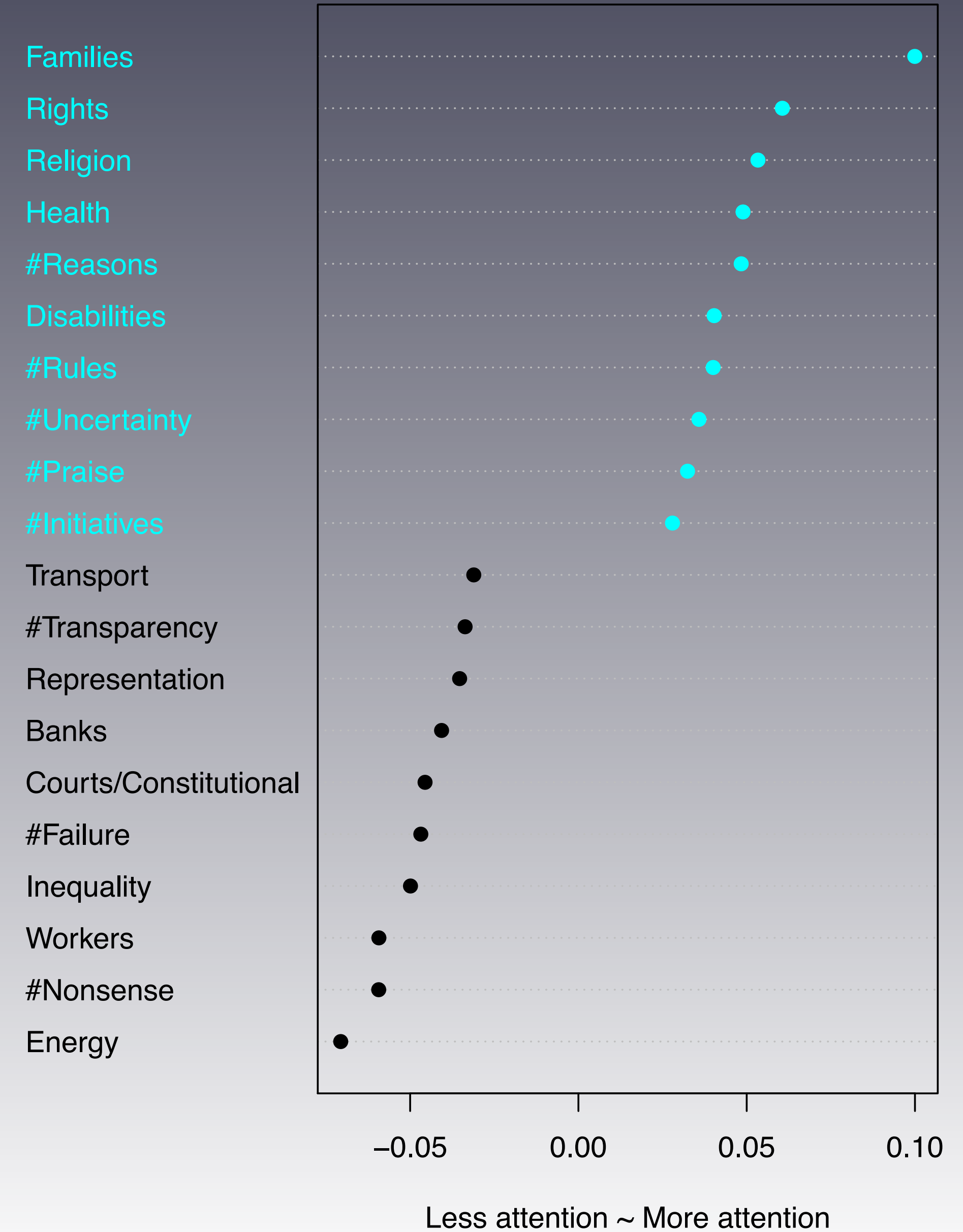


Less attention ~ More attention

### Liberal parties vs. others



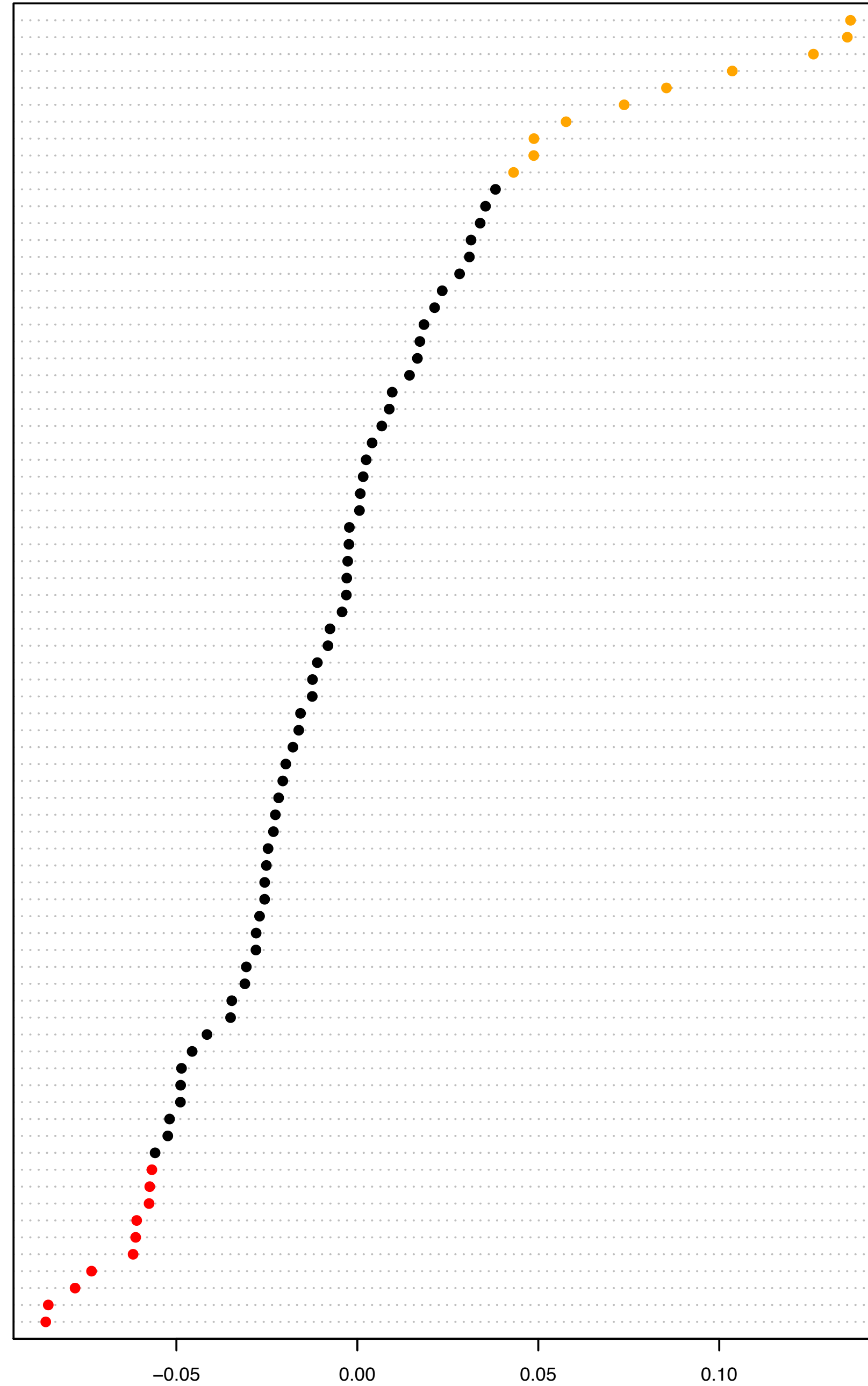
### Christian Democratic parties vs. others



# Nationalist

Bonus slide!

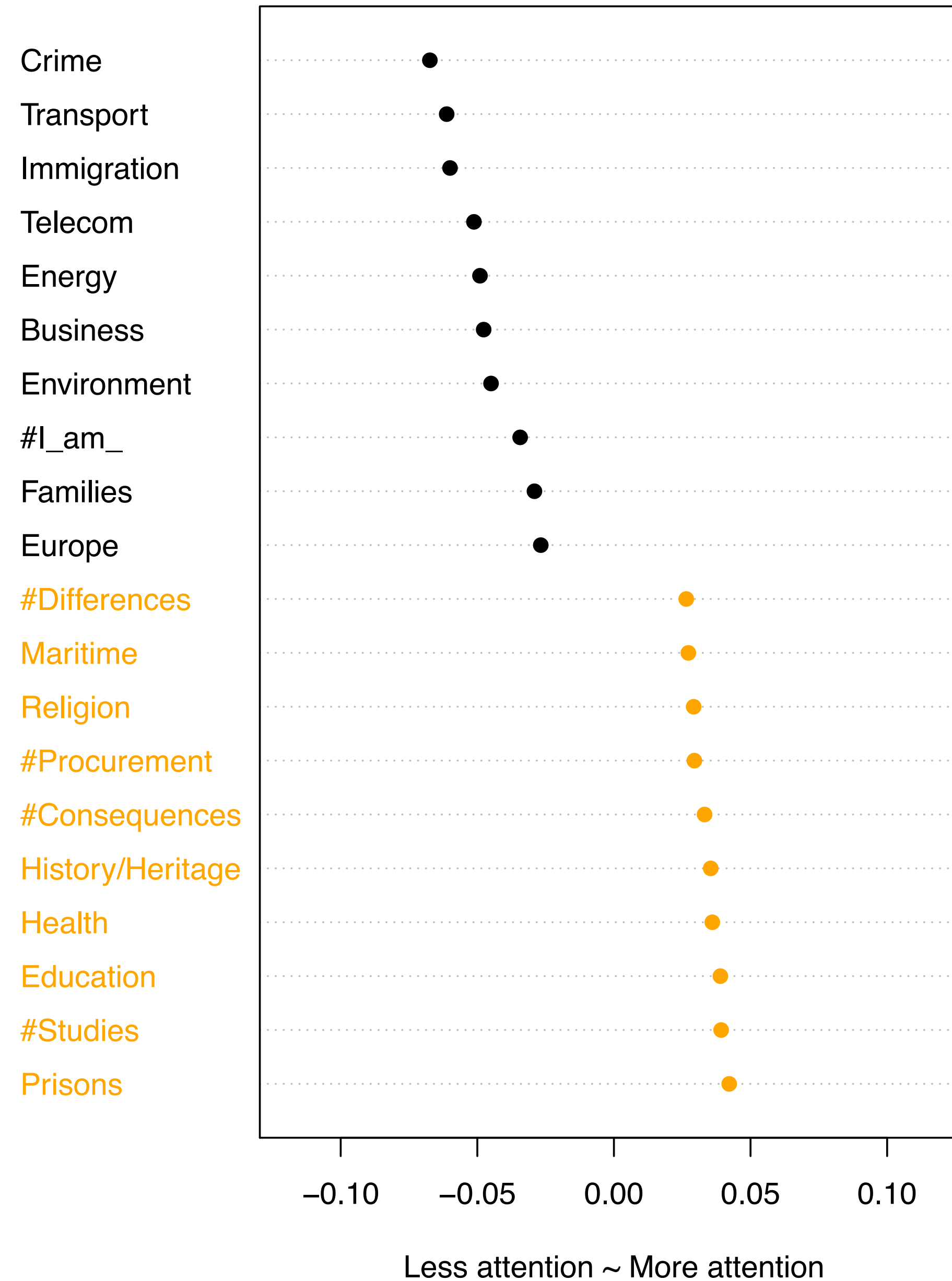
- Representation
- Immigration
- #Failure
- #Nonsense
- Crime
- Religion
- OECD/Trade
- Budget
- #Quotes
- #Disaster
- #Skepticism
- International Crises
- Maritime
- Professions
- Public Health
- Prisons
- Inequality
- Europe
- #I\_am\_
- #My
- Local/Regional
- Housing
- Defense
- #Standards
- Bureaucracy
- 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





Conservative

Changes in Conservative party attention  
*when there are nationalist parties present*



We can complete cross-national party analogies

(The embeddings capture linear ideological semantics.)

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

					Cosine
	$\mathbf{V}_{\text{Lab}} - \mathbf{V}_{\text{Con}} + \mathbf{V}_{\text{CDU/CSU}}$	is closest to	$\mathbf{V}_{\text{SPD}}$		0.60
	$\mathbf{V}_{\text{Lab}} - \mathbf{V}_{\text{Con}} + \mathbf{V}_{\text{Kok}}$	is closest to	$\mathbf{V}_{\text{SDP}}$		0.59
	$\mathbf{V}_{\text{Lab}} - \mathbf{V}_{\text{Con}} + \mathbf{V}_{\text{CDA}}$	is closest to	$\mathbf{V}_{\text{PvdA}}$		0.57
	$\mathbf{V}_{\text{Lab}} - \mathbf{V}_{\text{Con}} + \mathbf{V}_{\text{M}}$	is closest to	$\mathbf{V}_{\text{SAP}}$		0.49
	$\mathbf{V}_{\text{Lab}} - \mathbf{V}_{\text{Con}} + \mathbf{V}_{\text{PP}}$	is closest to	$\mathbf{V}_{\text{PSOE}}$		0.38
	$\mathbf{V}_{\text{Lab}} - \mathbf{V}_{\text{Con}} + \mathbf{V}_{\text{FI/PdL}}$	is closest to	$\mathbf{V}_{\text{PD}}$		0.27
	$\mathbf{V}_{\text{Lab}} - \mathbf{V}_{\text{Con}} + \mathbf{V}_{\text{ODS}}$	is closest to	$\mathbf{V}_{\text{CSSD}}$		0.23

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

					Cosine
	$\mathbf{V}_{Vas} - \mathbf{V}_{SDP} + \mathbf{V}_{PSOE}$	is closest to	$\mathbf{V}_{IU}$		0.63
	$\mathbf{V}_{Vas} - \mathbf{V}_{SDP} + \mathbf{V}_{SAP}$	is closest to	$\mathbf{V}_V$		0.49
	$\mathbf{V}_{Vas} - \mathbf{V}_{SDP} + \mathbf{V}_{CSSD}$	is closest to	$\mathbf{V}_{KSCM}$		0.48
	$\mathbf{V}_{Vas} - \mathbf{V}_{SDP} + \mathbf{V}_{PvdA}$	is closest to	$^*\mathbf{V}_{GL}$		0.46
	$\mathbf{V}_{Vas} - \mathbf{V}_{SDP} + \mathbf{V}_{PD}$	is closest to	$\mathbf{V}_{SEL}$		0.30
	$\mathbf{V}_{Vas} - \mathbf{V}_{SDP} + \mathbf{V}_{SPD}$	is closest to	$\mathbf{V}_{Linke}$		0.26
	$\mathbf{V}_{Vas} - \mathbf{V}_{SDP} + \mathbf{V}_{Lab}$	is closest to	$^{**}\mathbf{V}_{Green}$		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	$\mathbf{V}_{\text{FdV}}$		0.68
	$\mathbf{V}_{\text{Gruene}} - \mathbf{V}_{\text{SPD}} + \mathbf{V}_{\text{SAP}}$	is closest to	$\mathbf{V}_{\text{MP}}$		0.50
	$\mathbf{V}_{\text{Gruene}} - \mathbf{V}_{\text{SPD}} + \mathbf{V}_{\text{Lab}}$	is closest to	$\mathbf{V}_{\text{Green}}$		0.46
	$\mathbf{V}_{\text{Gruene}} - \mathbf{V}_{\text{SPD}} + \mathbf{V}_{\text{SDP}}$	is closest to	$\mathbf{V}_{\text{KD}}$		0.24
	$\mathbf{V}_{\text{Gruene}} - \mathbf{V}_{\text{SPD}} + \mathbf{V}_{\text{PvdA}}$	is closest to	$\mathbf{V}_{\text{MP}}$		0.13
	$\mathbf{V}_{\text{Gruene}} - \mathbf{V}_{\text{SPD}} + \mathbf{V}_{\text{PSOE}}$	is closest to	** $\mathbf{V}_{\text{ERC}}$		0.36
	$\mathbf{V}_{\text{Gruene}} - \mathbf{V}_{\text{SPD}} + \mathbf{V}_{\text{CSSD}}$	is closest to	** $\mathbf{V}_{\text{Usvit}}$		0.34

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

					Cosine
	$V_{LPPF} - V_{CDA} + V_{FI/PdL}$	is closest to	$V_{FDI}$		0.67
	$V_{LPPF} - V_{CDA} + V_M$	is closest to	$V_{NyD}$		0.31
	$V_{LPPF} - V_{CDA} + V_{Con}$	is closest to	$*V_{DUP}$		0.28
	$V_{LPPF} - V_{CDA} + V_{Kok}$	is closest to	$*V_{RKP}$		0.24
	$V_{LPPF} - V_{CDA} + V_{ODS}$	is closest to	$V_{Usvit}$		0.20
	$V_{LPPF} - V_{CDA} + V_{CDU/CSU}$	is closest to	$**V_{FDP}$		0.27
	$V_{LPPF} - V_{CDA} + V_{PP}$	is closest to	$**V_{PSOE}$		0.24

Closest Party Family Mean

	ECO	LEF	SOC	LIB	CHR	CON	NAT	ETH
ECO	sv_MP, it_FdV, fi_Vihr, uk_Green, nl_PvdD, de_Gruene	nl_GL						
LEF		sv_V, es_IU, it_PRC, it_PdCI, de_Linke, it_SEL, fi_Vas, nl_SP, cs_KSCM, it_RnP						
SOC			uk_Lab, fi_SDP, 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_TOP09, 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_SDLP, 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.

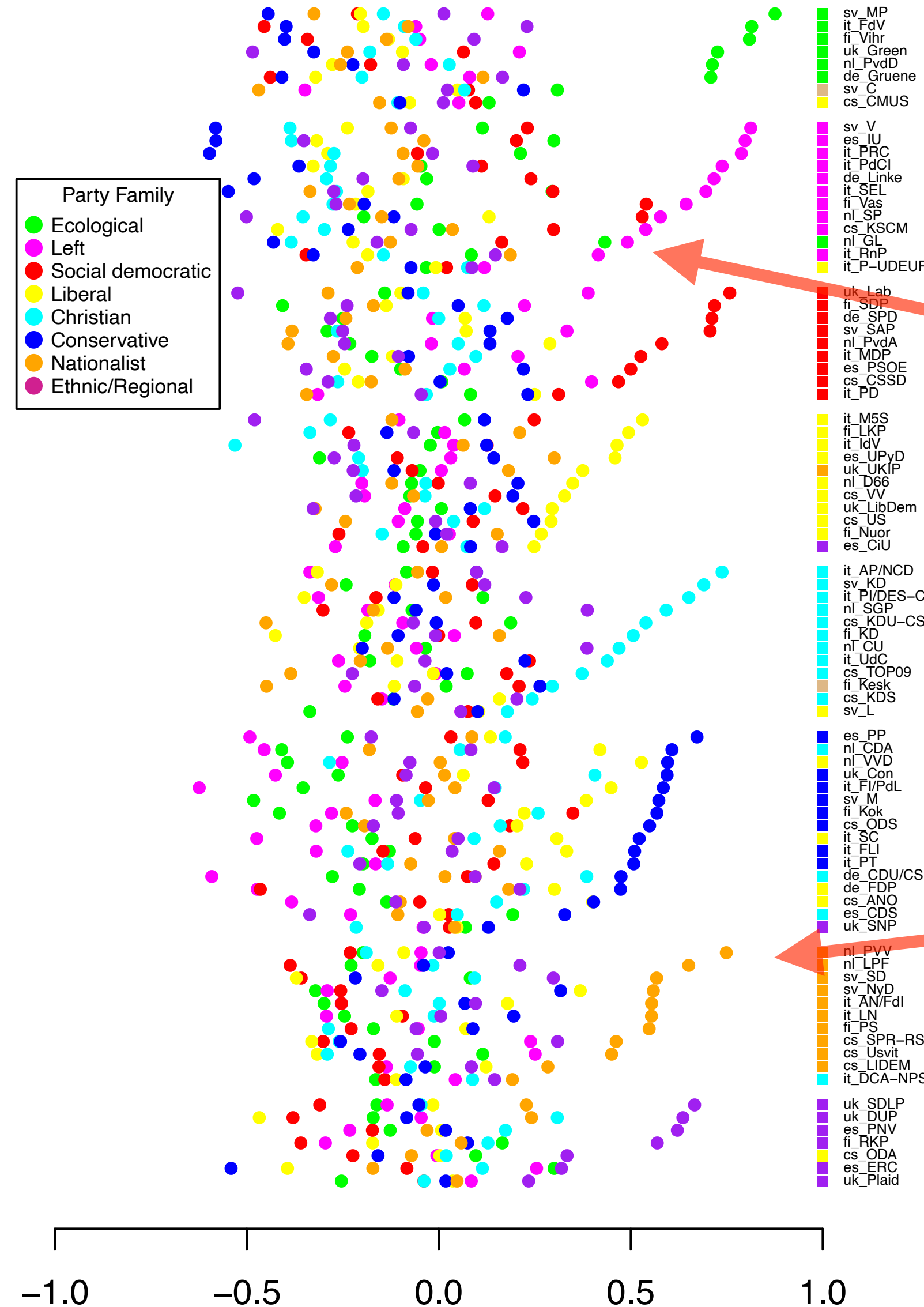
## Cosine Similarities with Party Family Means

# 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.



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.

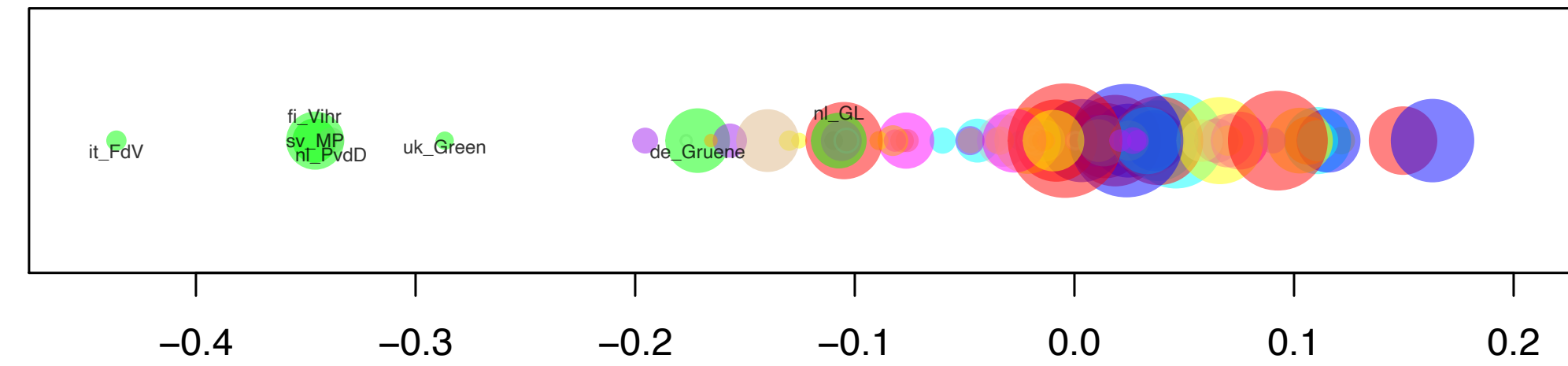
Our embeddings contain a topic model *and a spatial scaling model*.



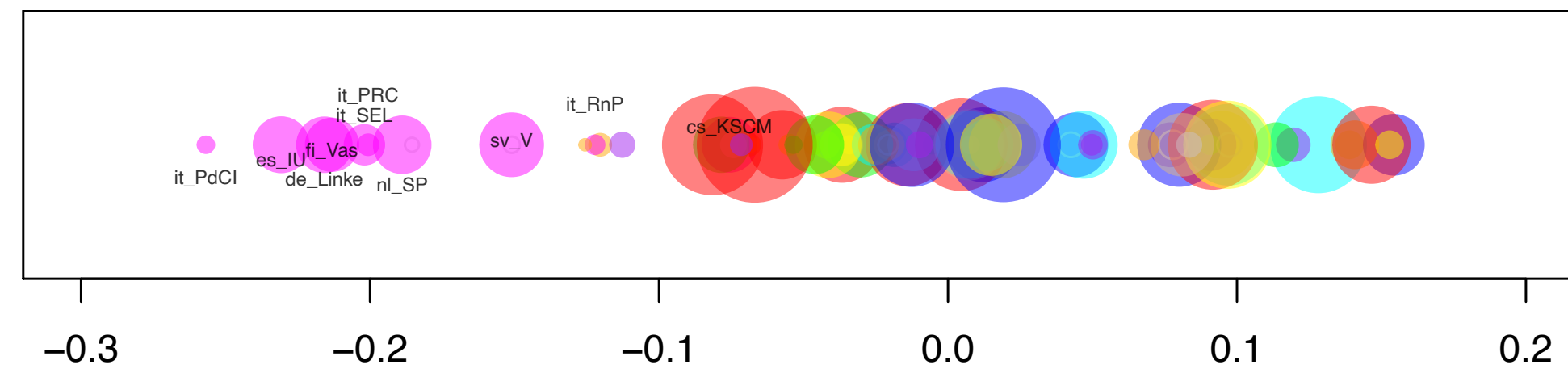
# Varimax rotation isolates the most distinctive ideological party families in the first five dimensions

Projections onto rotated dimensions

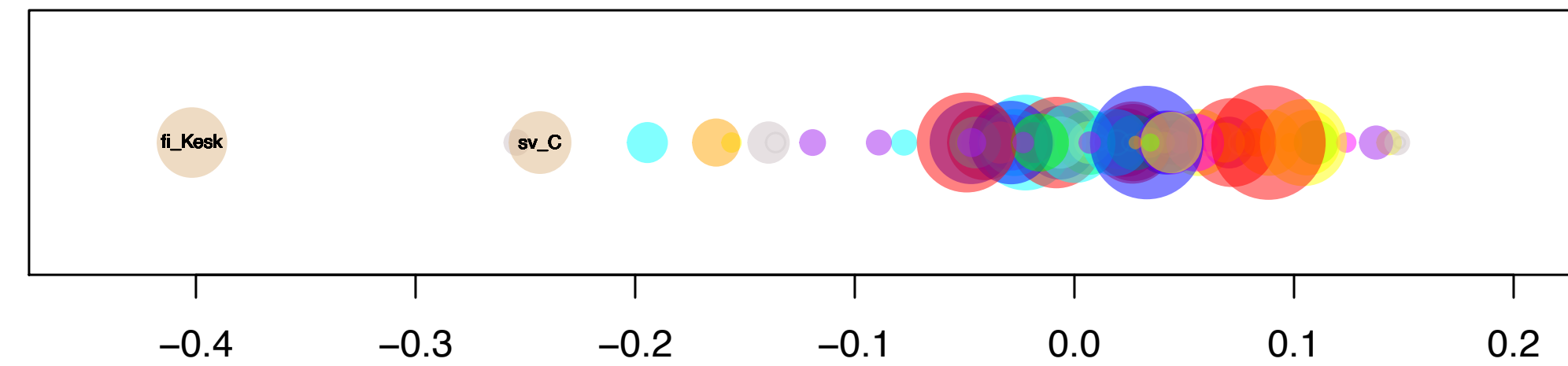
Green dimension



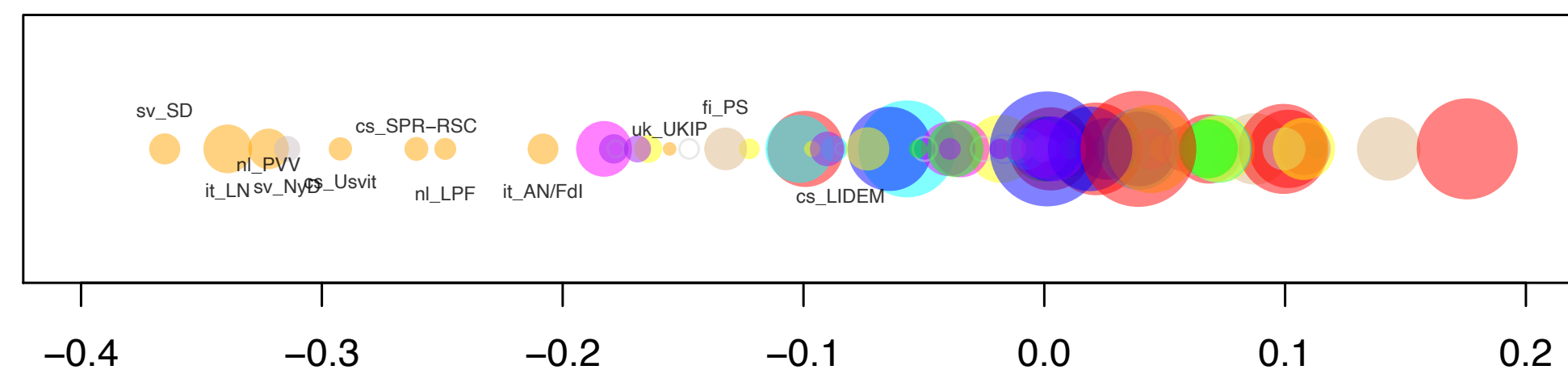
Left dimension



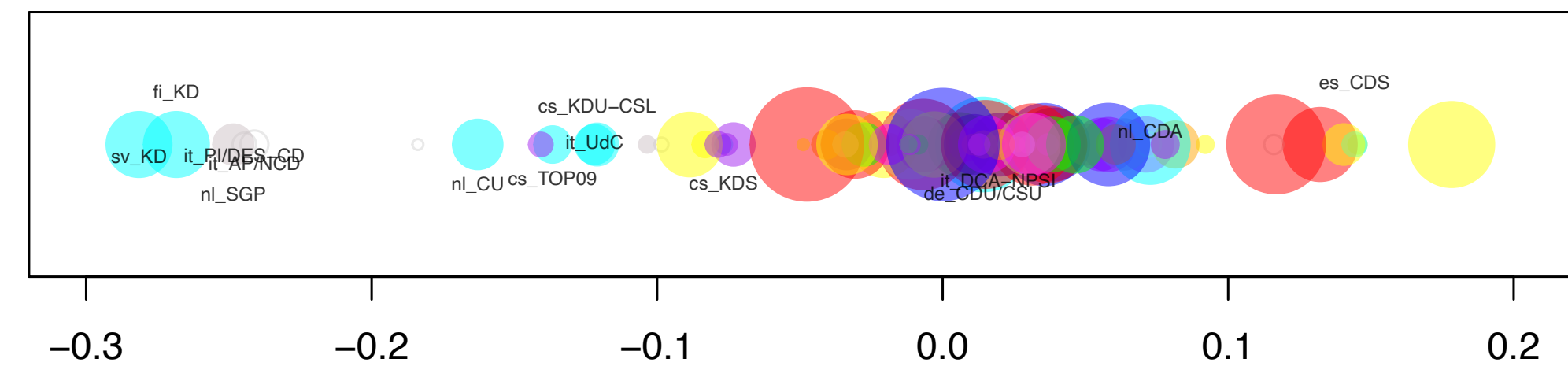
Agrarian dimension



Nationalist dimension



Christian dimension



# 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](mailto:burtmonroe@psu.edu)

# Machine Translation

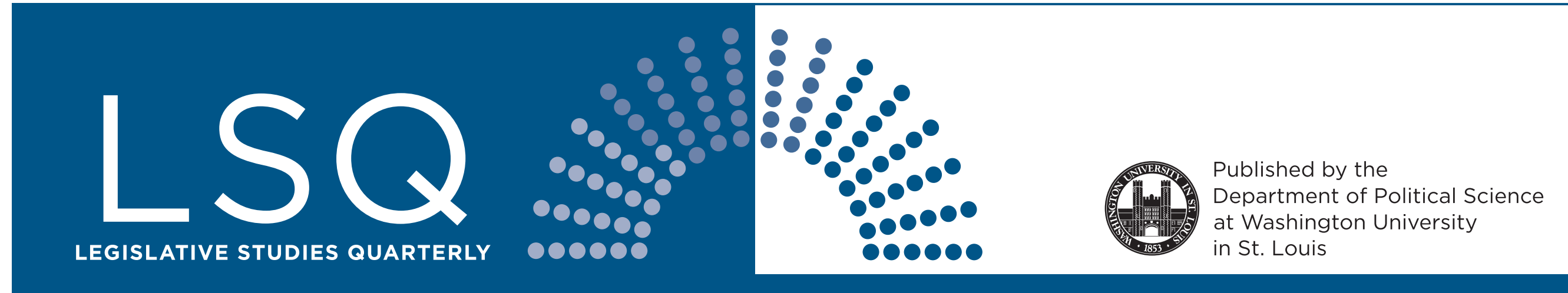
Single language topic models

## Top keywords for example topics, Knesset

Negotiations	Jobs	Religion / Identity	Education	Housing/ Immigration
אוסלו	העבודה	בשבת	החינוך	הקליטה
ערפאת	והרווחה	המשטרה	הספר	דירות
הפלשתינים	עובדים	הדתות	והתרבות	דיוור
ירושלים	התעסוקה	השואה	המורים	העלייה
השלום	שכר	היהדות	התלמידים	העולים
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



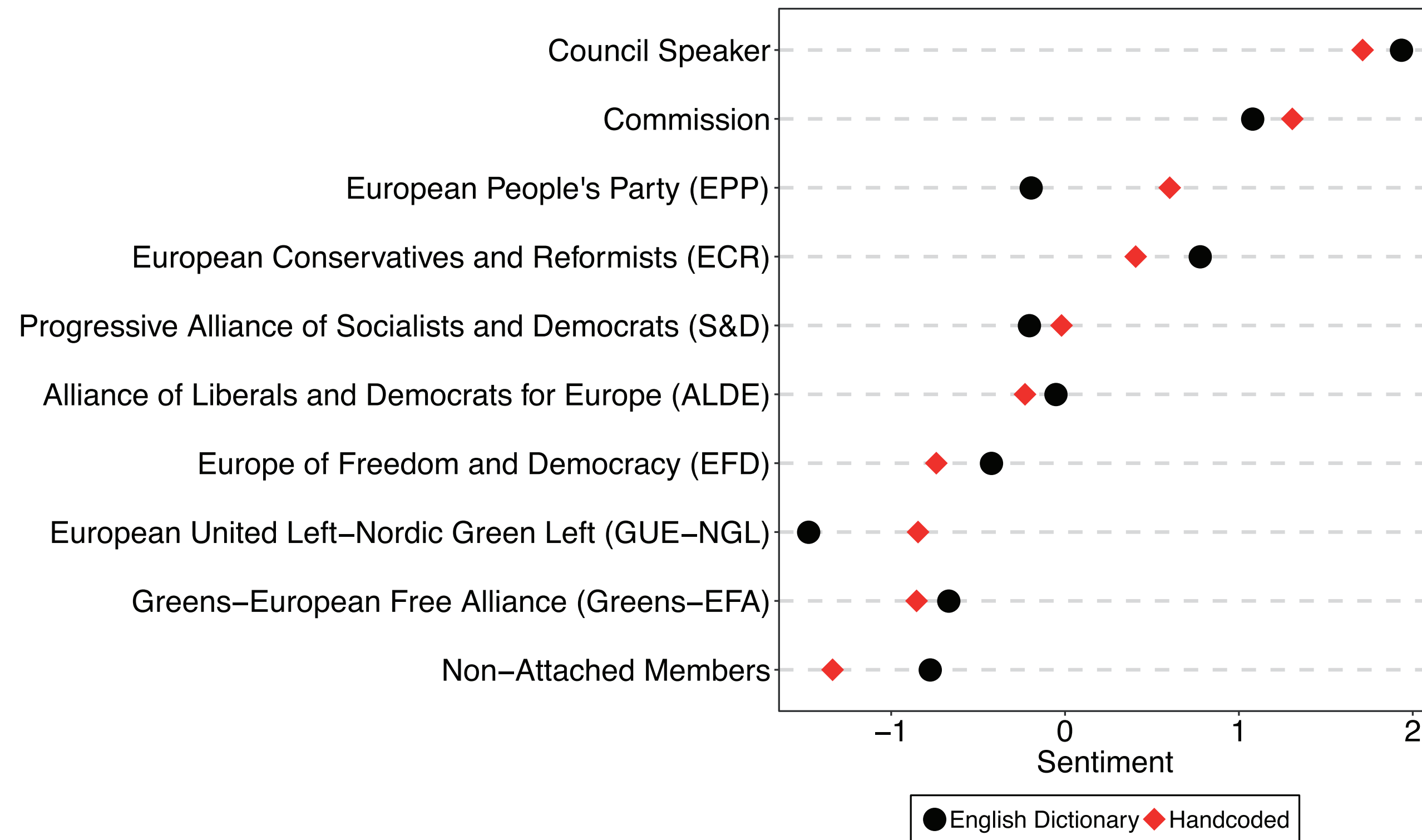
Translating sentiment dictionaries



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

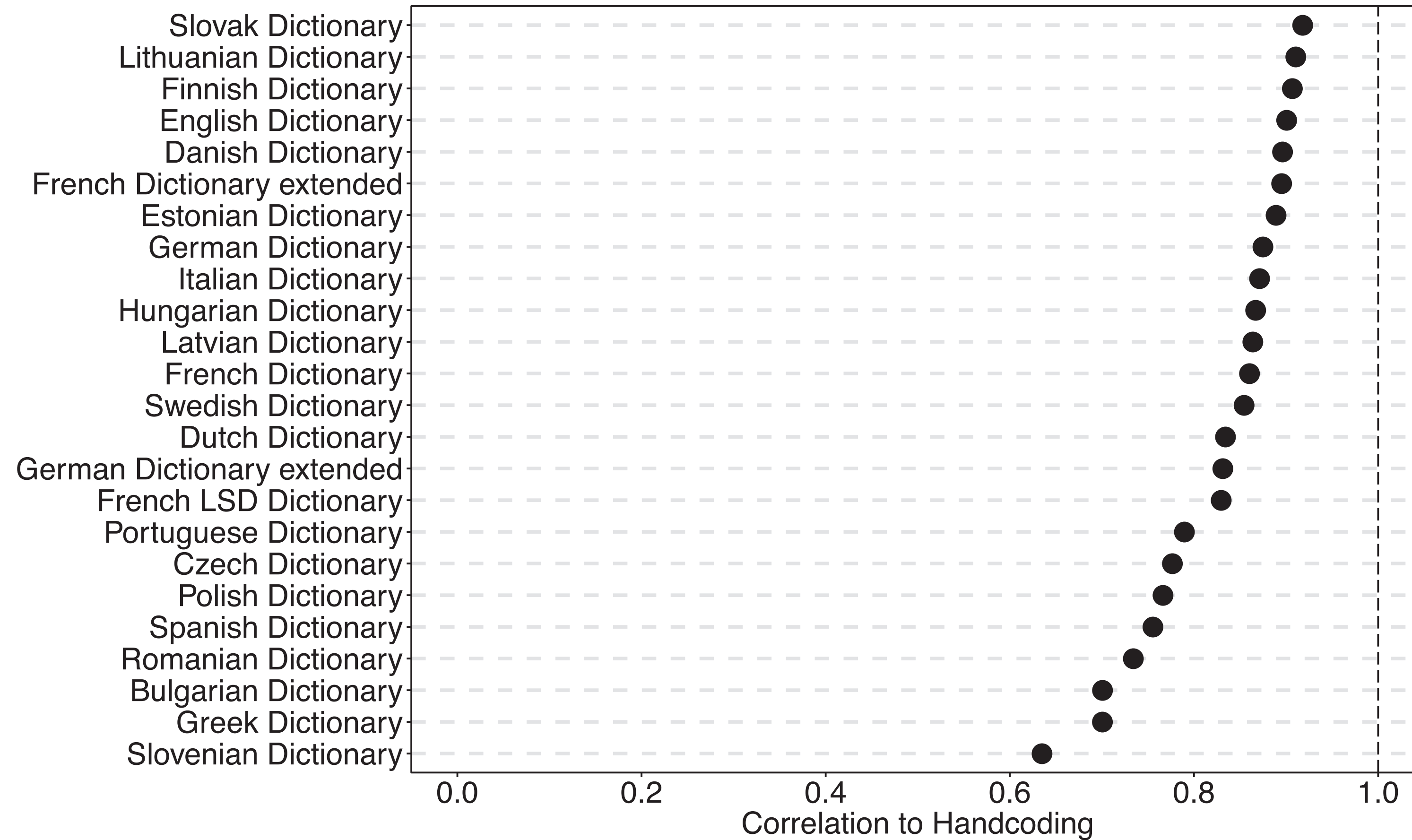
*Multilingual Sentiment Analysis: A  
New Approach to Measuring Conflict  
in Legislative Speeches*

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

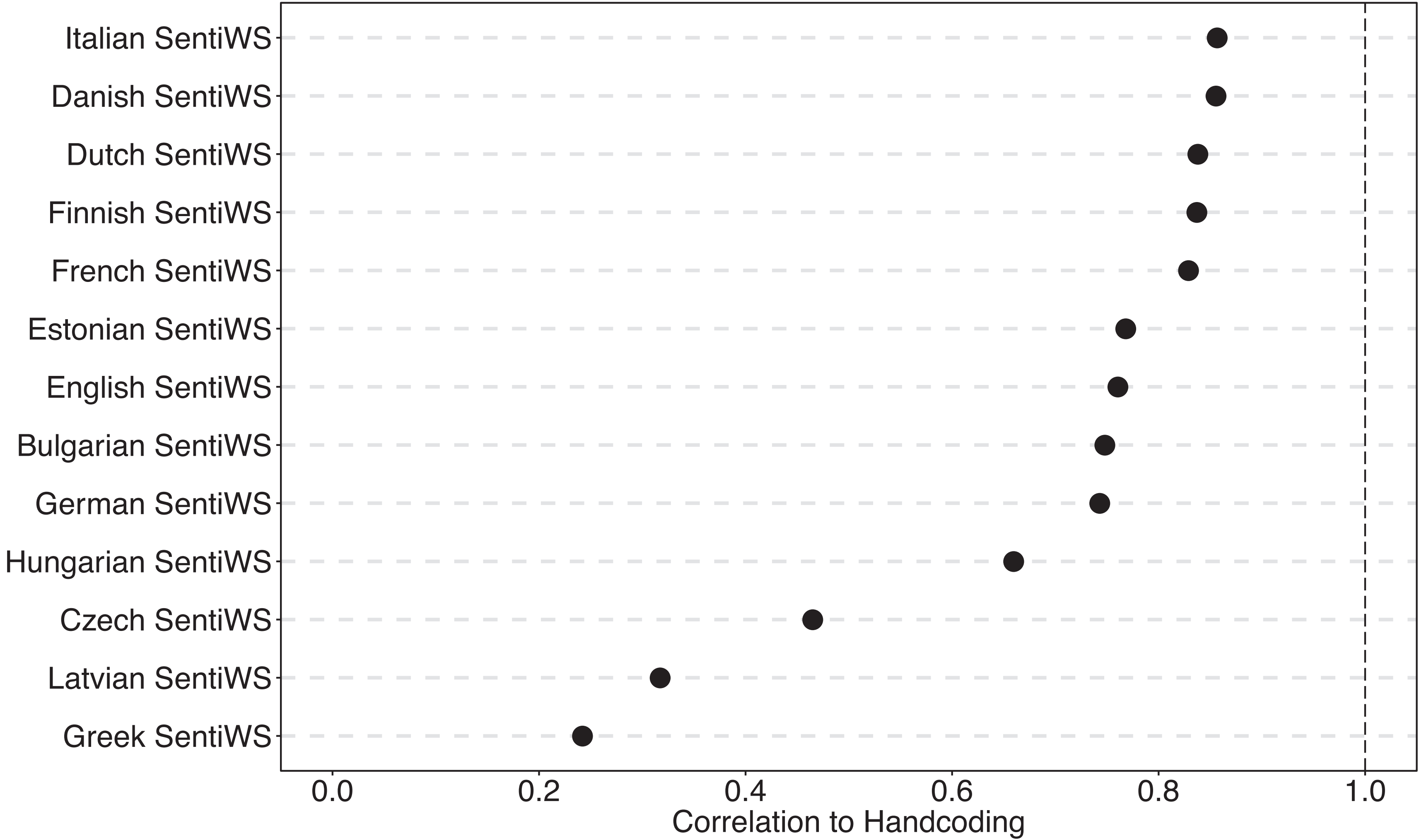


FIGURE 4  
 Sentiment in Irish Budget Debates, 2008–12 [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

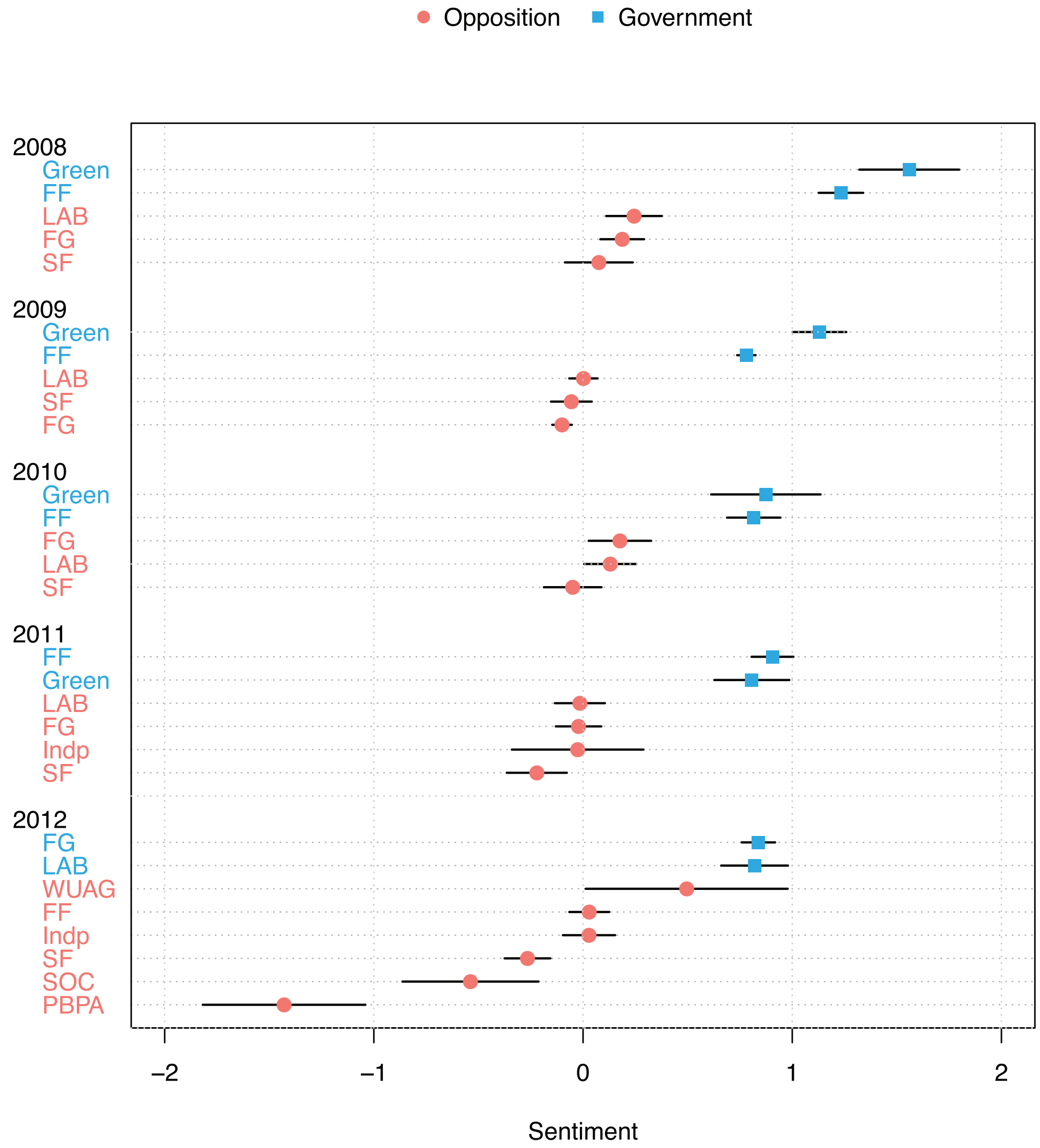
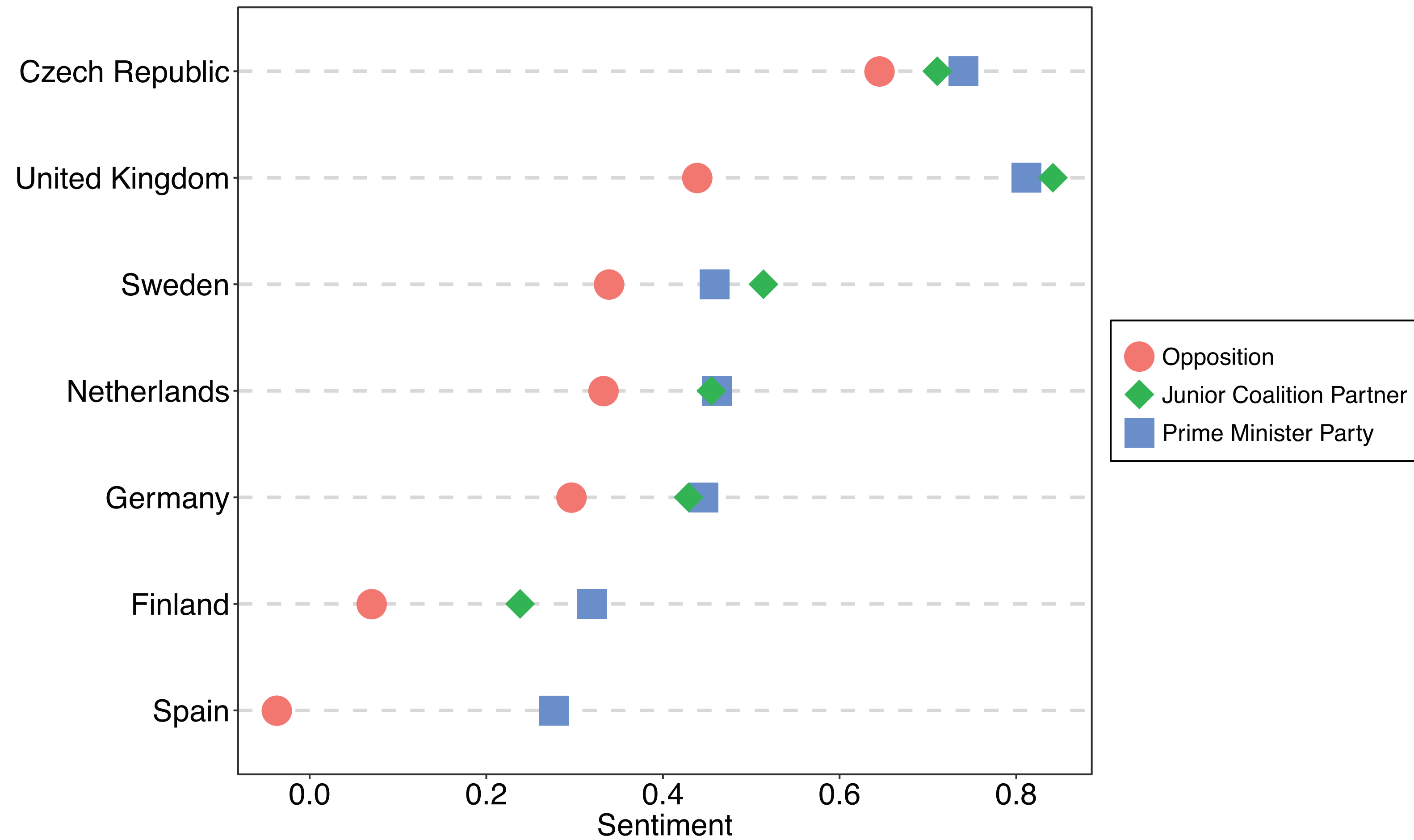


FIGURE 8  
Prime Minister Party, Junior Coalition Partner, and Opposition  
Sentiment in European Parliaments [Colour figure can be  
viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]



Bag-of-words generally?



 OPEN ACCESS  PEER-REVIEWED

RESEARCH ARTICLE

## Automated content analysis across six languages

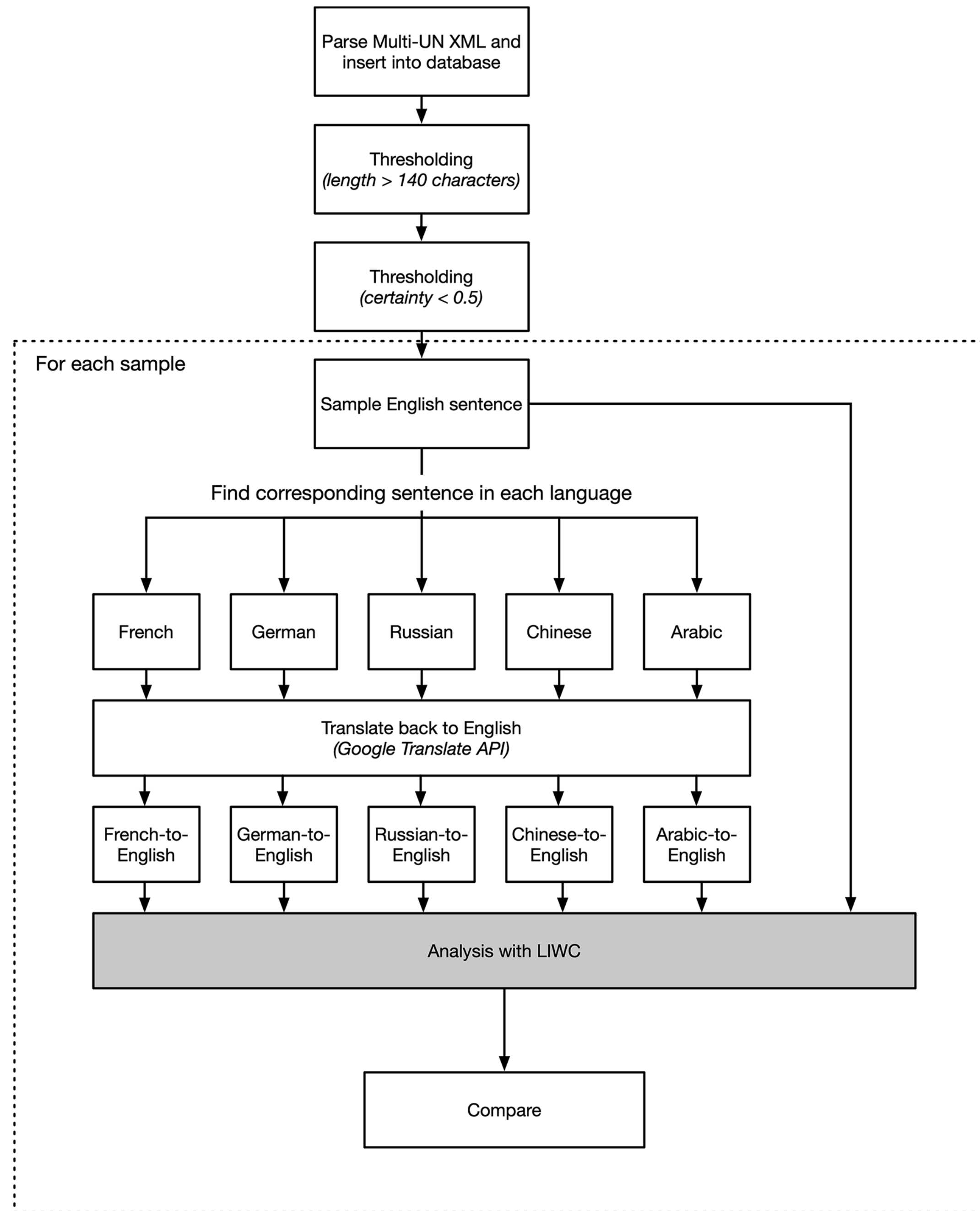
Leah Cathryn Windsor , James Grayson Cupit, Alistair James Windsor

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
<b>Russian</b>	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
<b>French</b>	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 humans are the subject of a thorough investigation
<b>Arabic</b>	ضرب طروتب ديفت يتا ريراقتا ازا هقلا نء سلجمدا برعيو تاكاهتذاو قريطخ مئارج يف ةيتياها ةينطولا ةطرشدا طابض ي ا يف ةلماك ةروصب قيقحتا ةرورض دكؤيو ، ناسذلا قوقحد . ناسذلا قوقحد تاكاهتذاو قريطخ مئارج باكترا باءاءا	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
<b>Chinese</b>	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
<b>German</b>	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



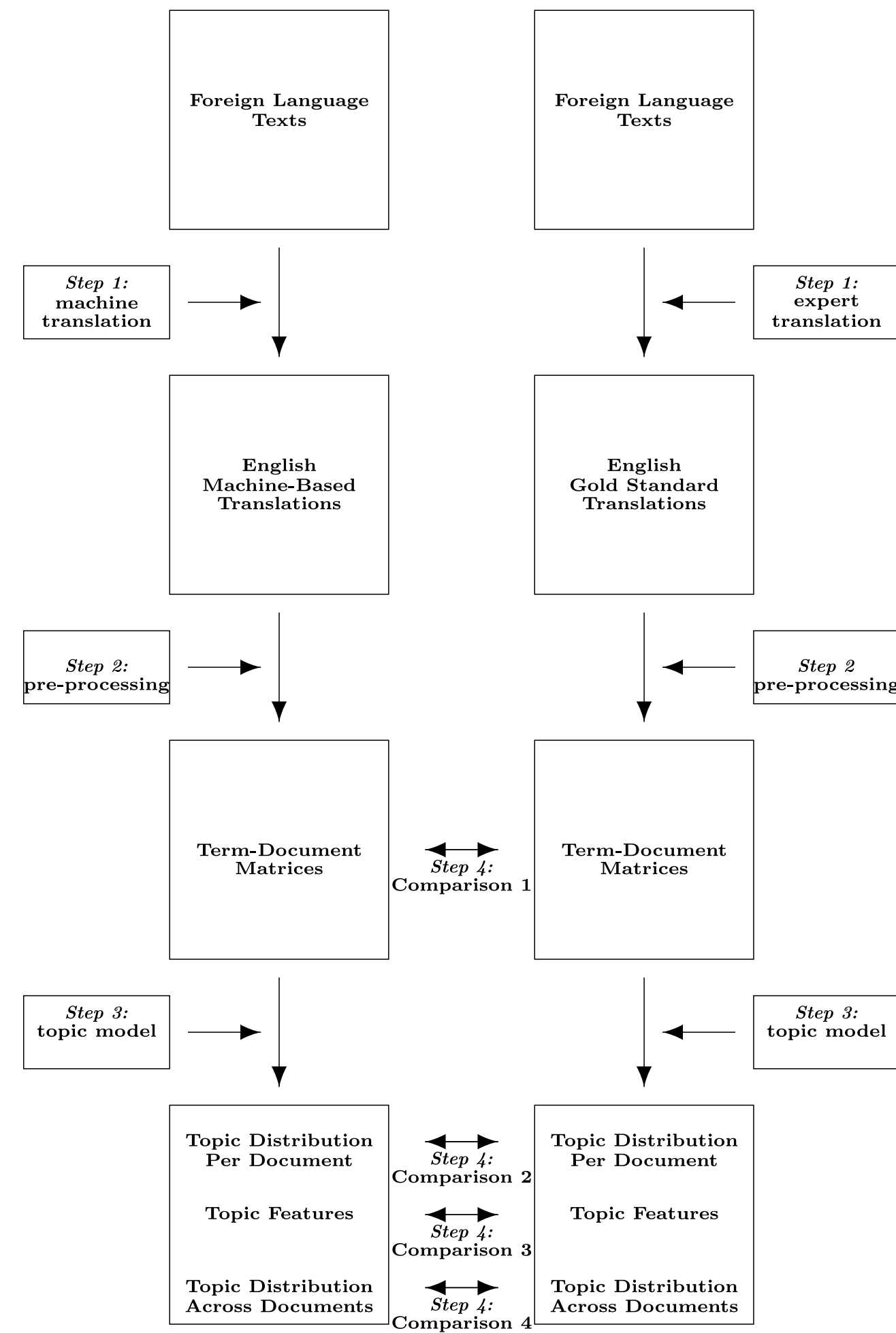
LIWC Category	Language translated from					Mean
	Arabic	German	French	Russian	Mandarin	
All	0.831	0.814	0.822	0.843	0.783	0.820
Summary	0.863	0.833	0.856	0.906	0.761	0.844
Linguistic Dim.	0.729	0.728	0.769	0.788	0.651	0.733
Other Grammar	0.829	0.784	0.783	0.856	0.724	0.795
Psych. Proc.	0.862	0.838	0.832	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 Vries<sup>1</sup>, Martijn Schoonvelde<sup>2</sup> and Gijs Schumacher<sup>3</sup>**



**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).

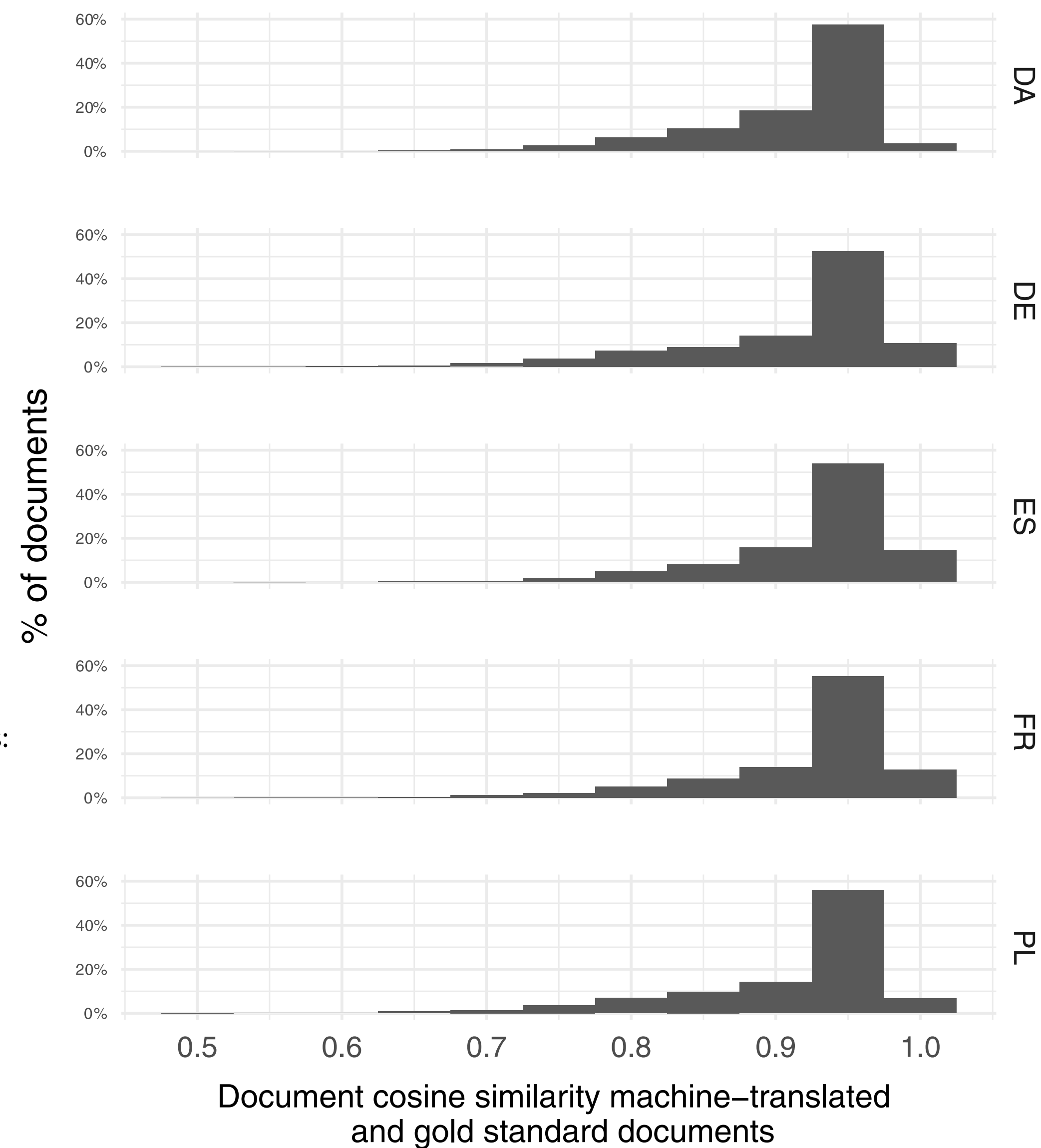
**Table 1.** Comparisons between gold standard and machine-translated data.

	<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.

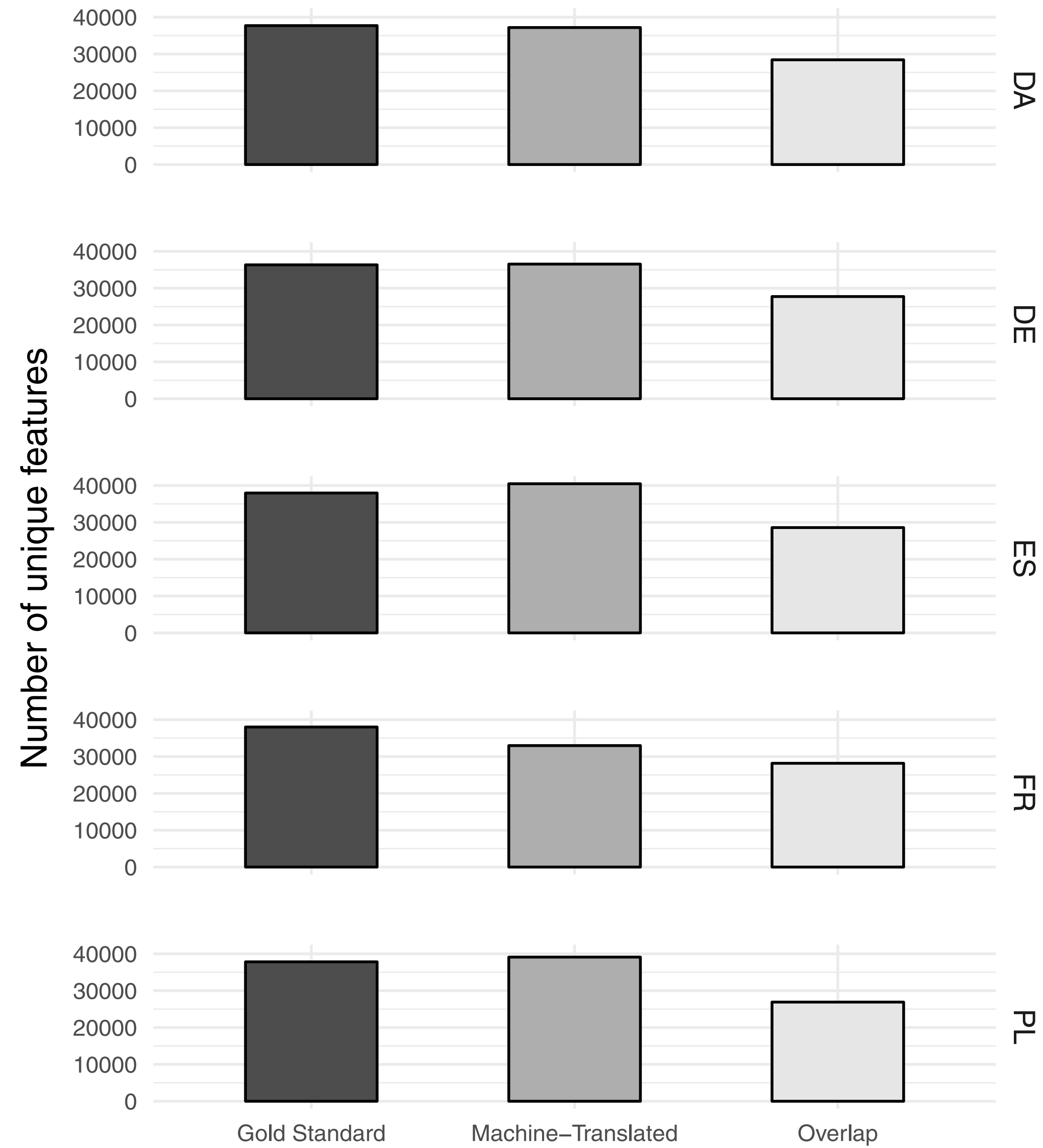
Language	N	Mean	St. Dev.	Min	Max
Danish	2,301	0.915	0.063	0.549	0.992
German	2,148	0.915	0.074	0.488	0.991
Spanish	2,335	0.929	0.059	0.483	0.991
French	2,347	0.925	0.064	0.564	0.989
Polish	2,338	0.913	0.073	0.475	0.989
Total:	11,469	0.919	0.066	0.475	0.992

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.



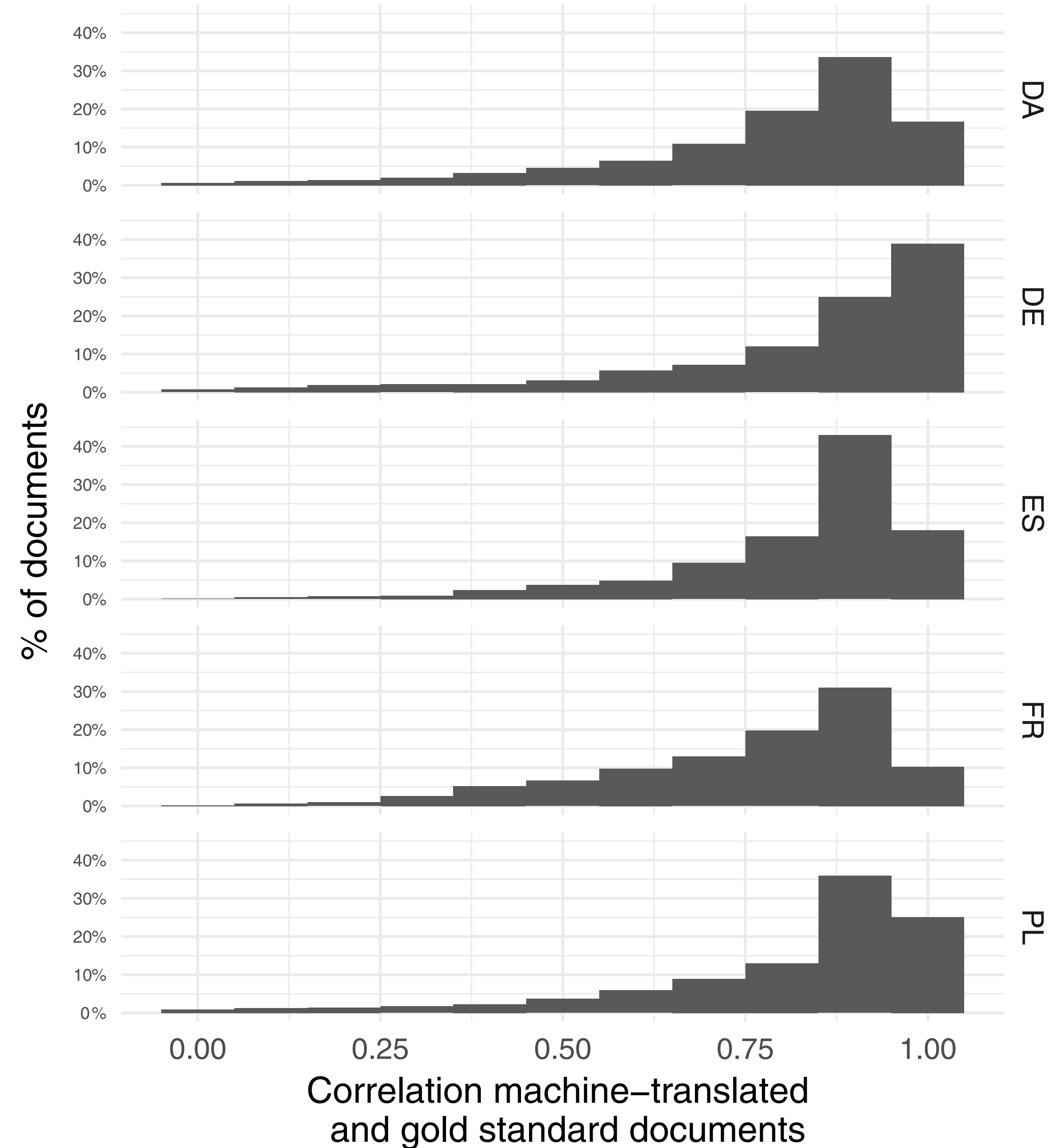


**Figure 4.** Unique TDM features for gold standard and machine-translated corpora. *Reading example:* for the French language, the amount of overlapping features is around 28,000, while the total number of features is around 33,000 for the machine-translated documents and around 38,000 for the gold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

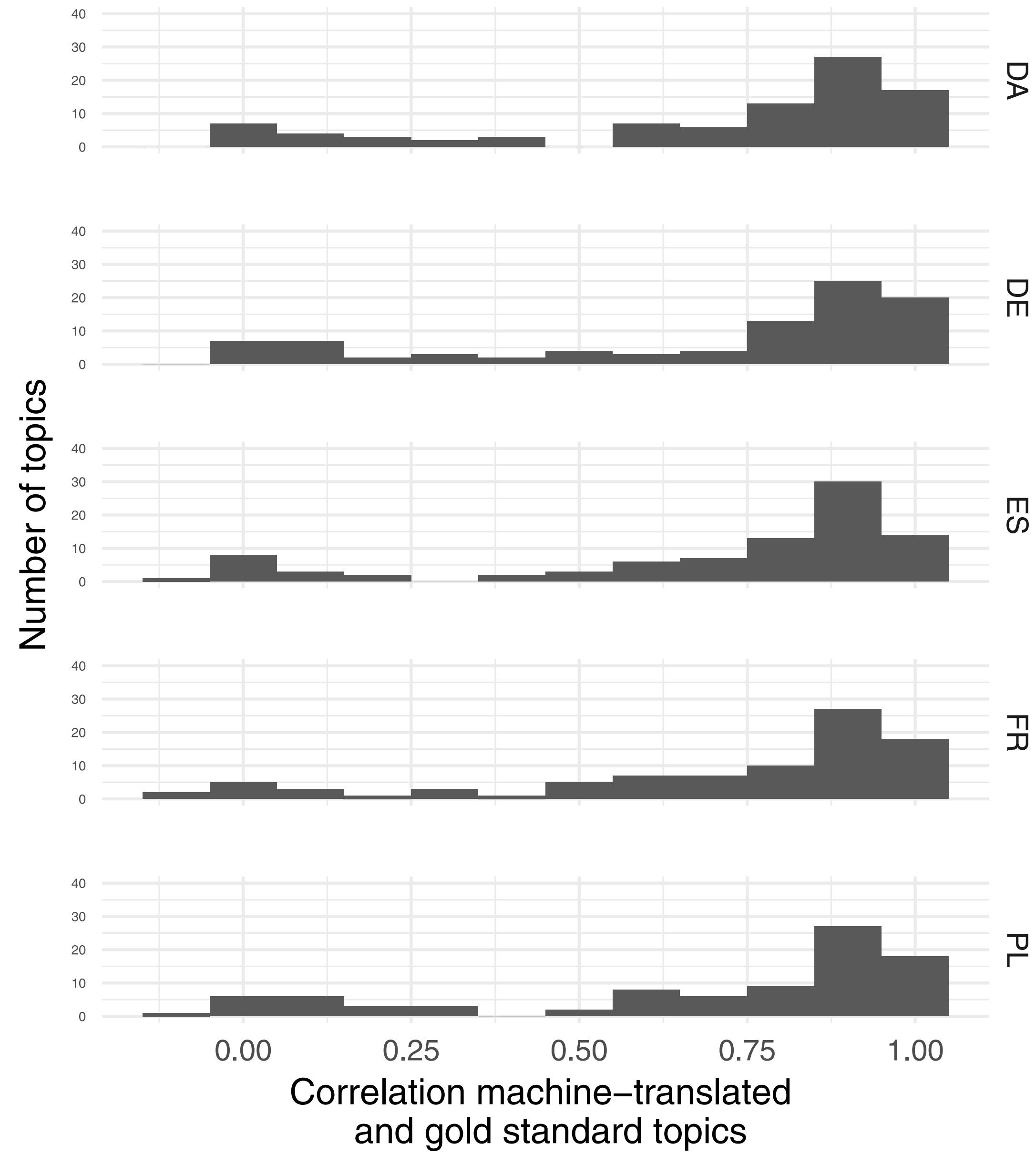
**Table 3.** Similarity of document-level topical prevalence with equal number of topics.

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

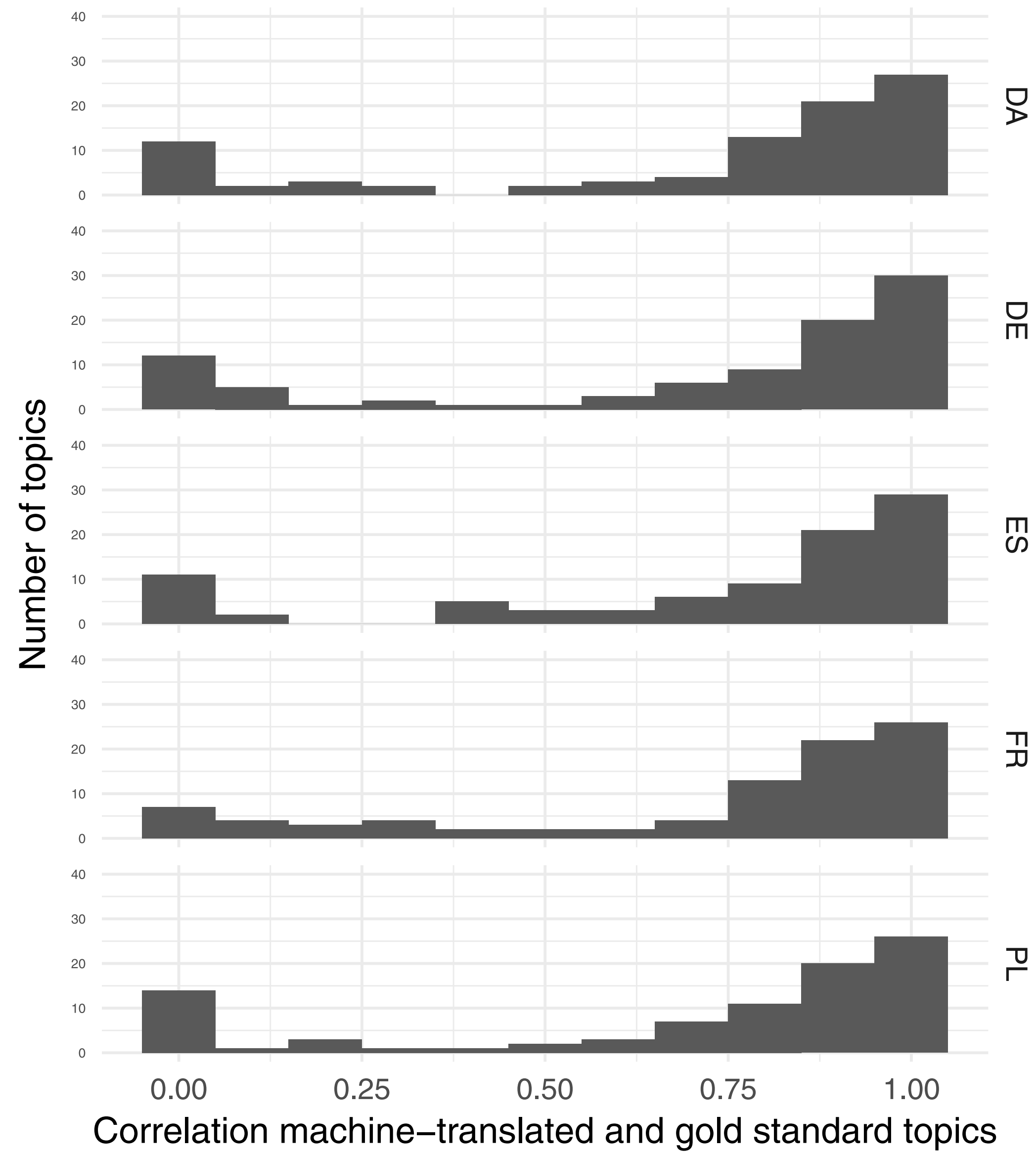
Note: ANOVA results:  $F(4, 11464) = 56.414, \rho < 0.001, \eta^2 = 0.019$ .



**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