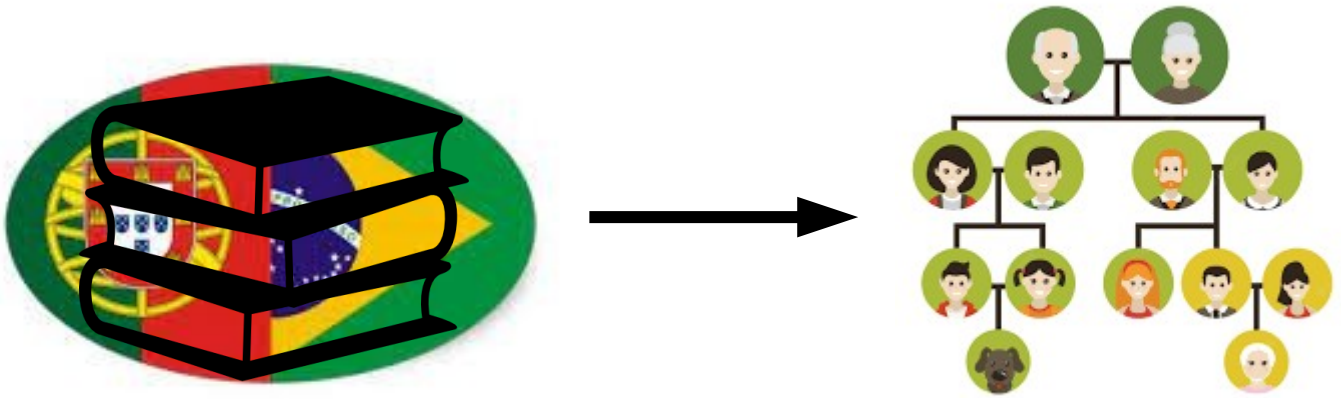


Automatic Cast Extraction in Portuguese and Brazilian Literature

Eckhard Bick

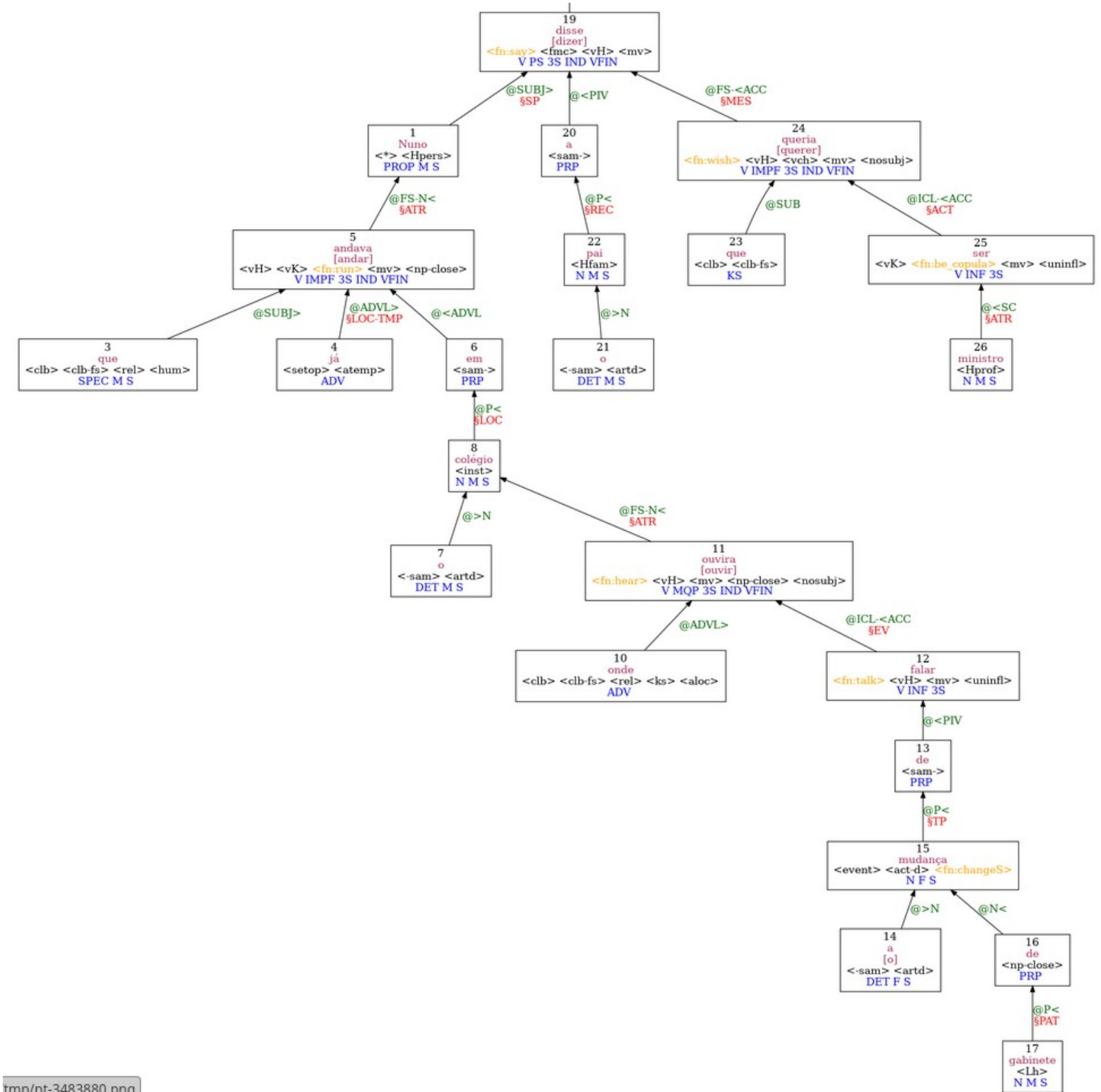
University of Southern Denmark
Institute of Language and Communication

The DIP task

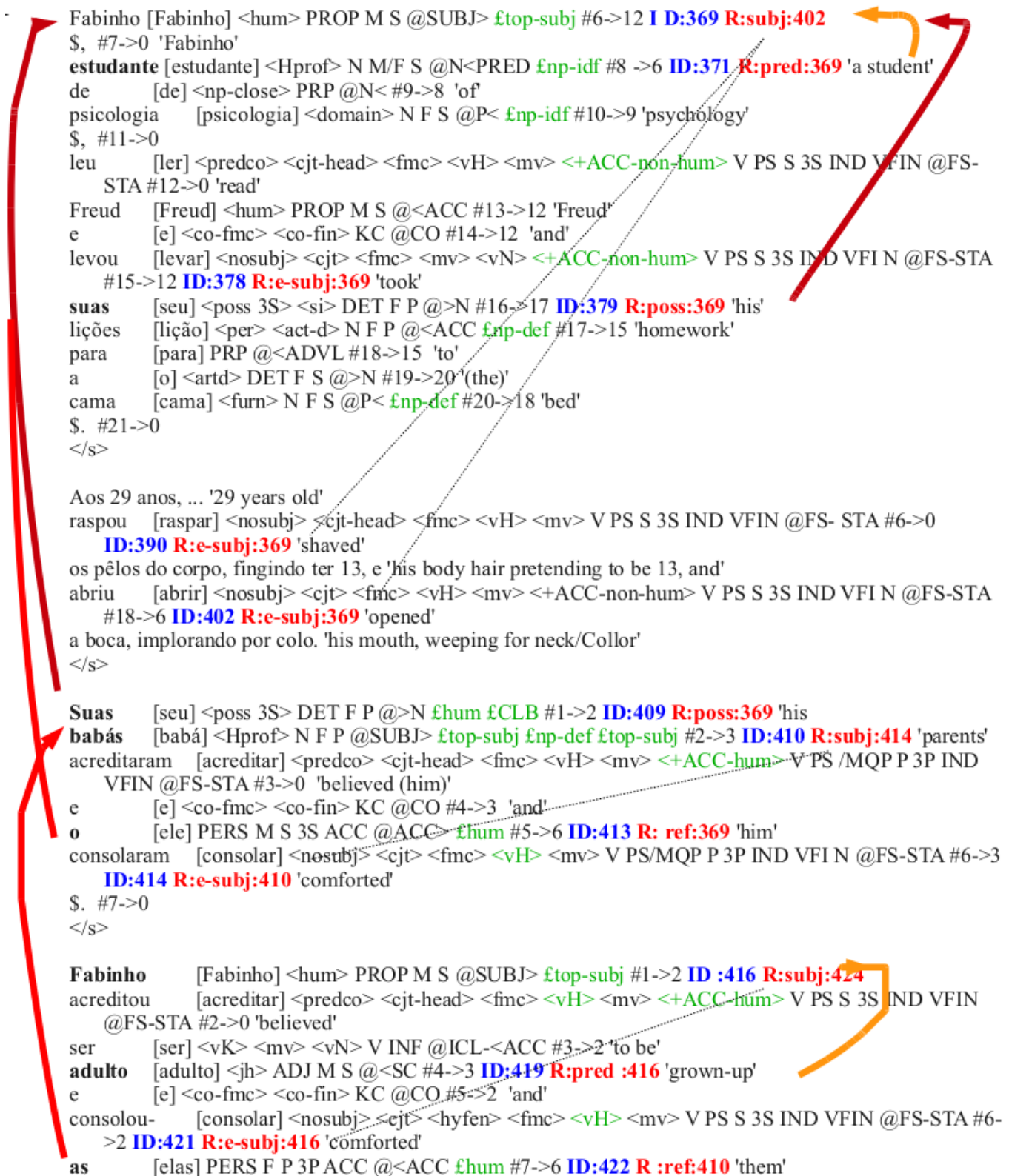
- Task challenges:
  - Automatically extract character names from Portuguese/Brazilian literature works
  - Distinguish between character and non-character person names (e.g. gods, poets, politicians, scientists)
  - Identify all variants of the individual character names, including titles and some family terms (e.g. *Tia X, Mano Y),* but not title-only references
  - Tag characters for gender
  - Identify the characters' professions and profession-like social roles, zero or multiple where relevant
  - Identify family relations between characters, multiple if changing over time
- Data challenges
  - Mostly historical texts, mostly 19th and early 20th century novels
  - Mixed authorship, mixed language varieties, mixed genre
  - May contain orthographical variations from modern Portuguese, affecting tokenization, upper-casing (NER problem), diacritics
  - May contain errors introduced by OCR scanning
  - May contain meta text or non-Portuguese text, sometimes SGML-marked as such
  - Contains a lot of direct speech without quotes

Implemented solution: Primary annotation

- Base annotation
  - Preprocess texts for tokenization and some orthographical normalization (PALAVRAS)
  - Tag texts with morphological, syntactic and semantic information (PALAVRAS)
  - Use existing NER to identify (all) person names (PALAVRAS-NER)
- Add long-distance relations
  - Add **anaphora** relations for pronouns and +HUM nouns, exploiting top-level subject-hood, definiteness, ±HUM, agreement, topic/focus
  - Link **zero-subject** verbs to a (preceding) surface subject
  - Link **vocatives** in direct speech to a surface discourse participant
  - Let all relations cross sentence boundaries (±6), propagate relations across "**stepping-stones**" (pronouns/verbs), optimally until a person name is reached
- Use semantic roles (§ATR, §ID) and framenet links to relate names to +HUM nouns (<Hprof>, <Htit>, <Hfam>)



tmp/pt-3483880.png


Fabinho [Fabinho] <hum> PROP M S @SUBJ> £top-subj #6->12 I D:369 R:subj:402
$, #7->0 'Fabinho'
estudante [estudante] <Hprof> N M/F S @N<PRED £np-idf #8 ->6 ID:371 R:pred:369 'a student'
de [de] <np-close> PRP @N< #9->8 'of'
psicologia [psicologia] <domain> N F S @P< £np-idf #10->9 'psychology'
$, #11->0
leu [ler] <predco> <cjt-head> <fmc> <vH> <mv> <+ACC-non-hum> V PS S 3S IND VFIN @FS-STA #12->0 'read'
Freud [Freud] <hum> PROP M S @<ACC #13->12 'Freud'
e [e] <co-fmc> <co-fin> KC @CO #14->12 'and'
levou [levar] <nosubj> <cjt> <fmc> <mv> <vN> <+ACC-non-hum> V PS S 3S IND VFI N @FS-STA #15->12 ID:378 R:e-subj:369 'took'
suas [seu] <poss 3S> <si> DET F P @>N #16->17 ID:379 R:poss:369 'his'
lições [lição] <per> <act-d> N F P @<ACC £np-def #17->15 'homework'
para [para] PRP @<ADVL #18->15 'to'
a [o] <artd> DET F S @>N #19->20 '(the)'
cama [cama] <furn> N F S @P< £np-def #20->18 'bed'
$. #21->0
</s>
Aos 29 anos, ... '29 years old'
raspou [raspar] <nosubj> <cjt-head> <fmc> <vH> <mv> V PS S 3S IND VFIN @FS- STA #6->0 ID:390 R:e-subj:369 'shaved'
os pêlos do corpo, fingindo ter 13, e 'his body hair pretending to be 13, and'
abriu [abrir] <nosubj> <cjt> <fmc> <vH> <mv> <+ACC-non-hum> V PS S 3S IND VFI N @FS-STA #18->6 ID:402 R:e-subj:369 'opened'
a boca, implorando por colo. 'his mouth, weeping for neck/Collor'
</s>
Suas [seu] <poss 3S> DET F P @>N £hum £CLB #1->2 ID:409 R:poss:369 'his'
babás [babá] <Hprof> N F P @SUBJ> £top-subj £np-def £top-subj #2->3 ID:410 R:subj:414 'parents'
acreditaram [acreditar] <predco> <cjt-head> <fmc> <vH> <mv> <+ACC-hum> V PS /MQP P 3P IND VFIN @FS-STA #3->0 'believed (him)'
e [e] <co-fmc> <co-fin> KC @CO #4->3 'and'
o [ele] PERS M S 3S ACC @ACC> £hum #5->6 ID:413 R: ref:369 'him'
consolaram [consolar] <nosubj> <cjt> <fmc> <vH> <mv> V PS/MQP P 3P IND VFI N @FS-STA #6->3 ID:414 R:e-subj:410 'comforted'
$. #7->0
</s>
Fabinho [Fabinho] <hum> PROP M S @SUBJ> £top-subj #1->2 ID :416 R:subj:424
acreditou [acreditar] <predco> <cjt-head> <fmc> <vH> <mv> <+ACC-hum> V PS S 3S IND VFIN @FS-STA #2->0 'believed'
ser [ser] <vK> <mv> <vN> V INF @ICL-<ACC #3->2 'to be'
adulto [adulto] <jh> ADJ M S @<SC #4->3 ID:419 R:pred :416 'grown-up'
e [e] <co-fmc> <co-fin> KC @CO #5->2 'and'
consolou- [consolar] <nosubj> <cjt> <hyfen> <fmc> <vH> <mv> V PS S 3S IND VFIN @FS-STA #6->2 ID:421 R:e-subj:416 'comforted'
as [elas] PERS F P 3P ACC @<ACC £hum #7->6 ID:422 R :ref:410 'them'

Stepping stones

1 "XXX" main referent: **top level** **@subject** (PROP, +HUM, np-def)
--> **R:be:6**
--> **<NA:Hprof/médico>**

...

2 subject pronoun **<REF:XXX>** <R:ref:1>

...

3 subject pronoun **<REF:XXX>** <R:ref1> **R:subj:4**

...

4 subject-less VFIN **R:e-subj:3**

...

5 subject pronoun R:ref:4 R:be:6
--> R:ref:4 --> R:ref:3 --> R:ref:1
--> <REF:XXX>

6 "médico" <Hprof> **@SC** **§ATR** R:n-attr:5
--> **R:n-attr:1**

Implemented solution: Secondary annotation

- Use Constraint Grammar to add DIP information
  - Map name ID's on +HUM nouns and pronouns
    - e.g. **<REF:Nuno>**
  - Map professions on person names, based on §ATR and §ID links, or through special rules exploiting other clues (e.g. verbs, <inst> or <org> tags)
    - e.g. **<NA:Hprof/ministro>**
  - Map named relations, for family relations, between names or +HUM nouns (the latter for later propagation to a name), e.g. *spouse_of, sibling_of*
    - use two-way relations with separate direction tags for asymmetrical family relations, e.g. *child_of* <-> *parent_of*, *grandchild_of* <-> *grandparent_of*
  - Map a name-of-relative tag on the names at both ends of a family relation link, exploiting the link's name and target, anaphora-propagating, if necessary, from noun targets to name antecedents
    - e.g. **<RI:parent_of:Capituzinha>**
  - Tag person names as safe non-characters (<cult>) or safe cast characters (<noncult>), based on morphological, semantic and contextual clues
  - Facilitate turntaking identification of speaker & addressee
    - <quoteopen> (--), <quo> (in speech), <nquo> (outside speech), <quote> (quoting verb)

Implemented solution: extraction script (ana2pers)

- read the annotated text, extract explicit/implicit information
- builds a data structure for all name IDs and name relations
  - For each person name token, store labels about gender and profession
  - For each person name token, store relation-name pairs
  - if not tagged explicitely: retrieve relative name from R:relative_of:ID links
  - if link ID leads to pronoun, <nosubj> verb or np anaphor:
    - use the target's <REF:...> tag to extract relation target names
    - in the absence of a <REF:...> tag, follow **stepping stones**, iterating over R:ref:ID links
  - ignore/avoid circular ID references and self-relations
- Coreference resolution:
  - For each name in the data structure, loop through all other names and decide which are aliases ("synonyms"), creating named *synsets* based on
    - shared core name elements (accept *-inho* suffixation, don't use ambiguous first names as synset names)
    - gender and profession unification
    - likelihood of occurring in/outside quotes (high ratio <quo>/<nquo> indicates character-hood)
  - weed out non-cast names based on <cult>/<noncult> tags, frequency and complexity (titles and first names indicate cast names, simplex surnames do not)
  - special treatment for the name "author" (e.g. 1. person outside quotes <snquo>)
- For family relations, note how often the same relation-name combination occurs in the annotated text

Implemented solution: logics & output script (ana2pers_logic)

- reads the data structure prepared by script1
  - concludes family relations from existing information, e.g.
    - family tree propagation: e.g. X child_of Y & Z parent_of Y --> X gchild_of Z
    - symmetry: X auncle_of Y --> Y nephie_of X
    - profession unification: e.g. X spouse_of *médico* & X spouse_of Y & Z child_of *méidico* --> Z child_of Y
  - resolves relation conflicts, e.g.
    - *spouse* through gender or shared *child,* e.g. X spouse_of Y & X parent_of Z --> Y parent_of Z
    - unique relation (*spouse*/*parent)* through frequency, e.g. X child_of_m Y = 3 & X child_of_m Z = 1 --> X ! child_of_m Z
- prepares output:
  - assigns running id's, strips *author* and *friend_of* information
  - adapts distinction betwee titles <Htit> and professions <Hprof>, e.g. militray <Htit> --> <Hprof>,
  - creates Portuguese fullforms, e.g. spouse_of_f --> marido, a(c)tor + F --> a(c)triz
- writes output files (*personagens.csv, relacoes.csv*)

System architecture

tokenization & NER 1

morphological CG

morphological analysis

core lexicon
noun ontology
name database
Framenet

tokenization & NER 2 (CG)

syntactic CG

dependency trees

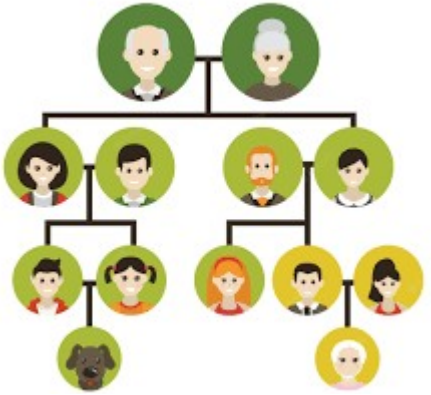
Framenet mapping

frame & role disambiguation (CG)

Relations CG
anaphora, zero constituents
DIP: family & ID relations, professions

ana2pers

ana2pers_logic



1. **Quando** [quando] <clb> <clb-fs> <*> <rel> <ks> ADV @ADVL> §LOC-TMP #1->2 ID:1 R:sd*-LOC-TMP:2
2. **íamos** [ir] <nosubj> <cjt-head> <fn:run> <mv> <vH> V IMPF 1P IND VFIN @FS-ADVL> §LOC-TMP #2->34 ID:2 R:sd*-LOC-TMP:34
3. **a** [a] PRP @<SA #3->2
4. **Andaraí** [Andaraí] <Ltown> <*> PROP M S @P< £civ §DES #4->3 ID:4 R:sd*-DES:2
5. **e** [e] <co-fin> KC @CO #5->2
6. **víamos** [ver] <nosubj> <cjt> <fmc> <vH> <fn:see> <mv> V IMPF 1P IND VFIN @FS-ADVL> §LOC-TMP #6->34 ID:6 R:sd*-LOC-TMP:34
7. **a** [o] <artd> DET F S @>N #7->8
8. **filha** [filha] <Hfam> <fn:spawn> <REF:Capituzinha> N F S @<ACC §STI %np-def #8->6 ID:8 R:soc-rel:10 R:soc-rel:12 R:np-ref:15 R:n-attr:15 R:sd*-STI:6
9. **de** [de] <np-close> PRP @N< #9->8
10. **Escobar** [Escobar] <cjt-head> <*> <nquo> <SR:filha> <RI:parent_of:Capituzinha> PROP M S @P< £hum §ORI #10->9 ID:10 R:have:8 R:sd*-ORI:8 R:parent:15
11. **e** [e] <co-prparg> KC @CO #11->10
12. **Sancha** [Sancha] <cjt> <Hsoc> <*> <nquo> <SR:filha> <RI:parent_of:Capituzinha> PROP M S @P< £hum §ORI #12->9 ID:12 R:have:8 R:sd*-ORI:8 R:parent:15
13. $, [$,] PU @PU #13->0
14. **familiarmente** [familiarmente] <deadj> ADV @>N #14->15
15. **Capituzinha** [Capituzinha] <*> <heur> <nquo> <NA:Hfam/filha> <RI:child_of:Escobar> <RI:child_of:Sancha> PROP F S @APP £hum §ID #15->8 ID:15 R:sd*-ID:8 R:be:8 R:child:10 R:child:12
16. $, [$,] PU @PU #16->0
17. **por** [por] PRP @<ADVL #17->6
18. **diferençar**- [diferençar] <fn:contrast> <vH> <mv> <uninfl> V INF 1P @ICL-P< §CAU %CLB #18->17 ID:18 R:sd*-CAU:6
19. **la** [ela] <REF:Capituzinha> PERS F S 3S ACC @<ACC §TH #19->18 ID:19 R:ref:15 R:sd*-TH:18
20. **de** [de] PRP @<ADVL #20->18
21. **minha** [meu] <poss> <1S> <REF:Author> DET F S @>N %hum #21->22 ID:21 R:have:22
22. **mulher** [mulher] <Hfam> N F S @P< §ORI %np-def #22->20 ID:22 R:soc-rel:21 R:sd*-ORI:18
23. $, [$,] PU @PU #23->0
24. **visto=que** [visto=que] <clb> <clb-fs> KS @SUB #24->26
25. **lhe** [ele] <REF:Capituzinha> PERS F S 3S DAT @DAT> §REC %hum #25->26 ID:25 R:ref:15 R:sd*-REC:26
26. **deram** [dar] <fn:give> <mv> <nosubj> <vH> V PS/MQP P 3P IND VFIN @FS-<ADVL §COND #26->18 ID:26 R:sd*-COND:18
27. **o** [o] <artd> DET M S @>N #27->29
28. **mesmo** [mesmo] <diff> <KOMP> DET M S @>N #28->29
29. **nome** [nome] <f> N M S @<ACC §TH %np-def #29->26 ID:29 R:sd*-TH:26

-- Boa pessoa, repetiu o major, boa criatura ...
Tonica [Tonica] <*> <Hpers> **<noncult>** **<quo>** **<SR:noivo>** PROP F S @**VOK** **£hum** **§VOC** #12->15 ID:12 R:sd*-VOC:15 **R:have:25** R:subj:21
,
vai [ir] <nosubj> <cjt> **<fn:future>** <fmc> <aux> **<quote>** <cjt-head> <vH> **<SUBJ:Tonica>** **<REF:Tonica>** V IMP 2S VFIN @FS-COM #14->0 ID:14 **R:e-subj:12**
buscar o retrato ...
Anda [andar] <nosubj> <cjt> <fmc> <*> <vH> **<fn:run>** <mv> **<quote>** **<SUBJ:Tonica>** **<REF:Tonica>** V IMP 2S VFIN @FS-COM #19->0 ID:19 **R:e-subj:12**
,
vai [ir] <nosubj> <cjt> **<fn:future>** <fmc> <aux> **<quote>** <vH> **<SUBJ:Tonica>** **<REF:Tonica>** V IMP 2S VFIN @FS-COM #21->0 ID:21 **R:e-subj:12**
buscar o
teu [teu] <poss> <2S> **<REF:Tonica>** DET M S @>N **%hum** #24->25 ID:24 **R:poss:12**
noivo [noivo] **<Hfam>** N M S @<ACC **§TH** %np-def #25->22 ID:25 **R:soc-rel:12** R:sd*-TH:22
... Dona Tonica foi buscar o retrato. Era uma fotografia; representava um homem de meia idade, cabelo curto, raro olhando espantado para a gente, cara chupada, pescoço fino e paletot abotoado.
-- Que lhe parece?
-- Muito bem
Dona|=Tonica [Dona|=Tonica] <*> <Hpers> **<nquo>** **<RI:gbfriend_of:Rodrigues>** PROP F S @SUBJ> **£hum** **§REC** **%top-subj** #1->2 ID:75 R:sd*-REC:77 **R:c-subj:81** R:sd-AG-EXP:81
recebeu o retrato e **fitou**-o alguns instantes; mas, tirou logo os olhos, e deixou-se estar sentada, enquanto a imaginação saiu a esperar o
Rodrigues [Rodrigues] <*> <Hpers> **<noncult>** **<nquo>** PROP M S @<ACC **£hum** **§EV** %np-def #21->19 ID:107 **R:subj:109** R:sd*-EV:105
.
Chamava- [chamar] **<fn:name>** <fmc> <*> <vH> <mv> <nosubj> **<SUBJ:Rodrigues>** **<REF:Rodrigues>** V IMPF S 3S IND VFIN @FS-STA #1->0 ID:109 **R:e-subj:107**
se [se] <obj> PERS M/F S 3S ACC @<ACC **§REFL** **%hum** #2->1 ID:111 R:refl:107 R:p-attr:112 R:sd*-REFL:109
Rodrigues [Rodrigues] <*> <Hpers> **<noncult>** **<nquo>** PROP M S @<OC **£hum** **§ATR** #3->1 ID:112 R:sd*-ATR:109 R:c-attr:111 R:pred:111
.

Rule example: relations through addressing

- ***Nuno**, que já andava no colégio, onde ouvira falar da mudança de gabinete, disse ao **pai** que queria ser **ministro**. **Teófilo** ficou sério. -- Meu filho, disse **ele**, escolhe outra coisa, menos ministro.*
- SUBSTITUTE:ri REPEAT (PROP) (VSTR:<RI:child_of:$1> PROP)
  TARGET PROP + £hum + @SUBJ + %top-subj
  (p V-SPEAK [disse-1]
  LINK *1 §REC + %np-def + &&GN& BARRIER VV OR CLB [pai]
  LINK 0 ("pai") OR ("mãe") [pai]
  LINK *1 <<<
  LINK *1W PROP + %top-subj BARRIER ALL-ORD [Teófilo]
  LINK 0 &&GN& LINK 0 ("([^<]+)"r) *harvesting $1 name*
  LINK *1 <<<
  LINK *1W HYFEN& OR QUOTE& OR <quote-edge> BARRIER ALL-ORD [--]
  LINK *1 V-SPEAK + VFIN [disse-2]
  LINK c @SUBJ + (PERS NOM) + &&GN& [ele]
) ;
- Nuno [Nuno] <*> **<NA:Hprof/ministro>** **<SR:pai>** **<RI:child_of:Teófilo>** PROP M S @SUBJ> **£hum §SP** %np-def %top-subj %CLB #1->19 ID:1 R:sd-AG:5 **R:c-subj:5** R:have:22 **R:subj:11** **R:subj:24** R:be:26 **R:sd*-SP:19**

Family relation types

SPOUSE_OF (F) - mulher
SPOUSE_OF (M) - marido
SPOUSE_OF_M (F) - mulher
SPOUSE_OF_M (M) - marido
SPOUSE_OF_F (F) - mulher
SPOUSE_OF_F (M) - marido
PARENT_OF (M) - pai
PARENT_OF (F) - mãe
AUNCLE_OF (M) - tio
AUNCLE_OF (F) - tia
COUSIN_OF (M) - primo
COUSIN_OF (F) - prima
CHILD_OF (M) - filho
CHILD_OF (F) - filha
CHILD_OF_M (M) - filho
CHILD_OF_M (F) - filha
CHILD_OF_F (M) - filho
CHILD_OF_F (F) - filha
GCHILD_OF (M) - neto
GCHILD_OF (F) - neta
GGCHILD_OF (M) - bisneto
GGCHILD_OF (F) - bisneta
GODCHILD_OF (M) - afilhado
GODCHILD_OF (F) - afilhada
FRIEND_OF (M) - amigo
FRIEND_OF (F) - amiga

GBFRIEND_OF (M) - noivo
GBFRIEND_OF (F) - noiva
GPARENT_OF (M) - avô
GPARENT_OF (F) - avó
GGPARENT_OF (M) - bisavô
GGPARENT_OF (F) - bisavó
GODPARENT_OF (M) - padrinho
GODPARENT_OF (F) - madrinha
GODSIBLING_OF (M) - compadre
GODSIBLING_OF (F) - comadre
ICHILD_OF (M) - genro
ICHILD_OF (F) - nora
ICHILD_OF_M (M) - genro
ICHILD_OF_M (F) - nora
ICHILD_OF_F (M) - genro
ICHILD_OF_F (F) - nora
IPARENT_OF (M) - sogro
IPARENT_OF (F) - sogra
ISIBLING_OF (M) - cunhado
ISIBLING_OF (F) - cunhada
NEPHIE_OF (M) - sobrinho
NEPHIE_OF (F) - sobrinha
SIBLING_OF (M) - irmão
SIBLING_OF (F) - irmã
WIDOW_OF (F) - viúva
WIDOW_OF (M) - viúvo

Creating a network data table
as a tool for literary text analysis

- Use narrative characters as nodes
  - standardized, unique id name as node name (u-name)
  - as-is "surface name" as attribute (s-name)
- Use family relations as directed arcs ("edges") between nodes
- Add un-named relations between co-occurring characters
  - if there are less than 3 semantic content tokens between them ("semantic" = carrying a semantic role or frame tag)
  - if there is no <addbreak> tag between them (recognized titles, chapter breaks)
- Create a .csv table with 6 fields for source node (SN), relation (REL) and target node (TN):
  - SN u-name, SN s-name, REL name, nil (rel-strength), TN u-name, TN s-name
- cat table | *uniq -c,* use count as relation strengths in field (4)
- import the table to *Cytoscape* for network visualizing and analysis

Data table

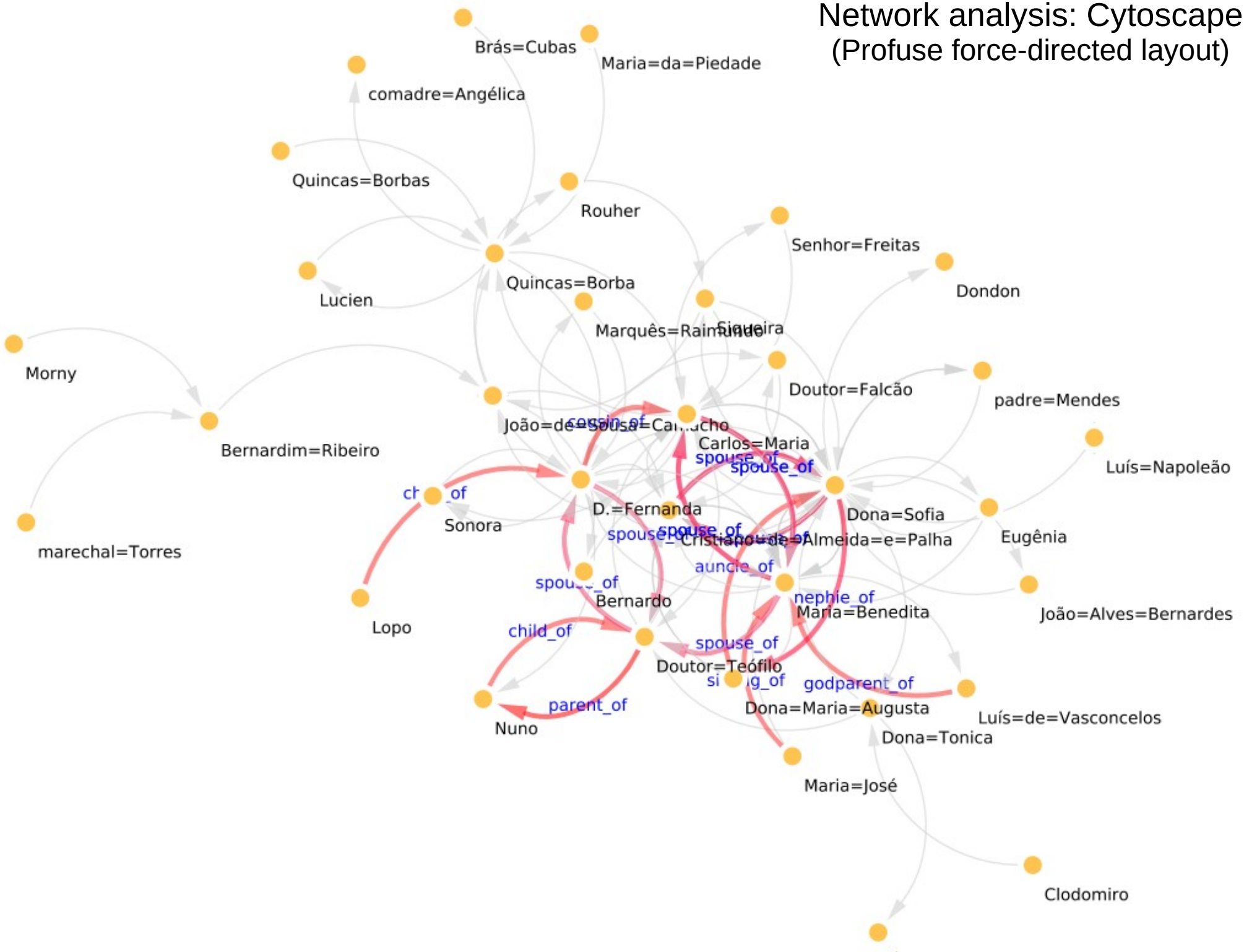
shared name	shared interactic	Source node attribute	Edge Attribu	Target node attribute
Carlos=Maria (spouse_of) Maria=Benedita	spouse_of	Carlos=Maria	ct=2	Maria=Benedita
Carlos=Maria (spouse_of) Maria=Benedita	spouse_of	Carlos=Maria	ct=1	Maria
Cristiano=de=Almeida=e=Palha (spouse_of) Sofia	spouse_of	Cristiano=de=Almeida=e=Palha	ct=1	Sofia
Cristiano=de=Almeida=e=Palha (spouse_of) Sofia	spouse_of	Palha	ct=1	Sofia
Dona=Fernanda (spouse_of) Doutor=Teófilo	spouse_of	Dona=Fernanda	ct=2	Teófilo
Doutor=Teófilo (spouse_of) Dona=Fernanda	spouse_of	Teófilo	ct=3	Dona=Fernanda
Maria=Benedita (spouse_of) Carlos=Maria	spouse_of	Maria	ct=1	Carlos=Maria
Maria=Benedita (spouse_of) Carlos=Maria	spouse_of	Maria=Benedita	ct=1	Carlos=Maria
Maria=Benedita (spouse_of) Doutor=Teófilo	spouse_of	Maria=Benedita	ct=1	Teófilo
Sofia (spouse_of) Cristiano=de=Almeida=e=Palha	spouse_of	Sofia	ct=1	Cristiano=de=Almeida=e=...
Sofia (spouse_of) Cristiano=de=Almeida=e=Palha	spouse_of	Sofia	ct=1	Palha
Maria=José (sibling_of) Maria=Benedita	sibling_of	Maria=José	ct=1	Maria=Benedita
Doutor=Teófilo (parent_of) Nuno	parent_of	Teófilo	ct=1	Nuno
Sofia (nephie_of) Dona=Maria=Augusta	nephie_of	Sofia	ct=1	Dona=Maria=Augusta
Luís=de=Vasconcelos (godparent_of) Maria=Bene...	godparent_of	Luís=de=Vasconcelos	ct=1	Maria=Benedita
Dona=Fernanda (cousin_of) Carlos=Maria	cousin_of	Dona=Fernanda	ct=1	Carlos=Maria
Lopo (child_of) Dona=Fernanda	child_of	Lopo	ct=1	Dona=Fernanda
Nuno (child_of) Doutor=Teófilo	child_of	Nuno	ct=1	Teófilo
Dona=Maria=Augusta (auncle_of) Sofia	auncle_of	Dona=Maria=Augusta	ct=1	Sofia
Bernardim=Ribeiro () Bernardim=Ribeiro		marechal=Pio	ct=1	marechal=Ribeiro
Bernardim=Ribeiro () João=de=Sousa=Camacho		Pio	ct=1	Camacho
Bernardo () Sofia		Bernardo	ct=1	Sofia
Brás=Cubas () Quincas=Borba		Brás=Cubas	ct=1	Quincas=Borba
Carlos=Maria () Dona=Fernanda		Carlos=Maria	ct=4	Dona=Fernanda
Carlos=Maria () Doutor=Teófilo		Carlos=Maria	ct=2	Teófilo
Carlos=Maria () Maria=Benedita		Carlos=Maria	ct=14	Maria=Benedita
Carlos=Maria () Quincas=Borba		Carlos=Maria	ct=1	Quincas=Borba
Carlos=Maria () Senhor=Freitas		Carlos=Maria	ct=5	Freitas
Carlos=Maria () Sofia		Carlos=Maria	ct=11	Sofia

safe link

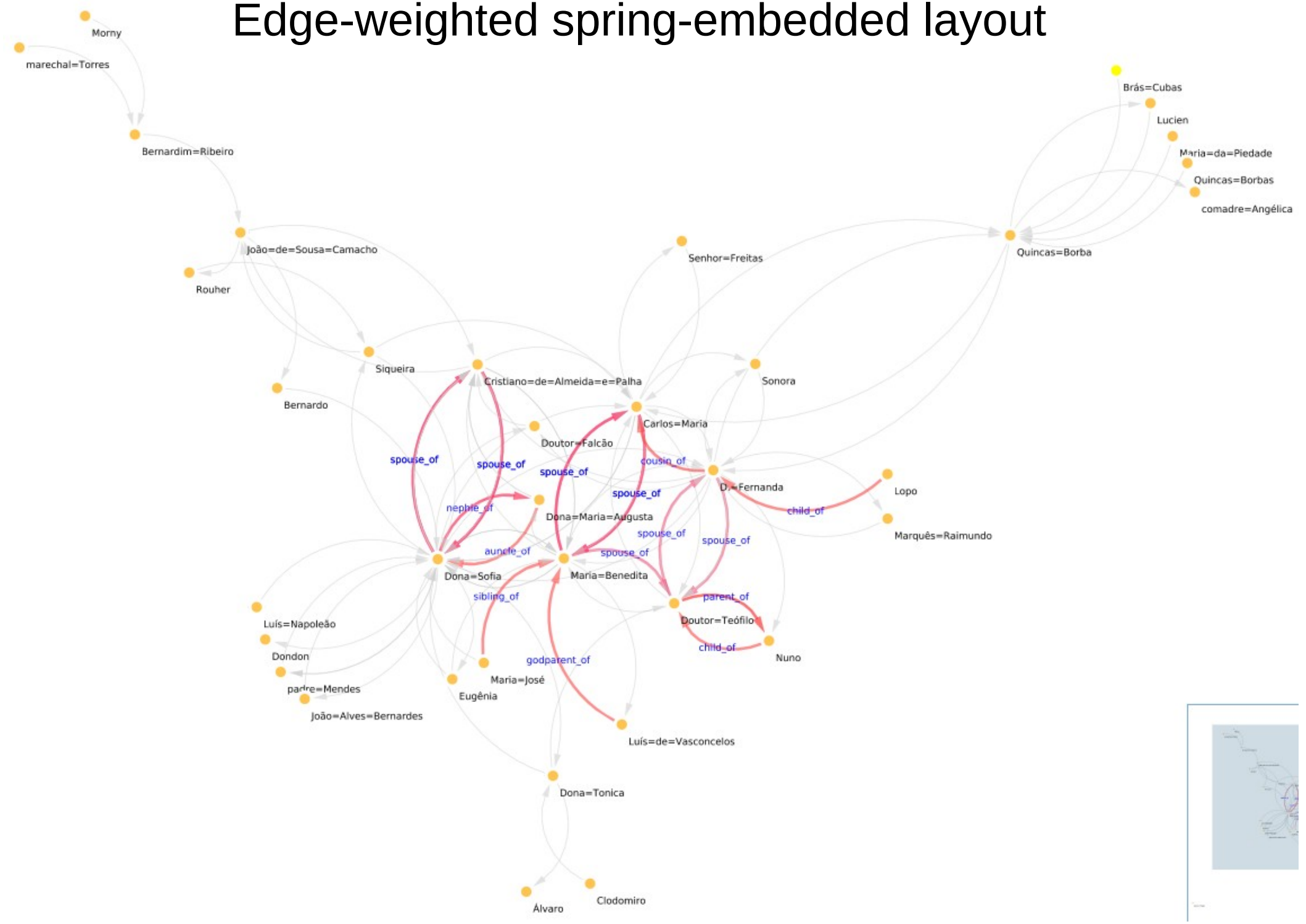
family?

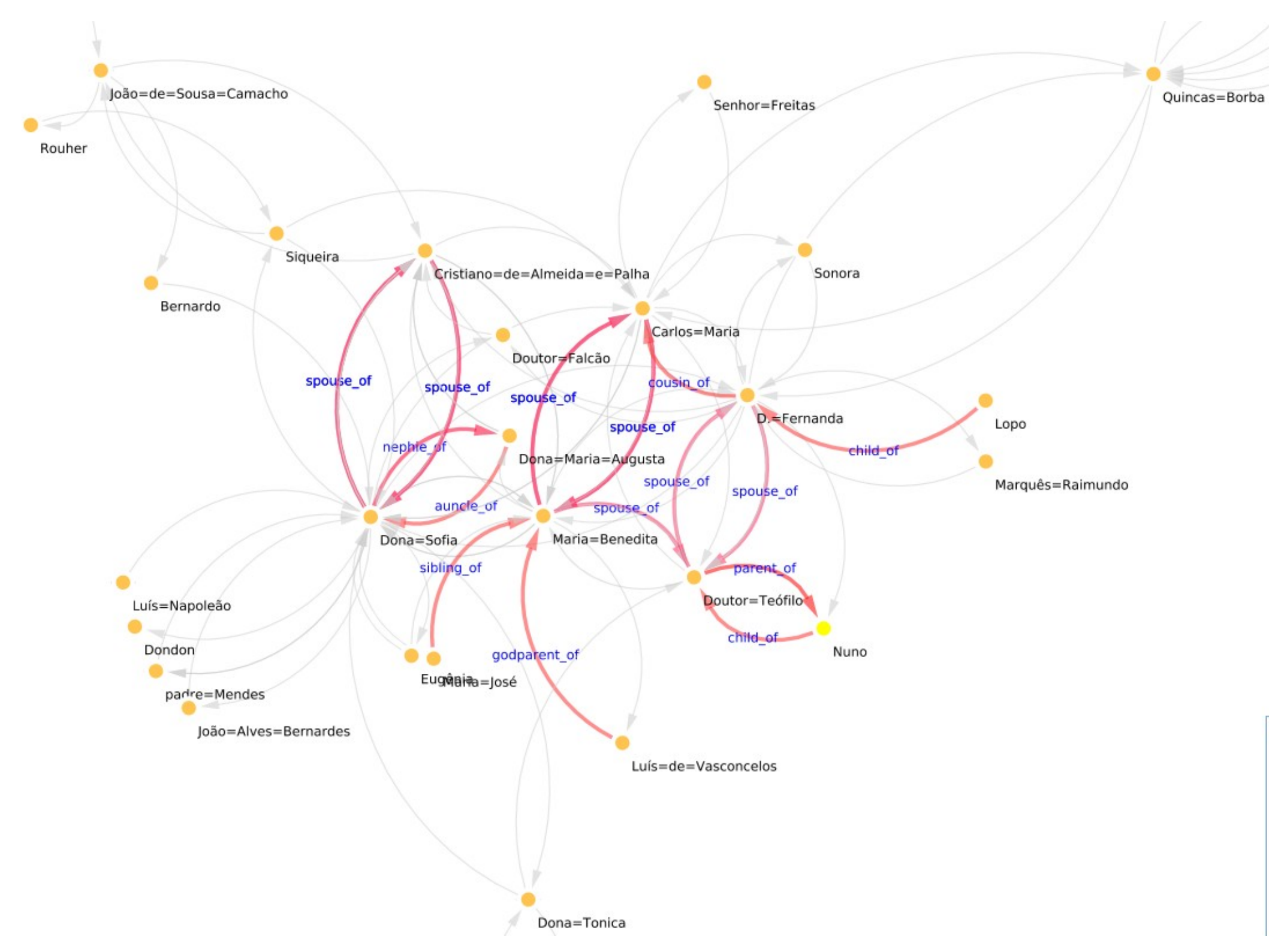
family?

Network analysis: Cytoscape
(Profuse force-directed layout)



Edge-weighted spring-embedded layout




João=de=Sousa=Camacho
Rouher
Siqueira
Bernardo
Senhor=Freitas
Quincas=Borba
Cristiano=de=Almeida=e=Palha
Sonora
Carlos=Maria
Doutor=Falcão
spouse_of
spouse_of
spouse_of
spouse_of
cousin_of
D.=Fernanda
Lopo
child_of
Marquês=Raimundo
nephie_of
Dona=Maria=Augusta
spouse_of
spouse_of
auncle_of
spouse_of
Dona=Sofia
Maria=Benedita
sibling_of
parent_of
Doutor=Teófilo
child_of
Nuno
Luís=Napoleão
Dondon
padre=Mendes
João=Alves=Bernardes
Eugénia
Maria=José
godparent_of
Luís=de=Vasconcelos
Dona=Tonica


Siqueira
Cristiano=de=Almeida=e=Palha
Sonora
Carlos=Maria
Doutor=Falcão
spouse_of
spouse_of
spouse_of
cousin_of
spouse_of
D.=Fernanda
Lopo
child_of
Marquê
nephie_of
Dona=Maria=Augusta
spouse_of
spouse_of
spouse_of
auncle_of
spouse_of
Dona=Sofia
Maria=Benedita
sibling_of
parent_of
Doutor=Teófilo
child_of
Nuno
godparent_of
Maria=José
Eugênia
lves=Bernardes
Luís=de=Vasconcelos

Network parameters

shared name	name	AverageS	Clustering	Closenes:	IsSinglel	PartnerOf	SelfLoops	Eccentrici	Stress	EdgeCount
Sofia	Sofia	1.76923...	0.08095...	0.56521...	☐	10	0	3	645	33
Dona=Fernanda	Dona=Fe...	1.76923...	0.2	0.56521...	☐	7	0	3	399	22
Carlos=Maria	Carlos=...	1.92307...	0.25555...	0.52	☐	6	0	4	386	21
Cristiano=de=Almeida=e=Palha	Cristiano...	2.07692...	0.4047619	0.48148...	☐	4	0	3	291	19
Maria=Benedita	Maria=B...	1.88461...	0.23333...	0.53061...	☐	7	0	3	287	23
Quincas=Borba	Quincas...	2.42307...	0.04761...	0.41269...	☐	3	0	4	269	10
João=de=Sousa=Camacho	João=de...	2.84615...	0.05	0.35135...	☐	1	3	4	248	12
Doutor=Teófilo	Doutor=...	2.42307...	0.35	0.41269...	☐	4	0	4	108	13
Dona=Tonica	Dona=To...	2.53846...	0.0	0.39393...	☐	1	1	4	106	7
Bernardim=Ribeiro	Bernardi...	3.74074...	0.0	0.26732...	☐	0	1	5	100	5
Siqueira	Siqueira	2.15384...	0.25	0.46428...	☐	1	0	3	67	5
prima=Sofia	prima=S...	2.76923...	0.33333...	0.36111...	☐	1	0	4	61	4
Bernardo	Bernardo	2.65384...	0.0	0.37681...	☐	0	0	4	44	2
Rouher	Rouher	3.07692...	0.5	0.325	☐	0	0	4	4	2
padre=Chagas	padre=C...	0.0	0.0	0.0	☑	0	1	0	0	2
Luís=Napoleão	Luís=Na...	2.70370...	0.0	0.36986...	☐	0	3	4	0	7
viúva=Mendes	viúva=M...	2.70370...	0.0	0.36986...	☐	0	0	4	0	1
Maria=José	Maria=José	2.51851...	1.0	0.39705...	☐	0	0	4	0	2
Doutor=Falcão	Doutor=F...	3.0	0.83333...	0.33333...	☐	0	0	4	0	3
Lopo	Lopo	2.70370...	0.0	0.36986...	☐	0	0	4	0	1
Brás=Cubas	Brás=Cu...	3.33333...	0.0	0.3	☐	0	0	5	0	1
Maria=da=Piedade	Maria=d...	3.33333...	0.0	0.3	☐	0	1	5	0	3
comadre=Angélica	comadre...	0.0	0.0	0.0	☐	0	0	0	0	1
Quincas=Borbas	Quincas...	3.33333...	0.0	0.3	☐	0	0	5	0	1
Clodomiro	Clodomiro	3.44444...	0.0	0.29032...	☐	0	0	5	0	1
Álvaro	Álvaro	0.0	0.0	0.0	☐	0	0	0	0	1
Morny	Morny	4.60714...	0.0	0.21705...	☐	0	0	6	0	1
marechal=Torres	marechal...	4.60714...	0.0	0.21705...	☐	0	0	6	0	1
Dondon	Dondon	0.0	0.0	0.0	☐	0	0	0	0	1
Atalaia	Atalaia	1.0	0.0	1.0	☐	0	0	1	0	1

Related work

- **Bamman, David; Underwood, Ted & Smith, Noah A. 2014.** A bayesian mixed effects model of literary character. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).
  - 74% of character occurrences are pronouns!
  - uses linear word distance and gender for pronoun anaphora resolution, performs automatic name clustering (co-reference resolution)
- Bornet, C & Kaplan, F. 2017. A Simple Set of Rules for Characters and Place Recognition in French Novels. Frontiers in Digital Humanities 4 (2017), 6. https://doi.org/10.3389/fdigh.2017.00006
  - Person NER on literary works, but no co-reference solution
- **Dekker, N; T. Kuhn & M. van Erp. 2019.** Evaluating named entity recognition tools for extracting social networks from novels. PeerJ Computer Science 5 (2019), e189. https://doi.org/10.7717/peerj-cs.189
  - NER systems worked ok on *1984, Huckleberry Finn, Ulysses* and *Game of Thrones* (80-90% F-scores for the best systems), but many had single-digit F-scores on *Brave New World*
  - uses simple co-occurrence for network analysis
- Goyal, Amit; Riloff, Ellen & Daumé, Hal III. 2010. Automatically producing plot unit representations for narrative text. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 77–86. Association for Computational Linguistics.
  - characters are presumed given in the title, identification is of affect states, not the characters themselves
- **Labatut, Vincent & Bost, Xavier. 2019.** Extraction and Analysis of Fictional Character Networks: A Survey. ACM Computing Surveys 52(5):89. https://doi.org/10.1145/3344548
  - meta article on character identification, character interaction as network node edges
- Mamede, Nuno & Chaleira, Pedro. 2004. Character identification in children stories. In Advances in natural language processing, pages 82–90. Springer.
  - links discourse utterances to characters (or the narrator)
- Paul, A., & Das, D. (2017). A Deep Dive into Identification of Characters from Mahabharata. ICON 1.
  - neural networks, word and phrase level features
- Valls-Vargas, Josep; Santiago Ontanón, Santiago & Zhu, Jichen. 2014. Toward automatic character identification in unannotated narrative text. In Seventh Intelligent Narrative Technologies Workshop
  - feature-vector using both linguistic information and external knowledgeuck

eckhard.bick@mail.dk

base annotation (PALAVRAS):

https://visl.sdu.dk/pt/parsing/automatic/dependency.php

--> Parser menu: frames & semantic roles