A Combined Approach to Gender Interpretation of Referential Personal Nouns in English

Thesis Project, to obtain the
Diploma of Advanced Studies (DEA)
PhD program in Linguistic Sciences and Applied Linguistics
Biennial 2004-2006

Directed by
Dr. Núria Bel Rafecas

Manuel Souto Pico
m.soutopico@gmail.com
Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra

Barcelona, 16th October 2007

Contents

Typographic conventions **vi**

Abbreviations **vii**

Introduction **1**

0.1 A brief note on terminology 2

1 Preliminary exploration **4**

1.1 Gender asymmetries between languages 5

1.2 Gender asymmetries and machine translation 9

1.2.1 Nouns 9

1.2.2 Adjectives 11

1.2.3 *One* pronoun 11

1.2.4 Verbs 12

1.2.5 Other parts of speech 13

1.3 Excluded cases 13

1.3.1 First and second persons 13

1.3.2 Non-referential contexts 14

2 Conceptual framework **15**

2.1 Bibliographical philias <= include in the introduction of this chapter, with no separate section 15

2.2 Language and the world 15

2.2.1 Reference 16

2.2.2 Non-referring expressions 17

2.2.3 Person reference 19

2.3 Language and gender 27

2.3.1 Types of gender 27

2.3.2 Noun classification systems 29

2.3.3 Expressing gender in English 32

2.3.4 Gender agreement 37

2.4 Language and the text 41
2.4.1 Lexical cohesion 41
2.4.2 Endophora and coreference 42
2.4.3 Realizations of nominal endophora 44
2.5 Equivalence in translation 48

3 Research purpose/outline/scheme/main points (NOT SURE ABOUT THIS CHAPTER HEADING, WHAT SHOULD THIS CHAPTER BE CALLED?) 50
3.1 Problem statement 50
3.2 The proposed task 52
3.3 Objectives 53
3.4 Hypothesis 53
3.5 Variables 53
3.6 Justification 54

4 Brief state of the art 56

5 Methodology 59
5.1 Gender interpretation vs. cohesion resolution 60
5.2 Preprocessing 61
5.2.1 Data collection 61
5.2.2 Corpus annotation 62
5.2.3 Preprocessing personal names 64
5.2.4 External resources 70
5.2.5 The tagset 71
5.2.6 Classes 72
5.3 Feature extraction 72
5.3.1 Feature +: supplementary apposition 73
5.3.2 Feature 8: marriage 74
5.3.3 Feature B: copula 75
5.3.4 Feature j: poss. + noun + as + role 76
5.3.5 Feature n: name + *as* phrase 76
5.3.6 Feature 1: rough *as* 77
5.3.7 Feature 2: possessive + partner 77
5.3.8 Feature x: gender-specific modifier 78
5.3.9 Feature O: orthographic overlapping 78
5.3.10 Feature J: justice: in his ruling 79
5.3.11 Feature ʃ: reported speech 79
5.3.12 Feature r: next adjoining (#CONTIGUO) pronoun 80
5.3.13 Feature R: previous adjoining (#CONTIGUO) pronoun . 80

5.3.14 Feature s: next rough pronoun 81
5.3.15 Feature S: previous rough pronoun 82
5.3.16 Feature p: next adjoining (#CONTIGUO) name or role . 83
5.3.17 Feature P: previous adjoining (#CONTIGUO) name or role 83
5.3.18 Feature q: next rough name or role 84
5.3.19 Feature Q: previous rough name or role 84
5.3.20 Feature &: coordination 85
5.4 Classification 86
5.5 Evaluation OR Main problems 88

6 Conclusions **90**
6.1 Proposal of future research for the thesis 90
6.1.1 Recurrent people's names database 90
6.1.2 Exploitation of images in web pages 91

Appendices **91**

Bibliography **93**

List of Tables

2.1 Reference items 26
2.2 Expression of referential gender in different language types . . 31
2.3 Greek gender-specific patronyms 36
2.4 Gender-specificity of titles 38
2.5 Gender expression in person reference in English 38
2.6 Gender-specificity of *one*'s translation according to its type . . . 47

5.1 Evaluation for masculine and feminine given names 69
5.2 Tagset 72

List of Figures

2.1 Direction of possible gender-biased translation. 32
2.2 Relation between referential gender and agreement. 40

5.1 Processing pipeline. 59

Typographic conventions

*asterisk	For ungrammatically or extreme abnormality. (#REPLACE WITH # AO FINAL)
blue box	For full NP cues.
blue text	For the constituent which is the cue, the grammatical features of which have been generalized to the whole NP.
bold italics	For emphasis.
bold type	For important terms when first introduced; for family names in some examples.
“double quotation marks”	For quotations from other authors.
[GRAMMAR INFORMATION]	For non-overt grammatical information, e.g. mother[F].
(GRAMMAR INFORMATION)	For overt grammatical information, e.g. kid (F).
gray box	For negative cues or irrelevant expressions for the case in question.
green text	Discourse patterns used as features.
italics	For citation forms when not set as displayed examples.
?question mark	For semantic oddness.
****gender-biased translation***	For gender-biased machine translations.
red box	For targets.
red text	For the constituent which is the target within a target NP.
‘single quotation marks’	For metalinguistic reference and for meanings.
SMALL CAPITALS	For grammatical information in glosses.
[± SMALL CAPITALS]	For semantic components.

Abbreviations

All abbreviated terms will be given in full form on first appearance but it is useful to present here which abbreviations will be used.

AdjP	adjetive phrase
AmE	American English
Ar.	Arabic
Ca.	Catalan
En.	English
FEM/f	feminine
Fi.	Finnish
Fr.	French
GenLessLang	genderless language(s)
GramGenLang	grammatical gender language(s)
Gl.	Galizan
Gt	Google Translate
INV	invariable
MASC/m	masculine
MT	machine translation
NatGenLang	natural gender language(s)
NP	noun phrase
PC	predicative complement
PL	plural
POS	part of speech
poss.	possessive
Pt.	Portuguese
REFL	reflexive pronoun
SG	singular
SL	source language
ST	source text
Sp.	Spanish
TL	target language
Tl	Translendium
TT	target text
VP	verb phrase

Introduction

BASTANTE PENDENTE DE ACABAR, ESBOÇO...

Personal nouns are typically gender-indefinite or dual-gender[1] in English but their equivalents might not be in the target language when the text in which they appear is subject to translation in referential contexts. Thus, the gender information (either feminine or masculine) related to the individuals they refer to is underspecified in the personal noun and, for the sake of the translation's quality, the translator must get that information somewhere else, either from the very text or from world knowledge. A human translator will typically resort to other expressions possibly present in the co-text that refer to the same person and which might be gender-specific. If none is found, the second strategy will be to rely on the gender stereotypes she has got regarding the class of people that personal noun denotes (e.g. nurses will be females, mechanics will be males, etc.)[2] or resort to the masculine as a default gender. On the other hand, automatic translation engines tend to implement the second strategy only, especially when the gender cues are to be found out of the sentence that contains the personal noun at stake.

JUSTIFICATION: Given two purportedly coreferring expressions, if their grammatical genders clash, the reader might either misunderstand and assign the two expressions to two different referents, or realize the translation error and overcome the mismatch, what involves a time-consuming correction of the initial representation.

LINGUISTIC DOUBT: assign a noun to a gender OR a noun must be assigned to its gender

the more precise intended meanings of these terms will become clearer later

[1] I will use "gender-indefinite noun" and "dual-gender noun" as synonymous expressions throughout this dissertation. I will use fairly general terms and keep the technical jargon to a minimum until I get down to explaining the main concepts in Ch. 2.

[2] Some research indicate that gender stereotypes are automatically activated in the absence of disambiguating information. "Violating stereotypes: eye movements and comprehension processes when text conflicts with world knowledge." duffy&keir.pdf

0.1 A brief note on terminology

Two terms are used in the literature as regards the gender that an expression must have. **Gender resolution** refers to the rules speakers use to specify the form of agreeing elements when the controller consists of conjoined NPs with conflicting genders, whereas **gender assignment** refers to the mechanism by which, based on either semantic or morphological factors, native speakers of a language consistently know what gender a controller noun has so that they can produce the correct agreements (Corbett 1991: 7, 261).

For the purposes of coining an appropriate terminology for our research, we could enlarge the concept 'assignment' so as to include also the means by which a reader of a text establishes the gender of a controller noun based on the gender of the elements which agree with it. However, we would like to distance from the above-mentioned terms in as much as they depict a process on the side of the producer of the text, whereas we care only about the speaker that has that text as input, not output. For that reason, we prefer to designate by **gender interpretation** the process of interpreting the gender of a gender-indefinite noun with resort to the form of all elements with agree with it.
The important point here is that, on the one hand, referential gender is represented somehow in virtually all languages of the world but, on the other hand, personal nouns are not gender-specific in all languages. So when a personal noun is not gender-specific, normally there are other elements in the text which do reflect the gender of the referent.

English gender is a covert category -> problem...

The model assumes a two-stage process. First, balblabla´. Second, blablablá.

Combined because it combines heuristics and machine-learning techniques:

Heuristics: names… , ML: nouns

The aim of this dissertation is a proof of concept. This is a description of a certain method used to execute the task proposed in order to demonstrate its feasibility and that it is probably exploitable in a useful manner.

****

Different languages present different features and a feature present in one language does not necessarily appear to be present in other languages. When translating between two dissimilar languages in this regard, the translator has to be warned as far as these differences are concerned and make up somehow for the information that is missing or redundant in the target language. Gender is one such a linguistic feature.

The translator must take into account how and if languages express or mark gender when translating between from one language with gender marking into one without it, or the other way round. In this dissertation, I will only deal with the first case.

Examples are, whenever possible, taken from real texts from the Web, preferably from the New York Times, and only exceptionally from quoted examples in reference books.

Chapter 1

Preliminary exploration

The absence of a one-to-one correspondence between a graphical form and a meaning (i.e. homonymy or polysemy) or reference leads sometimes to ambiguity problems for computational applications such as information retrieval or machine translation. For example, the field of word sense disambiguation deals with lexical ambiguities such as those brought about by the underlined words in the following examples:

(1) Era una política agresiva.[1]

(2) Je connais cet avocat, il est très bon !

(3) Al meu cap no li senta bé tanta farra.

With a view to machine translation (or translation in general, we could say), the underlined words in the examples should be disambiguated because there might be a asymmetry between the source and the target languages as regards lexicosemantic univocity (the number of words corresponding to the intended meaning). For example, the equivalents in English as a target language for Sp. *política*, Fr. *avocat* or Ca. *cap* could be either *(female) politician* or *policy* (or even *politics*), either *lawyer* or *avocado* and either *head* or *boss*, respectively.

Similarly, grammatical asymmetries exist between languages as regards the possible number of values of the grammatical features, such as gender or number, that words encode. For example, the uncountable words *sheep* or *fish* (uninflectable in number) might correspond to plural or singular nouns in other languages, e.g. Sp. *oveja* vs. *ovejas* or *pez* vs. *peces*. On the other hand, as far as gender is concerned, such nouns in English as *student* or other languages such as Turkish *orangi* 'student' might have more than their equivalent in the target

[1] All the examples used are taken (or, sometimes, adapted) from real texts, preferably from the *New York Times* but also from other media online.

languages depending on the context, e.g. German *Student* (m) ∼ *Studentin* (f).[2] In this regard, my expectation was that, when this gender asymmetry affects referential personal nouns or any word which must agree with one, an ambiguity is created which will have to be resolved by machine translation engines so as to prevent the quality of the target text from being down graded.

The first stage of my research was, then, to explore that asymmetry, both comparing how two given languages (or types of languages) encode gender and predicting in what way that asymmetry may be a problem for machine translation. As far as possible, real translation examples were obtained from freely accessible MT engines online. This exploration is necessary to reduce the scope of the problem, that is, to see what exactly is what translation engines do wrong and must be worked on.

1.1 Gender asymmetries between languages

The following are some typical examples of the asymmetry we have just exposed (and the problem that it poses for machine translation). Let us first consider the case I have just referred to, in which an English noun can correspond to either a feminine or a masculine noun in the target language. For example, the ex. (4) shows a sentence in English and two possible translations, either of which would be correct if the source sentence were translated in isolation of any other context: the noun *teacher* and the title *Dr.* are gender-indefinite so they could be translated either by feminine or masculine Catalan equivalents.

(4) a. Dr. Jones is a good teacher.
b. **El Dr.** Jones és **un bon professor**.
c. **La Dra.** Jones és **una bona professora**.

However, if we use a NP with a specific gender (such as *Ms.*, *mother* or *she*) to refer to Dr. Jones instead, one of the two proposed translations cannot be correct. Consider examples (5) and (6):

(5) a. She is a good teacher.
b. Ella és ___*un bon professor___.
c. Ella és **una bona professora**.

(6) a. Dr. Jones is a good mother.
b. ___*El Dr.___ Jones és una bona mare.

[2] This might not always be the case even when translating into some gender-rich languages, e.g. Portuguese *estudante* (m/f), but the gender duality will be expressed by other means, such as the phrase dependents, e.g. ***o** estudante* ∼ ***a** estudante*.

c. **La Dra.** Jones és una bona mare.

Even in cases in which it is considered that the same term must be used in the target language regardless of the gender of the person referred to, such as *modelo*, *testigo* or even, for certain people, *juez* or *obispo*,[3] finding out whether it refers to a male or a female might be essential to produce a correct translation.

(7) a. Do you think that Judge Rosenda (f) Sarmiento acted well?
b. ¿Cree usted que **_*el_ juez** Rosenda Sarmiento actuó bien?
c. ¿Cree usted que **la juez** Rosenda Sarmiento actuó bien?

Not only English nouns can have equivalents of both genders in the target language. Indeed, other words, such as the pronoun *one*, are gender-indefinite in English but might refer to a female or to a male. As the referent's gender of "one" in ex. (8) is not known, both translations provided can be considered correct.

(8) a. Most primary school teachers quit, and I know one who was sent to prison.
b. La majoria dels professors[4] d'escola primària abandonen, i en conec **un/una** que va ser enviat a la presó.

In the following examples, however, only a translation with a female-specific equivalent for *one* would be correct:

(9) a. Most primary school female teachers quit, and I know one who was sent to prison.
b. La majoria de les professores d'escola primària abandonen, i en conec **_*un_** que va ser enviat a la presó.

(10) a. Some women are designers, writers, accountants, teachers, housewives, secretaries, cinematographers. I know one who's a landscape architect, one who runs an art gallery.
b. Algunes dones són dissenyadores, escriptores, comptables, professores, ames de casa, secretàries, cinematògrafes. En conec **_*un_** que és un arquitecte paisatgista, **_*un_** que porta una galeria d'art.

[3] The Royal Spanish Academic (or RAE)'s *Diccionario de la lengua española* (DRAE Real Academia Española 2001), which is allegedly the most authoritative dictionary of Spanish, does not include the feminine form in the entry of *obispo* and many people would consider the word *obispa* incorrect and object to its usage, in spite of the fact that the RAE itself says (Real Academia Española 2005) that those personal nouns whose masculine form ends in *-o* should normally form the feminine replacing this vowel with an *-a*. In the case of *juez* ∼ *jueza*, for example, both are admitted as correct.

[4] We will only consider singular nouns for the moment.

On the other hand, while the third person singular personal pronoun in English is always a cue as regards the gender of the gender-indefinite noun in question, as we saw in the ex. (5), it is not in other languages, such as Turkish, Finnish, Farsi, Chinese or Quichua (Karlsson 1999), in which there exists only one 3rd person singular pronoun to refer to an individual (i.e. there is no gender distinction). Indeed, the pronouns *hän*, *o* and او /ū/ refer to either a female or a male. The pronoun will therefore pose a translation problem, too, if their referent's gender is not know. See examples (11-13):

(11) Hän sai kandidaatintutkinnon. (Finnish)
PRON.3RD.SG got the bachelor's degree
'**She/He** got the bachelor's degree.'

(12) O orangi. (Turkish)
PRON.3RD.SG student
'**She/He** is a student (f/m).'

(13) او جغرافيا مى خواند. [5] (Farsi)
ū joghrafia mīkhūānæd
PRON.3RD.SG geography studies
'**She/He** studies geography.'

NPs are not the only source of gender asymmetry between two different languages. For example, the form of other lexical categories such as adjectives might also be invariable in one language but gender-dependent in another, but we will not go into them because they will always depend on the NP's head's gender. Similarly, in some languagaes, such as Arabic, verbs need to be inflected in accordance with the gender of, at least, one NP (generally in subject position). I will just give one example of each case for illustration. In ex. (14), the adjectival phrase headed by past participle *born* stands in an appositional copular relation to the subject of the main clause, so in some languages the equivalent of the former has to agree to agree with the equivalent of the latter. In ex. (15a), the name "Robinson" poses no problem of translation but if it is not translated as feminine name, the form of the agreeing verb risks being the masculine one يقود in the target text, instead of the correct feminine one تقود, in agreement with the named entity referring to the first female President of Ireland Mary Robinson.

(14) Born in 1930 in Tokyo, she was married to the actor Noboru Nakaya from 1954 until their divorce in 1978.

[5] I follow the standard DIN 31635 for the transliteration of the Arabic alphabet.

(15) a. Robinson (…)[6] lidera un proyecto internacional. (Spanish)

b. روبنسون يقود مشروعا دوليا. * (Arabic)
rūbinsūn *yaqūdu mušrūʿan dawlīyyan
Robinson leads-MASC project international

So far, we have considered the singular number only. Plural nouns, on the other hand, might not need express their referent's gender if they refer to a mixed group in which there are both males and females (instead, their gender might be determined by other factors, such as the language in question, social conventions, individual usage, etc.). When all the members in the group share the same gender, however, they might be referred to by a gender-specific plural noun and, depending on the language, any item inflectable in gender (verbs, adjectives, etc.) which refers to them will reflect their gender.

In the following example, the fact that both Dr. Lisa Gibbs and Dr. Laura Mosqueda are women implies that the equivalent of the verb "established" in a target language with gender-inflection in verbs (e.g. Arabic) should be in the feminine as to agree with its plural subject.

(16) But Dr. Lisa Gibbs said the wife had not intended harm. (...) Dr. Laura Mosqueda and Dr. Gibbs established that mental impairment, depression and isolation had made all three men vulnerable.

It is interesting to see how the form of the verb depends on the subject-referent's gender in some languages. In English, neither do the plural personal pronoun *they* nor verbs inflect in gender in accordance with the subject-referent's, but in some Slavic languages they do. For example, knowing that "they"'s antecedent is "women" or that "they"'s referent's gender is feminine does have a relevance when translating into some Slavic languages such as Czech or Polish (unlike in Russian, Belorrusian or Bulgarian). In the real example (17a), the referent of "they" is a colective entity composed of only women, whereas in the adapted example of (18a), the referent of "they" is a collective entity composed of only men. We can see that the Czech and the Polish verbs, *stát se* / *stávat se* (perfective and imperfective aspect) and *stawać się*, respectively, reflect the referent's gender. As you can see, misinterpreting or overlooking the referent's gender in the source sentence might lead to a mistranslation in Czech and Polish. This would also be the case in other tenses such as the future or the present for Czech.

[6] Full sentence: "Además de la primera mujer que presidió la República de Irlanda, Robinson fue Alta Comisionada de las Naciones Unidas para los Derechos Humanos y en la actualidad lidera un proyecto internacional en defensa de la ética de la globalización y el desarrollo sostenible."

(17) a. They (f) have become agents of modernization.[7]

b. **Staly** se nositelkami modernizace. (Czech)
became.FEM REFL agents.FEM modernization

c. **Stały** się nośnikami modernizacji. (Polish)
became.FEM REFL agents.FEM modernization

(18) a. They (m) have become agents of modernization.

b. **Stali** se nositeli modernizace. (Czech)
became.MASC REFL agents.MASC modernization

c. **Stali** się nośnikami modernizacji. (Polish)
became.MASC REFL agents.MASC modernization

1.2 Gender asymmetries and machine translation

Let us see now some real examples of how well machine translation engines performs when faced with the asymmetry I have just briefly described. I give the translations carried out by two engines (Google Translate[8] and Translendium[9]), always with English as a source language and French as a target language. I will use examples in which some element shows the referent's gender so as to make evident the anomaly (or absence thereof) in the target sentences, starting with syntactically simple examples and adding complexity progressively.

1.2.1 Nouns

In a source sentence in which there is a gender-specific element in the left side of the copula, the translation of an NP in the predicative complement by both translation engines is right regardless of whether the verb is *be* or another copulative verb and regardless of whether the predicate is depictive (e.g. *seem* or *remain*) or resultative (e.g. *become*), what indicate that they must have a rule that generalize the subject's grammatical gender to the predicative complement. See ex. (19-20).

(19) a. She is/seems/remains/becomes a good treasurer.

b. Elle est/semble/reste/devient **une bonne trésorière**. (Gt/Tl)

(20) a. My sister is a good treasurer.

b. Ma sœur est **une bonne trésorière**. (Gt/Tl)

[7] The unabridged sentence is "Not only are women influenced by modernity, as a highly educated professional group, they REFL have become significant agents of change and modernization."

[8] http://www.google.com/translate_t

[9] http://www.translendium.com

However, if we reverse the positions of subject and predicative complement (e.g. transforming the ascriptive copular clause in ex. (20) into a specifying one) we see that only one of the engines, Google Translate, still generalizes the grammatical gender of the NP in the predicative complement to the NP in the subject function. See example (21).

(21) a. The treasurer is my sister.
b. **La trésorière** est ma sœur. (Gt)
c. *****Le trésorier*** est ma mœur. (Tl)

When the predicative is oblique, as with the complements of an *as*-phrase, again Translendium seems not to take into account the gender information associated with the subject-predicand, leading to a gender-biased translation. See exs. (22-23):

(22) a. She works as a treasurer.
b. Elle travaille en tant que **trésorière**. (Gt)
c. Elle travaille comme *****trésorier***. (Tl)

(23) a. She served as treasurer.
b. Elle a servi de **trésorière**. (Gt)
c. Elle servait comme *****trésorier***. (Tl)

We have seem examples in which the gender-specific element is in subject or predicative position of intransitive clauses. If we consider now transitive clauses with the gender-specific element as object, we can see that now it is Translendium which handles it well (ex. (24)), whereas Google Translate produces a gender-biased translation (ex. (25)).

(24) a. I consider her a good driver.
b. Je la considère *****un bon trésorier***. (Gt)
c. Je la considère **une bonne trésorière**. (Tl)

(25) a. They chose her as treasurer.
b. Ils l'ont choisie comme *****trésorier***. (Gt)
c. Ils la choisissaient comme **trésorière**. (Tl)

On the other hand, we see that Translendium takes into account the information provided by a gender-specific element in the left side of the copula as long as it is not a personal proper name. See example (26):

(26) a. Linda Miller is a treasurer.
b. Linda Miller est *****un trésorier***. (Gt/Tl)

We have seen that both engines might, depending on the context, exploit the gender information provided by the cue, but only provided the gender-indefinite element and the cue are within the same clause. interestingly, if they are not, as in the example (27), the result is a gender-biased translation:

(27) a. The treasurer was so angry she called me to her office.
b. ****Le trésorier*** était si fâché elle m'a appelé à son bureau. (Gt)
c. ****Le trésorier*** était si fâché qu'elle m'appelait à son bureau. (Tl)

1.2.2 Adjectives

If we consider now predicative complements with the form of an AdjP instead of an NP, we will see that none of the engines considered experiment any problem to inflect the adjective in the correct gender, as can be observed in examples (28-29):

(28) a. She, in this context, is no longer relevant.
b. Elle, dans ce contexte, n'est plus **appropriée**. (Gt)
c. Elle n'est plus, dans ce contexte, **pertinente**. (Tl)

(29) a. She seems/remains/becomes happy.
b. Elle semble/reste/devient **heureuse**. (Gt/Tl)

In other cases, such as when the predicative is an adjunct rather than a complement (optional instead of obligatory), none of the engines seems to have a rule to generalize the gender feature from the subject to the predicative adjunct.

(30) a. She arrives tired.
b. Elle arrive ****fatigué***. (Gt/Tl)

On the other hand, we will not consider other kind of predicatives, such as a locative complement in the form of a PP ("She put it on the table") because we do not know of any language in which such PP must agree with the subject or the object.

1.2.3 *One* pronoun

The performance of both translation engines is similar as regards the translation of the pronoun *one*, and as with NPs the results seem to depend just on the syntactic complexity of the sentence we feed to the engines (the translation is gender-biased as soon as we go beyond the boundaries of the clause).

(31) a. She is one in a million.

b. Elle a **une** ans dans million. (Gt)

c. Elle est **une** en un million. (Tl)

(32) a. As one who has been a Republican in a heavily Democratic state, she has always been independent.

b. Comme ****un*** qui a été ****un Républicain*** dans un état lourdement Démocratique, elle a toujours été indépendante. (Gt)

c. En tant qu'****un*** qui a été ****un républicain*** dans un état fortement démocratique, elle a toujours été indépendante. (Tl)

1.2.4 Verbs

We have seen that, in some languages, verbs have gender inflection, generally in agreement with the subject NP. Therefore, in some languages it is not only for the sake of it being rendered in the appropriate gender-specific form in the target language that the NP must be correctly interpreted in the source language, but also because there are other elements which need to be made to agree with that NP. For example, in Arabic (as a target language) a 2nd or 3rd person singular subject and the verb need to be made to agree in gender,[10] which might be a problem if the source verb is gender-indefinite. When translating the example (33a), the system does not take into account that the human named entity "Ms. Walter" is feminine, and the verb "argues" is translated into the masculine form "يقول" in the target text, instead of the correct feminine one "تقول" in agreement with the subject.

(33) a. Ms. Walker argues that the time is ripe for elevators.

b. السيدة ووكر يقول ان الوقت حان لمصاعد. * (Gt/Arabic)
al-sayyida wuwkir ****yaqūlu*** ʾan al-waqt hāna limaṣāʿid
Mrs. Walker argues-MASC that the-time came for-elevators
(Arabic)

A similar thing happens with the example (34), in which the feminine given name *Jeannette* is not exploited to infer the referential gender of "vice president" and the verb "says", and thus they are translated into the masculine:

(34) a. Jeannette Horan, a president, says she thinks she knows why.

b. يقول جنّتّ هوران ، رئيس ، هو يفكّر هو يعرف لماذا (Gt/Arabic)
****yaqūlu*** jeannette horan, raʾisun, huwa yufakkiru
says-MASC Jeannette Horan, a president-MASC, he thinks-MASC
huwa yaʿrifu limāḏa
he knows-MASC why

[10] #ARABIC GRAMMAR

1.2.5 Other parts of speech

In other cases, agreement might be necessary between the the NP headed by a personal noun and other elements, such as the possessive pronoun *whose*. See what would happen when translating from English into Quichua, if the gender of the NP in the subject is not accounted for:

(35) a. (...)[11] said Cherifa (f) Kheddar, whose brother and sister were killed by Islamic extremists in 1996.

b. (...) Cherifa Kheddar nin, **_*turanwan_** **_*ñañanwan_**
(...) Cherifa Kheddar said, brother-him-with sister-him-with
ancha chiqnikuq islamistakuna 1996 watapi
very xenophobous islamists 1996 year-in
wañuchirqanku. (Quichua)
assassinate-3P.PL.PAST.ACTIVE
'(...) Cherifa (f) Kheddar said, whose (m) brother and whose (m) sister the xenophobous islamists killed in 1996.'

1.3 Excluded cases

1.3.1 First and second persons

Regarding person, I chose to deal only with third person (considering all nominal references to extralinguistic human entities as third person). There are several reasons for this, the most important of which is that gender inflection or gender contrast is rare in the first and second persons, as it is a characteristic mainly of the third person. Siewierska (2004) argues that out of 133 languages in her study, 129 (97%) have gender in the third person as opposed to only 24 (18%) in the second and 3 (3%) in the first. Therefore, it is worth devoting efforts primarily to interpreting third person gender to the detriment of the other persons because it is the most frequent case.

Another reason, as far as the second person is concerned is that deciding the gender of the equivalent in the target language is not a NLP problem because the MT engine might never know whether the reader is a woman or a man, unless a webcam or a user query is used (except in dialogues, quoted text or texts written only for women or for men or in text addressed to one only and known person, which is not the general case in press reports or informative texts in the Web). It is rather an issue of politically correct language policy which should be solved by other means.

[11] Full sentence: '"We've reached a dangerous point when the criminals are out of prison and the people who don't agree with it are arrested," said Cherifa Kheddar, whose brother and sister were killed by Islamic extremists in 1996.'

1.3.2 Non-referential contexts

See the example (36) of translation from English into Spanish and Catalan, in which "the user" is not a referring expression but a generic reference and the anaphor is double because the potential user might be either female or male.

(36) a. A user may only remove his or her own print jobs, but not those of other users.

b. Un usuari només pot treure pròpies feines d'impressió | marca **_*seves o seves_**, però no aquells d'uns altres usuaris. (Tl)

This problem happens the other way round, too. In the example (37), the inclusive language used in Spanish is not correctly rendered into English:

(37) a. Los trabajadores y trabajadoras no estamos dispuestos a seguir recibiendo este trato.

b. The **_*workers and workers_** we are not arranged to continue receiving this treatment. (Gt)

We see that this kind of gender-inclusive language in non-referring contexts must be taken into account in machine translation, but this is different from finding out whether a noun refers to a male or a woman.

Chapter 2

Conceptual framework

In this chapter we will try to explain and interrelate the concepts involved in the research and which we will subsequently rely upon. Gender has to do both with relations between the text and the world, because the gender of a human referent is reflected in language, and with relations between several parts within the text itself, because gender is expressed by means of agreement relations. On the other hand, the knowledge we need to resort to can also come either from the text itself or from the extra-linguistic world. Hence, we draw mainly on the areas of gender studies and text linguistics, although some concepts from the field of translation theory will be introduced too and some reflection will be made about reference.

2.1 Bibliographical philias <= include in the introduction of this chapter, with no separate section

# FILIAS: (Halliday & Hasan 1976) is by far the most comprehensive monography on the topic of cohesion and is in fact a stardard reference. Halliday and Hasan blabla

2.2 Language and the world

It is important to devote some thought to what reference[1]is and what is not reference because, as will presented further below, only a referring expression points to a particular individual in the world and only under these conditions it makes sense to be concerned with the linguistic gender of an expression. It must, hence, be clear when an expression is referential or not.

2.2.1 Reference

We use language, among many other things, to communicate about the world (by 'world' meaning whatever a speaker wants to talk about, be it tangible or cognitive). The relation between expressions of the language and the world is what we call **reference**. For Huddleston & Pullum (2002: 399), reference is the relationship between, on the one hand, a linguistic expression (normally a NP) which a speaker utters on a particular occasion with the intention of picking out a independently distinguishable entity in the world and, on the other hand, that entity which the expression picks out or stands for on a given occasion of its utterance.[2] See also Lyons (1977: 174) and Moore (1993).

It is important to bear in mind that, as argued by Lyons (1977: 176), reference is utterance-dependent or context-dependent: such sentence as "the expression x refers to ..." must be understood as "by means of uttering x the speaker is referring to ...", because, as we will see in XX, some personal nouns seem not to have a gender class in abstract but will always have a covert linguistic gender according to the agreement relations that they establish when they are used in a real text.

Let us now define the key terms that I will be using regarding reference. A referential or **referring expression** is, then, an expression uttered in a particular speech act which picks out a particular (i.e. independently distinguishable) entityin the world cognitively accessible to the speaker[3] that she or he wants to communicate about. On the other hand, the entity in question is what we will call the **referent** of the expression. And finally **reference** is this relationship between the referring expression and the referent, in which the former points to the latter (by pointing meaning the fact that it is the writer/speaker's intention to make the reader/hearer think of that entity when she uses that expression) (Huddleston & Pullum 2002: 399; Lyons 1977: 177).

Although we will see this in more depth below (XX), we can add here that

[1] It must be warned here that there are two senses for the term 'reference'. On the one hand, by reference some mean the symbolic relationship that an exophoric expression has with its referent in the world or the situational context (as we will see below). On the other hand, by reference some (Halliday & Hasan 1976: § 2) mean the endophoric relationship of one linguistic expression to another, in which one provides the information necessary to interpret the other (see section 2.4). We will call this latter concept *coreference*, following Brown & Yule (1983).

[2] An utterance is the product of a speech act and reference must therefore be understood as related to a particular speech act, that is, context-dependent.

[3] I will use the terms 'speaker' and 'writer' or 'hearer' and 'reader' interchangeably, so the fact that I use one (say, 'speaker'/'hearer') doesn't exclude that what is valid for it might also be valid for the other ('writer'/'reader') taking part in verbal communication.

when two referential expressions have the same referent in the world they are called **corefering** or **coreferential**.

2.2.2 Non-referring expressions

While referential expressions account for a good part of all the NPs that must be considered in our research, there are other expressions which are considered not referential in their context of use but to which attention must be devoted too when they are headed by a personal noun. In general, we could say that an uttered expression is referential only when it is possible to use the existential quantifier ($\exists\ x$) when expressing it in predicate logic. Consider the following examples: in (38) the noun *a doctor* refers, whereas in (39) it does not:[4] Obviously, the gender of the personal noun in the non-referring NP must not be interpreted because it refers to no particular person in the world; the automatic distinction of non-referential expressions such as this can be problematic, however.

(38) **A doctor** and two chemists have patented what they say is a simple and inexpensive test to detect oral cancer.

(39) How often have you left **a doctor**'s office wondering just what you were told about your health (...) ?

On the other hand, there is another kind of non-referential expressions headed by a personal noun whose gender must be interpreted, despite being traditionally considered non-referential (according to the canonical notion of reference we have seen): for example, indefinite NPs[5] in predicative complements (PCs) (either intransitive or transitive).[6] PCs characteristically express a property and therefore usually have the form of an AdjP, although the same meaning can be conveyed by means of a NP (cfr. *She is* *clever* vs. *She is* *a clever person*).

This takes us to the distinction established by logicians between names and predicates (understood in a logical sense): names select entities in the world by naming them and predicates describe or attribute properties or qualities to those entities (Nelson 1992: 3, 8). For instance, in the copular clause in (40) the

[4] See Huddleston & Pullum (2002: 399-410) for many other examples of what are deemed to be non-referring expressions.

[5] Here we focus on indefinite NPs only as problematic cases because definite NPs are normally referential expressions (except in their generic use).

[6] Mitkov (2002: 6), for example, argues that an "indefinite predicate nominal" cannot be considered as having the same referent as the subject-name because it is not specific enough. Huddleston & Pullum (2002: 252), too, considers that predicative complements with the form of an NP are non-referential.

referring expression "Ayman al-Zawahri" is a name which refers to the person Ayman al-Zawahri and "an Egyptian" is not a name but a predicate with an ascriptive use[7] which attributes the property of being Egyptian (but is not canonically considered to refer) to the subject-referent.

(40) Ayman al-Zawahri is an Egyptian.

The similar example (41) shows a complex-intransitive clause with the predicative complement "a mourner", which applies[8] to the subject-referent "Ms. Martin".

(41) Ms. Martin, appearing stricken and dressed in black, looked like a mourner at a funeral having trouble keeping her composure.

Other examples with occupational terms: e.g. in *She served as treasurer, They chose her as secretary*, the predicates (the underlined NPs) all apply to the predicands (the subject-referent or the object-referent) but they do not refer to them.

Although an indefinite NP in a PC and a NP in a subject or object position cannot corefer *stricto sensu*, they must grammatically agree (see gender agreement #X) and therefore the gender of the personal noun heading such a PC should be interpreted. For the practical purposes of disposing of a clear designation for the expressions we deal with, we will enlarge the notional extension of what is understood by reference so as to also consider as referring expressions those NPs in attributive sense which apply and assign some property to the subject or object referent (even though they do not denote any particular individual).[9] Thus, α will be considered to refer to i (and corefer with any other expression which refers to i) if there exist an individual i and a property p expressed by a nominal expression α used attributively such that i is p.

There are some precedents for this convenient notional licence: see, for example, the criteria adopted in the Message Understanding Conference (MUC) coreference task (Hirschman & Chinchor 1997), for whose aims coreference is allowed to be recorded when the predicate nominal is marked indefinite.

[7] "In the ascriptive use, PC denotes a property and characteristically has the form of an AdjP or a non-referential NP; the subject is most often referential and the clause ascribes the property to the subject-referent. (...) The specifying use defines a variable and specifies its value: 'The victim was his sister'" (Huddleston & Pullum 2002: 266).

[8] We use the term *application* in the sense of (Nelson 1992: 18) to designate the relationship between the predicate and the name in the predicand: the expression "an Egyptian" *applies* to the referent of the name "Ayman al-Zawahri".

[9] Whether one considers this kind of expressions coreferential or not should not have important methodological consequences for the purposes of our research (the difference would be simply a terminological one), as we will deal, all the same, not only with expressions which are bound to agree with the gender of the person referred to, regardless of whether they are normally considered referential or not.

2.2.3 Person reference

Reference can be made to any particular entity, as we said, but it is reference to human beings what we care about here. We are not interested in first- and second-person reference, because they are phrased almost exclusively in the form of pronouns and are mainly used to refer deictically to the participants of the situational context (the speaker or the addressee). Nor do we care about collective entities, for reasons that we have seen (#WHERE). We will hence focus on third-person singular reference, which can be made by a wide range of nominal expressions (namely descriptions headed by personal nouns, personal names and third-person personal pronouns), possibly deictically to a participant in the situational context but mainly to another person in the world (in written texts). Third-person person reference can be expressed by descriptions, personal names and pronouns.

2.2.3.1 Descriptions

Person-referring (in)definite descriptions are ordinary non-pronominal NPs whose head is normally a personal noun.[10]

2.2.3.2 Personal proper names

Proper names are capitalised[11] expressions with the syntactic status of a definite NP (although they can also function as attributive modifiers or heads of larger NPs) which the members of a speech community use by convention to refer to a particular entity, such as a person, an animal, a place, an organization, an event, etc. (Huddleston & Pullum 2002: § 20) and which, therefore, may potentially refer to any entity that hold that name. For example, the personal name *Ms. Rice* refers to a particular individual which, given the appropriate context, might easily be interpreted to be the first African American woman to serve as US Secretary of State, and fictional proper names such as *Idomeneo* refer not to a particular person in the world but to an imaginary character who exists in an opera by Mozart and in the minds of the people who know it. See # for a full review of personal names.

We will only deal with the proper names used for people (anthroponyms), which we will call **personal names**, in the US press, namely because that is the nature of the corpus all examples and texts are taken from for the thesis project (i.e. *The New York Times*). Personal names uniquely identify a person within a group and both differentiate her from, and relate her to, other people from the

[10] Personal nouns are equivalent to what (Irmen 2007) or (Huddleston & Pullum 2002) call 'role names' or 'bare role NPs', respectively, e.g. *wife*, *friend* or *patient*.

[11] I consider written discourse only.

same community. By using a personal name as a referring expression, then, a writer expects her adressee "to construct a representation of that referent" (or retrieve an existing one) (Mulkern 1996: 239). In general, as we will see, most personal names are gender-specific, but their recognition is not always straightforward because they seldom have a univocal gender mark, as well as due to their structural diversity and, therefore, unlike pronouns or other nouns which encode gender lexically, the gender of personal names is not always easily inferred.

The task of automatized gender interpretation needs to rely on all sorts of gender expression in texts: personal names can be one of these means, and that is why it is so important to understand how named reference is made to individuals in written discourse: i.e. what kind of constituents personal names are made of, as well as what structural patterns are used to combine them according to the name holder's cultural background (people from any cultural background in the world can be referred to in the US media, thus increasing the complexity of the task). On the other hand, some name constituents (i.e. given name, patronyms, titles) can be gender-specific and serve therefore as cues. On the other hand, some constituents (i.e. typically the family name) can be used instead of the full name on second and subsequent references and therefore serve to establish named entity chains. However, these key constituents' position within the personal name is variable according to the naming tradition.

Constituents of personal names Personal names might have a simple head (many Indonesians have only one name, as Gen. Suharto) but will most typically, especially on first appearance, consist of a composite head, consisting normally of at least one given name (plus one or more middle names) plus a family name, although given (and middle) names might be collapsed to an initial letter or omitted altogether, especially in formal written press. They can still have some elements which are sometimes considered adjacent (not forming part of the name) but that we will consider as pre-head or post-head modifiers within the personal name NP, such as titles, embellishments, etc.

A **given name**[12] is a name that is accorded to, and assumed by, a person at birth and is used to refer to that person individually and identify her or him within a group or family. As opposed to a family name, it is generally not inherited. In Europe and North America, where the given name precedes the family name, given names are called first names or forenames (also Christian

[12] The designations *first name* or *forename* and *last name* or *surname* will be avoided at the expense of *given name* and *family name*, respectively, because of the different positions they can occupy to designate individuals with different cultural backgrounds. By first and last names reference will be made to the position within the personal name, not to the quality of the name.

names). On the other hand, a **family name** (also called surname or last name) is the part of a person's name that indicates to what family he or she belongs. It is passed from one generation to the next and serves as a means both to distinguish among several individuals with the same given name and to group together members of the same family. Also, in many cultures a woman adopts her husband's family name when they are married (Campbell 2007).

Given and family names can be used as constituents of a full personal name but they can also be used on their own, depending on pragmatic factors (e.g. the genre of the written discourse in question,[13] the referent's status, etc.), and provided that the speaker expects the addressee to already have in memory a representation of the person referred to by that name (for the application of the 'givenness hierarchy' to personal names, see Mulkern 1996). That is, full names can be used appropriately to refer to an individual under any discoursive circumstances, although they will be used most frequently to refer to an allegedly new human entity, whereas single names (a given name or a family name alone) are normally used for familiar or already given referents only (McShane et al. 2005), as Grice (1975)'s maxim of quantity predicts.

Although the basic pattern of personal names is given plus family name, in some cultures there are variations to this: another name (or its initial letter which stands for it)[14] is sometimes inserted between the given and the family name, and is therefore called **middle name**. In some cultures, for example among the East Slavs, the Greeks or the Arabs, the middle name is a **patronym** or patronymic (i.e., a name derived from the given name of the father or another paternal ancestor which might indicate descent, e.g. by means of an affix (Campbell 2007; Miller 2007)).

Titles or **appellations** are elements of personal names which could be analyzed as pre-head modifiers. They express the condition of the individual referred to. There are several kinds of titles (Huddleston & Pullum 2002: 519), some of which indicate the gender of the individual referred to. They include: the default set of courtesy titles (*Mr./Mr, Ms./Ms, Mrs./Mrs, Miss, Master*); titles of royal/aristocratic office or rank (*King* ∼ *Queen, Emperor* ∼ *Empress, Archduke* ∼ *Archduchess, Her Majesty, Prince* ∼ *Princess, Earl, Lord* ∼ *Lady, Count, Sir* ∼ *Madam*, etc.); clerical office (*Pope, Archbishop, Sister* etc.); military and police rank (*Private, Captain, Squadron Leader, Admiral, Inspector*, etc.); political office (*President, Senator/Sen., Governor/Gov., Councillor, Representative/Rep., Republican, Conservative*, etc.); judicial office (*Judge, Justice, Solicitor, Barrister*, etc.);

[13] The journalistic genre seems to influence the choice of name constituent to refer to a person, e.g. in sports news it is more common to refer to, say, football players by their given name.

[14] Sometimes, it is the given name that is replaced by its initial, such as in Germany or sometimes the UK.

academic status (*Doctor/Dr./Dr, Professor/Prof.*, etc.). On the other hand, there are some post-head modifiers: adjectives such as *Junior/Jr.* or *Senior/Sr.* as in *Edward B. MacMahon Jr.*and apposited fused modifer-head NPs, such as ordinals or other labels for monarchs and popes: *Peter the Great, George the Fifth, Pope John Paul II.*

On the other hand, it is important for us to take into account another set of pre-head[15] elements that Huddleston & Pullum (2002: 519-520) call **embellishments**, because they are frequently target nouns which must be interpreted. They are very similar to titles and sometimes the only apparent difference might be the capitalized initial of titles, but, unlike titles, embellishments's motivation for appearing adjacent to the personal name is more discoursive/informative than social, as can be observed in the underlined modifiers in *the commanding Dutch avant-gardist Reinbert de Leeuw, the pianist Oleg Malov, the baritone Sergei Leiferkus,the radical imam Hassan Mustafa Osama Nasr*or *the syndicated columnist Robert Novak,*which are nominal modifiers but not too different from adjectives such as *the 22-year-old Andrei Ryabov,the late David Wax,the statuesque Natalya Trokai*or *the dryly intense Mr. Malov*.

The structure of personal names Personal names can have different internal structure according to what naming tradition the name's bearer comes from. The pattern used in Western countries is very wide-spread, but there are still some cultures which follow other naming practices, which might be maintained in English texts. The main practices are the Western, the Eastern (except Vietnamese) and the Vietnamese (which is slightly deviant from the Eastern rule). We have done a shallow exploration and aim at giving a brief description of the different syntax that personal names might be found to have in English media according to a number of factors.

The basic pattern for full personal names for **Western** people referred to in anglophone media, as have been seen, is given name (followed by an optional middle name) plus the family name. Therefore, the given name is always the first name and the family name is always the last name. This is the standard

[15] Huddleston & Pullum (2002: 520) distinguish between constructions consisting of an omissible nominal dependent (an embellishment) plus a personal name head, such as *architect Normal Foster*, and definite constructions consisting of the article plus a personal noun as head plus a proper name as an omissible appositive dependent, such as *the architect Norman Foster*. In the examples provided (see below), the proper name would not be a composite head but an apposition, according to their view. For practical purposes I will reserve the term 'apposition' for NPs or nominals embedded within commas and will consider personal nouns (together with the article) in this position as removable pre-head modifiers which form part as 'embellishments' of the NP which has the personal name as head (or will consider the head as a weak proper name), regardless of whether the NP is determined or not.

practice in Western countries but by the 20th century, it had already been imposed on many other places in the world, such as parts of Africa or among North American Indians. Therefore, many non-Western cultures do not diverge significantly from Western countries in the internal structure of personal names. Check out the Western order in two references to an Ethiopian entrepreneur in ex. (42):

(42) "I've always loved Starbucks, the ambiance of it," said Tseday **Asrat**, the proprietor of Kaldi's [a coffee roasting company] (...) "Coffee is part of every Ethiopian's life," Ms. **Asrat** said. "We discuss life over coffee."

Some cultures, such as the Arabs, the Slavs and others, follow that basic pattern but they might use one or more middle names or patronyms. Thus, traditional Arabic names[16] include a given name (*ism*), a patronym (*nasab*) and a family name (or *nisba*, which might begin with the article *al-*/*el-*) (Anderson 2007: 302), but in some cases, as for legal or very formal purposes, an individual may have more than one middle name so as reflect the genealogy on the father's side: i.e. the name of the father, the grandfather, the great-grandfather, etc. (Omar 1975: 247-250; Campbell 2007). The important point here, however, is that if the given name is used, it would appear at the beginning of the name and the family name will be the last constituent in the name, regardless of the number of middle names in between.

In other naming systems, including those in most of **East Asia** as well as in Hungary, the given names traditionally come after the family name (Dari 2006), although the native practice might not always be adopted when people from these cultures are referred to in the Anglo-Saxon media. For example, in Hungary it is common practice to write one's family name first, followed by one's given name (van Nijmegen 2002), but most often the order of Hungarian names is normally switched when they write or are written about in English in the foreign media (some press style guides such as Goldstein (2000) and Jordan (1976) recommend to follow the person's naming preference). See the example (43), in which the person Laszlo **Solyom** is referred to by given + family name on first appearance but by title + family name on second appearance:

(43) Those tensions were evident in Budapest today as the Hungarian president, Laszlo **Solyom**, welcomed Mr. Bush to a gilded chamber in the Sandor Palace. (...) In brief remarks, Mr. **Solyom** said Hungary's commitment to democracy is coupled with a respect for human rights — a possible reference to Guantánamo.

[16] Arabic or Muslim names are names used in the Arabic-speaking and Muslim countries, which closely coincide, although there are some Muslim names which are not Arabic (i.e. Iranian or Turkish).

The Chinese name system is the original pattern of names in Eastern Asia, namely, the one that all communities in Eastern Asia have followed. Today, the number of Chinese family names actually used is not higher than one thousand, but less than hundred cover most of Chinese people.[17] The variety in Chinese names therefore depends greatly on given names rather than family names. Similarly to Hungarian, Chinese have their family names precede their given names (e.g. ***Mao*** *Tse-tung*) but Chinese names used in Western countries may be rearranged when written to avoid misunderstanding. Some names remain in the traditional order even in English media, though (Dari 2006; Goldstein 2000: 142, 42).

Regarding other countries, the Chinese practice of putting family name first is followed in most East Asian names (e.g. Korean or Laotian), even when writing in English (Jordan 1976: 111).[18] For instance, Korean former President **Kim** Il Sung was President **Kim**. See how two Korean people are referred to in first and second appearance in the example (44)). In the past, however, some Koreans, like Syngman **Rhee**, westernized their names but this practice seems to be unproductive nowadays. Vietnamese names follow the same order but are treated differently on second or subsequent reference and will be described apart (see below).

(44) The case of **Lee** Chun-hak, the 19-year-old who fled the North on June 28, is a typical one. (...) [H]is mother, **Kim** Myung-shim, 46, visited him from Seoul the other day. Mrs. **Kim** fled to South Korea in 2003, remarried and began working to arrange the defection of Mr. **Lee** (...).

On the other hand, two trends might be found in Japan: Japanese people generally have their names in Western order but, if born before the Meiji period, they can be referred to in Japanese order, see ex. (45).

(45) The title of "world's oldest person" is now apparently passed to Yone **Minagawa** of Japan, who (...) turned 114 this month. (...) [But] Ms. **Minagawa** may not hold the crown for long.

In other cases, the full names might be used in all references, such as for Cambodian or Burmese names. Because of their different ethnic backgrounds, Cambodians' names assume a varied range of forms and, with a view to consistency and certainty, full names are recommended in all references (Jordan

[17] "Statistics argue that there are about Chinese 5,600 surnames (...), of which about 1,000 are most frequently used. (…) The top ten surnames are used by about 40% of Chinese. (…) So in total, more than 70% of the Chinese population uses the same 45 surnames." (Jeff 2003).

[18] Unlike Chinese practice, the Korean given names (e.g. *Il Sung*) are not hyphenated in our corpus. This is, nevertheless, a rule prescribed for articles from the New York Times but which might not apply to Korean names appearing in other media.

1976: 32). See for instance ex. (46). In headlines, however, the surname can be used on its own if it is certain that it is a surname or if it is used alone by Cambodians themselves, e.g. King (or Prince, when he was it) Norodom Sihanouk, see ex. (47).

(46) Prime Minister Hun Sen has hardly run Cambodia as a democracy. (…) Seth Mydans reported in The International Herald Tribune recently that Mr. Hun Sen is waging political war against human rights groups and political opponents, largely through misuse of defamation laws.

(47) The Senate approved a bill early today in which it signaled a willingness to allow the sending of American arms to Prince Norodom **Sihanouk**'s guerrilla army in its fight for power in Cambodia. (…) The Senate vote will give a real boost to **Sihanouk** and will be very helpful in facilitating a successful outcome to the talks between him and Hun Sen.

A **Vietnamese name** usually consists of the family name followed by two given names. In the name ***Vo** Van Kiet*, for example, **Vo** is the family name, *Van* is the middle name and *Kiet* is the given name. However the normal practice to refer to a Vietnamese, if their whole name is not used, differs from the pattern in other East Asian cultures in that one the last part of the name (the second given name) is used as if it were a surname, preceded by the appropriate title.[19] Thus, ***Vo** Van Kiet*, for example, would subsequently be referred to as *Mr. Kiet* (Jordan 1976: 218; Geotravel 2007, quoted in ThingsAsian.com 2007). See examples (48) and (49):

(48) The Prime Minister of Vietnam, **Vo** Van Kiet, wants to restrict soft-drink and beer imports.(…) All new joint ventures in beverages will have to be cleared by the Prime Minister to encourage domestic investment and local production, the daily said, quoting a statement from Mr. Kiet's office.

(49) "The meaning of the group," said **Nguyen** Thi Sau, 29, whose husband has already died from AIDS complications, "is so that when you die you are less lonely." (…) "Some days I have to take care of four people who have died in the hospital," said Ms. Sau, who worked at a shoe factory until she was fired.

On the other hand, somewhere in the middle between the standard Chinese-based Eastern and the Vietnamese practices is the **Thai name practice**. The Thai have names normally composed of a given name followed by a surname, but in

[19] An exception was Ho Chi Minh, who was and is commonly referred to in Chinese fashion as *Ho*.

second or later references the given name is used with the honorific, according to the guidelines (Jordan 1976: 206). Thus *Thaksin **Shinawatra*** is *Mr. Thaksin*, as can be appreciated in ex. (50):

(50) The election will take place in a significantly changed political context after the banning of the party of former Prime Minister Thaksin **Shinawatra**, who was ousted in the coup and remains in self-exile in London. (…) Mr. Thaksin faces an arrest warrant on corruption charges if he returns home.

Personal pronouns Personal pronouns are context-dependent pro-forms with the syntactic status of an NP which stand for and/or refer to one (Huddleston & Pullum 2002: 1461). Third-person pronouns in particular are typically endophoric,[20] as they normally, but not exclusively, refer to another expression in the text (their antecedent), with which they normally always corefer. The English personal 3rd-person singular pronouns are gender-specific: i.e. *she* and *he*, with all their declinated or reflexive derivatives. We must also mention the 3rd-person gender-indefinite pronoun *one* (in red in the Table 2.1), which is frequently used non-referentially as indefinite (or, as Halliday & Hasan (1976: 38, § 2.3.3.2) calls it, 'generalized person' but which can also be used as a referring expression, either as the only constituent or as the head of an NP (see examples (56-64) in §§ 2.4.3, p. 44).

Let us recall the classification of personal pronouns established by Halliday & Hasan (1976) in Table 2.1 (in blue the elements we will mainly deal with).

	Speech roles		**Other roles**		
	Speaker	Addressee	Specific		Generalized Human
			Human	Non-human	
one	*I me* *mine my*	*you you* *yours your*	*he him* *his his* *she her* *hers her*	*it it* [*its*] *its*	*one*
more than one	*we us* *ours our*		*they them* *theirs their*		

Table 2.1: Reference items, from Halliday & Hasan (1976: 44)

[20] We prefer to talk about 'endophora' rather than 'anaphora' so as not to exclude cataphora, which is much rarer but nevertheless possible.

2.3 Language and gender

Given the relative terminological dispersion in the field of language and gender, we must delimit the domain of our concerns and define the concepts we handle in relation to nominal classification and the social categorization of extra-linguistic referents. Moreover, the term *gender* is used in many different contexts, yet within linguistics, to refer to substantially different, though somehow related, things, so some clarification is needed. This chapter draws heavily on Corbett (1991)'s excellent monography about gender and especially on Hellinger & Bußmann (2001)'s critical article.

2.3.1 Types of gender

For the purposes of our research, it is important to distinguish, then, as well as relate to each other, the key concepts of referential gender and linguistic gender. Social gender must be devoted some thought too. These three kinds of gender can be thought of as referential, denotative and connotative, respectively.

Referential (or referent's) gender This is the extra-linguistic gender, that is, the constructed dichotomous trait of being (perceived as), and / or behaving primarily like, a female or a male[21] that a person identifies with and / or that has been or is attributed to them within the society they are part of, although some authors (Curzan 2003) consider referential gender only in terms of what is traditionally regarded as "biological sex."[22]

Linguistic gender Linguistic gender (also called grammatical gender), according to Lyons (1968: § 7.3.3), Corbett (1991), Curzan (2003: 12), Foley & van Valin (1984: 325) or Ibrahim (1973: 50), is a system of noun classification that, regardless of whether or not the noun varies formally according to its gender class, is manifested compulsorily by regular and systematic morphosyntactic means, i.e some gender-variable words such as adjectives, determiners, pronouns, etc. related syntactically or discoursively to a noun experiment changes

[21] It is not always clear what referential gender people not conforming to the dichotomy female / male (Stryker 2004) #Hall & Buchotz (1995)??) might have but this issue has only a marginal interest for the topic of this dissertation, so we will not go into it here.

[22] Often, biological sex matches the socially constructed gender. However, this might not always be the case (e.g. transgender individuals), neither is always clear what a person's sex is (e.g. intersex people). Therefore, unlike Frank et al. (2004), for whom "sex –a quality of the world– coincides with gender –a grammatical category–, when talking about humans", when I use the word (referential) *gender* (of a person) I will be referring to an individual's gender identity or gender role and not to their genitalia-defined sex (see Kessler & McKenna 1978).

in their form according to the gender (referentially, semantically or grammatically) associated with the noun they are related to. Linguistic gender is, therefore, reflected in the behaviour of related words and it is this dependent behaviour what allows linguistic gender to be expressed. Contrarily to stereotypical gender, it is conveyed in the denotational meaning of gender-specific nouns.

Stereotypical gender Stereotypical gender (also called social gender) is the connotation associated with certain words. It has to do with "stereotypical assumptions about what are the appropriate roles for women and men", including expectations about who can be a member of a particular class of people (Hellinger & Bußmann 2001: 11). In other words, stereotypical gender refers to the fact that an addressee might think of a referent of a particular gender according to the addressee's stereotypes and expectations even though the noun used to pick out the referent is gender-indefinite. This is not a neutral phenomenon, indeed an androcentric bias underlies the majority of personal nouns, so in most cases (but not always, cfr. *nurse*) the speakers of most languages will think of a male human referent if the referential gender is not specified (Silveira 1980; Hamilton 1991, referred in Stahlberg et al. 2007; Braun 2001).[23] # RELATE TO MT PROBLEM. # CITE experiments with Finnish

It must be borne in mind that the whole theoretical point of our proposal of gender interpretation is that there is a correspondence between linguistic gender and referential (extra-linguistic) gender as far as referring expressions headed by personal nouns are concerned. This might not be an obvious fact in English if one adopts a perspective focused on the lexicon because English nouns generally have a covert gender (Corbett 1991: 62-63), but a discourse-based view allows one to see that gender-specific pronouns show gender inflection according to their antecedent noun's referential gender (see 2.3.4). A personal noun's referential gender (or sex of the referent, as other authors prefer to call it) is, therefore, an indicator of that noun's gender class in English, and there is indeed a correspondence between the two.

The most outstanding authors on this domain agree on this view. Corbett (1991), for example, affirms that sometimes noun classification corresponds to what he calls biological distinctions of sex, and Curzan (2003: 11-12, 21) admits that there is "a clear correlation between masculine and feminine nouns and biological traits in the referent" in the natural gender system of Modern English.

[23] Perhaps as a result of an underlying gender hierarchy, "masculine/male expressions (…) are the default choice for human reference in almost any context" (Hellinger & Bußmann 2001: 19).

As for grammatical-gender languages,[24] Curzan (2003) argues that grammatical gender is not always arbitrary because "there often is a correlation between grammatical gender and biological sex for nouns describing human beings".[25]

2.3.2 Noun classification systems

Different authors have different approaches at organizing the world languages' systems of classifying nouns. For example, Corbett (1991) establishes a distinction between semantic gender systems and formal gender systems. In the former, the noun's meaning (i.e. features of the referent, although not necessarily sex features) determines its gender class, with no needed reference to its form, whereas in the latter systems, the gender assignment of many nouns is determined by morphological or phonological factors (although there always remains a semantic core to the system, especially regarding personal nouns). Modern English would be a gender language with a nearly perfect[26] semantic gender system, as the classification of nouns mostly corresponds to extra-linguistic distinctions of male, female and non-human/inanimate.

It is interesting to also point out here that Corbett (1991: 63) asserts that formal gender assignment systems are never exclusively formal: "[pure formal] system[s are] not found in any natural language: gender always has a basis in semantics." Indeed, even in languages with a formal system, such as, say, Spanish, in which the majority of gender assignments rely on morphological or phonological factors and in which there is no apparent or known semantic motivation for the gender of inanimate objects, there is a semantic core to the system, especially regarding personal nouns. Hence, the fact that the noun *maestra* has a feminine gender and the noun *maestro* has a masculine gender is not arbitrary at all, but is referentially motivated (as long as they are used in referring expressions).

On the other hand, Hellinger & Bußmann (2001: 5-6) establish a distinction between noun class languages and gender languages, based on whether there is a correspondence between nouns' linguistic gender and their referents' "biological sex", as well as other secondary elements: noun class lan-

[24] On the difference between natural gender systems and grammatical gender systems, see section 2.3.2.

[25] There are cases, however, in which lexical (or morphological) gender clashes with referential gender, as in the case of Sp. *zorrón* (m) 'slut', which always refers to a woman or Gl./Pt. *sentinela* (f) 'sentry', which might likely refer to a man. Correfering expressions from outside the NP might, however, most likely agree with the referential gender and not with the lexical/morphological gender of the noun (cfr. "Menudo zorrón es ella" or "Como uma sentinela ele era melhor que qualquer de os cachorros").

[26] Except for the inanimate nouns that can take gendered pronouns and the human nouns that can take *it*.

guages, such as Dravidian and New Guinean languages, show no obvious correspondence between class membership and a personal noun's specification as female-specific or male-specific, whereas in gender languages, such as most in the Indo-European and Semitic families, there is a correspondence between the specification of personal nouns and the referential gender.[27]

Craig (1994)

Those classical approaches are interesting but still we need a different perspective, which we will take from the organization adopted by Stahlberg et al. (2007) or Curzan (2003). This perspective relies on whether and/or to what extent referential gender (or biological sex, as they call it) is reflected in grammatical structures of a language. Indeed, there is a range of different forms of representation of gender, from its encoding on every nominal, verbal and sometimes other phrases to it absence altogether. The important point here is that, on the one hand, referential gender is represented somehow in virtually all languages of the world but, on the other hand, personal nouns are not gender-specific in all languages. We will distinguish three levels of representation (summarized in Table 2.2 below):

Grammatical gender languages Every noun is assigned feminine or masculine (and in some cases also neuter) grammatical gender. The grammatical gender of nouns denoting/referring to inanimate objects might be arbitrary from a referential point of view, but there is a correspondence between the grammatical gender of virtually all personal nouns and the gender of their referents. Semitic and most Indo-European languages, for example, belong to this group.[28]

Natural gender languages In the so-called natural gender languages there tends to be no grammatical encoding of referential gender in nouns. Any personal noun can therefore be used to refer either to a female or to a male and, therefore, we can consider that personal nouns in this language type are **dual-gender** or **gender-indefinite**[29] but we could also consider that they are invari-

[27] The category of noun class languages seems to correspond roughly to Corbett's semantic gender system, although semantics can also play a role in grammatical gender systems.

[28] Although Stahlberg et al. (2007) seem not to, we will include in this type noun class languages in which female-specific and male-specific personal nouns belong to different classes, such as, say, Tsakhur or Dyirbal. On the other hand, noun classes in which female-specific and male-specific personal nouns belong to the same class could be equated with the genderless language type (see below).

[29] Other authors might use the term *epicene* for this, but we prefer to avoid it because it is normally used for nouns with the same grammatical form (which might be of either gender) to refer both to males and females in languages where nouns are gender-specific, e.g. use *el ser humano* (m) instead of *el hombre* in Spanish to refer generically to humanity.

able **dual-gender** nouns which are realized in real usage in texts as belonging to a particular gender class according to the referential conditions.[30] Personal pronouns, on the other hand, do match their antecedent noun's referential gender, varying accordingly, and are the main device for expressing gender. English is the paradigmatic member of this group[31] (Scandinavian languages could also be included here).

Genderless languages In these languages, neither do personal nouns have grammatical gender nor are personal pronouns sensitive to femaleness or maleness, but there are other means to express gender, such as lexical gender (kinship terms, for example). All the same, referential gender is reflected in language very seldom, compared to the other language types. Not many languages belong to this group, some well-known examples being Turkish, Finnish, Farsi, Quichua or Chinese.[32]

	Grammatical gender languages	**Natural gender languages**	**Genderless languages**
Frequency	High	Middle	Low
Necessity	Often	Sometimes	Rare
Linguistic forms	Lexical, pronominal, grammatical	Lexical, pronominal	Lexical

Table 2.2: Expression of referential gender in different language types

Table 2.2 (taken from Stahlberg et al. 2007: 166) summarizes the comparison among these three language types. We see that all languages can express

[30] We could draw an analogy with how a particular meaning of individual lexical items (which out of context have just a potential for meaning) is realized through association with other lexical units in a particular context of communication.

[31] Hellinger & Bußmann (2001: 6) consider that English lost its gender system and hence belongs to the group of languages which do not exhibit a gender system or nominal classification whatsoever, allegedly because these authors take into account only nouns, which in English have in their majority no formal gender marking. Contrarily to this view, we think that English has a quite eroded gender system because (apart from the fact that some morphological means of marking gender can still be found in English personal nouns), if one takes Corbett (1991: 147-150)'s view that gender classes are determined by (or must be considered to match) agreement classes, and if one considers not only personal nouns but the language system as a whole, it must be concluded that English is a gender language. Although it might be thought that a given personal noun does not belong to any particular grammatical gender *per se* out of context, it is true that in real usage such a noun, if referring, will agree with other co-referring elements in the text, which will show which grammatical gender class the noun in question has in that particular speech act.

[32] We include in this language type noun class languages which assign all personal nouns to the same gender class, regardless of whether they denote/refer to a female or a male, as translation into them would not reproduce the problem we present here: e.g. North American Indian languages, which have two genders (animate and inanimate) or Swahili, which has 11 noun clases, one of which (I: *m*-) is for animate beings, including both female and male humans (Lyons 1968: § 7.3.5).

gender by some means or another, at least through general terms with lexical gender such as *woman*, *father*, etc. This is the only means genderless languages use. Natural languages use both that and gendered pronouns and finally grammatical gender languages use those two means as well as grammatical gender for all nouns. There is, therefore, a relation among the three language types which we could express as *GramGenLang* > *NatGenLang* > *GenLessLang*, where '>' means 'with significantly more gender-specific personal nouns than'.

The problem object of our research could be materialized between two languages belonging to different types out of these three as long as the direction of the translation is from a source language with comparatively more gender-specific personal nouns to a target language with comparatively less gender-specific personal nouns. For instance, from a natural gender language to a grammatical gender language or from a genderless language to any of the previous two (as shown by Figure 2.1). Obviously, the less means of expressing gender the source language has, the more difficult it is to interpret the gender of personal nouns.

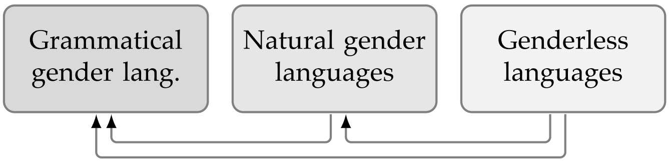

Figure 2.1: Direction of possible gender-biased translation.

English fits particularly well the purposes of our research as a source language, for in English, as a natural gender language, personal nouns have a covert gender, but elements such as pronouns, proper personal names etc. have overt gender (i.e. they do express their antecedent noun's referential gender).

2.3.3 Expressing gender in English

If we focus on English now, then, it is worth pointing out at what levels (or through which categories) this language (particularly the US variant) can and does (or does not) express gender (i.e. can relate to the extra-linguistic category of referential gender) when referring to individuals.

2.3.3.1 Lexical gender

Most personal nouns in English are **gender-indefinite** or **gender-neutral**, e.g. *surgeon*, *analyst*, *judge* or *student*, but there is a number of non-inflected nouns which are **gender-specific** (i.e. lexically specified as carrying the semantic property [+FEMALE] or [+MALE] and, therefore, female-specific or male-specific) and which are used to refer to either a female or a male, but not to a person whose gender is not known and who could actually be either. Examples: most kinship terms (*mother* ∼ *father*, *niece* ∼ *nephew*, *daughter* ∼ *son* etc.[33]), basic or general terms (*boy* ∼ *girl*, *woman* ∼ *man*, *guy* ∼ *gal*, *Sir* ∼ *Madam*, *husband* ∼ *wife*, *lady* ∼ *gentleman* etc.), sex- or sexuality-related terms (*lesbian*,[34] etc.), some occupational terms (e.g. *nun* ∼ *monk*, *king* ∼ *queen*, *soprano*, *mezzo-soprano*, *alto / contralto* vs. *sopranist*,[35] *countertenor*, *tenor*, *baritone*, *bass-baritone*, *bass* etc.), etc.

Women's generalized access to the labour market has broken down the gender segregation that existed in many professions and therefore a new regard on the stereotypical gender of some occupational terms is necessary. For example, *nurse* is today a term which might refer either to a man or a woman. A more extreme example: now that women can become clerical leaders in some religious communities around the world (not without strong debates, though), the clergy nouns *rabbi*, *priest*, *vicar*, *pastor*, *bishop*, *deacon*, *imam*, *mullah*, *hafith* (and even other religious-related nouns such as *mujahideen* or *jihadist*) cannot be considered male-specific anymore. However, it cannot be ignored that people holding one of those occupations are most likely males, which sometimes leads writers to specify the referential gender of a personal noun by means of a *wo/man* or *fe/male* compound when the referential gender contradicts the stereotypical gender, see example 51:

(51) The group had delayed ending the fast until late today as a courtesy to (...) a female rabbi.[36]

2.3.3.2 Morphological gender

While, as we just said, most personal nouns in English are **gender-indefinite**, it cannot be ignored that English possesses an increasingly low[37] productive

[33] Some kinship terms are indefinite, though: *cousin*, *sibling*, *parent*, *spouse* etc.

[34] The noun *gay* would be expected to be gender-indefinite and, in theory at least, it can refer to either men and women but in fact it is mainly used to refer to men, whereas the noun *lesbian* refers exclusively to women. As an adjective, *gay* is gender-indefinite and may normally apply to women too: "She's gay". It is important to be sure to what extent a noun might refer to any of the sexes so as to know whether that noun must be a target of gender interpretation or a cue (see Irmen 2007: 168). # INCLUDE IN THE TEXT (AND NOT AS A FOOTNOTE)? #HOST-HOSTESS

[35] A sopranist (or sopranista or sopranite) is a male classical singer with a voice-type and register equivalent to that of a female soprano.

[36] Source: "Arab Women End Their Fast" In *New York Times* August 5, 1982.

[37] Gender-specific occupational titles are vanishing from AmE (Bailey 2004): for example, *server* is increasingly used instead of *waitress* and *waiter*. *Actress* is still used, but Americans would not find

pattern of gender inflection and compounding for nouns, which are marked with(out) suffixation or compounded with the formants *-woman* or *-man* according to the referent's gender, i.e. *actor* ∼ *actress*, *blond* ∼ *blonde*, *brunet* ∼ *brunette*,[38] *widow* ∼ *widower*,[39] or *congresswoman* ∼ *congressman*.

Some considerations must be made regarding gender-marking morphology. While all the *-ess* marked nouns can be safely assumed as apply to females, the same cannot be said of their unmarked counterparts. On the one hand, nouns such as *abbot*, *count*, *duke*, *emperor* and *prince* are restricted to males and nouns such as *abbess*, *countess*, *duchess*, *empress* and *princess* are restricted to females. On the other hand, however, while professional/occupational nouns such as *actress*, *authoress*, *manageress*, *poetess*, *sculptress* or *stewardess* are restricted to females (but rarely used), their (much more frequent) unmarked counterparts can apply either to female or male individuals: an authoress, for example, is an author, a manageress is a manager, etc. This asymmetry is reinforced by the feminist-compelled trend to avoid sexual bias and discriminatory terms.[40] For more details, see Huddleston & Pullum (2002: Ch. 19, § 5.3).

There are other, less used feminine marking suffixes, such as *-ette*, *-ine* and *-(tr)ix*, which are somewhat obsolete now, as all the nouns they can form have gender-neutral counterparts which are preferred in emotively neutral contexts (i.e. those feminine markers have a pejorative connotation). Some exceptions which might still be used nowadays are *majorette*, *usherette* and *heroine*, although it is not uncommon to find *hero* referring to a woman too (see ex. 52).

(52) Ms. Schumaker denies that she was a <u>hero</u>.

On the other hand, suffixes such as *-ine* or *-ette* are still found today in a number of feminine proper names, e.g. *Bernardette*, *Bernadine*, *Pauline*, *Clementine*, etc.

it strange to hear that "Meryl Streep is a wonderful actor." *Stewardess* has been replaced by *flight attendant*, and so on and so forth. See below for more comments on gender-marking morphology.

[38] As happens with the pair *gay* ∼ *lesbian*, there is a asymmetry in the usage of these nouns describing a person's hair colour (the same does not apply to the homographic adjective, though). Although the *Merriam-Webster Online Dictionary* establishes the spellings *blond*/*brunet* when used of a male and "usually" *blonde*/*brunette* when used of a female, the fact is that those nouns seem to be relegated almost exclusively to describing females. Although it would be theoretically possible to read "He's a blond," it would be much more likely to see "He has blond hair" or no reference to his hair colour at all. On the other hand, the masculine spellings *blond* and *brunet* can also be found as referring to females. This implies that *blonde* and *brunette* must be treated as female-specific nouns but that *blond* and *brunet* should not be blindly taken for granted as male-specific. # INCLUDE IN THE TEXT (AND NOT AS A FOOTNOTE)?

[39] This is one of the few examples in which the inflected, marked form corresponds to the masculine.

[40] Some exceptions might perhaps be the case of *waiter* ∼ *waitress*, as the former is still predominantly used for males and therefore the latter is not easily avoidable, and the pair *actor* vs *actress*, whose gender differentiation is motivated by the different roles played by males and females.

2.3.3.3 Pronominalized gender

When the referential gender is known and no generics are needed, third-person singular pronouns (used referentially) are always gender-specific in English (the speaker must choose among *she* or *he*) and therefore reveal or confirm the referential gender of possible co-occurring gender-indefinite or gender-specific personal nouns, respectively.

2.3.3.4 Anthroponymical gender

So far we have seen what constituents a personal name might be composed of and how they are combined according to the different traditions in the world and the (lack of) adaptation that is made in the US media. Let us now turn to the elements in a personal name which are of primary interest to us, namely those elements which can be, directly or indirectly, perceived as gender-specific and can therefore be exploited as cues for a gender interpretation process.

Given names Given names are frequently gender-specific (some exceptions are unisex names *Sam, Leslie, Gael, Dominique, Alison, Chris, Robin, Claude, Jean, Joan*[41] etc.). The recognition of their gender specificity need, however, some source of external knowledge or a probabilistic estimation, because they do not generally have any reliable formal mark of being male-specific or female-specific. Japanese given names, however, typically bear one ending or another depending on whether the person named is a male or a female (Abe 2007): male typical suffixes include *-aki, -fumi, -go, -haru, -hei, -hiko, -hisa, -hide, -hiro, -ji, -kazu, -ki, -ma, -masa, -michi, -mitsu, -nari, -nobu, -nori, -o, -rou, -shi, -shige, -suke, -ta, -taka, -to, -toshi, -tomo, -ya, -zou*, etc. and female typical suffixes include *-a, -chi, -e, -ho, -i, -ka, -ki, -ko, -mi, -na, -no, -o, -ri, -sa, -ya, -yo*, etc.

Family names and patronyms Family names, on the other hand, are normally gender-indefinite in many languages (for example, in Anglo-Saxon or Iberian cultures a brother and a sister will have exactly the same family name). However, in some languages, such as Slavic languages or Greek, not only given names but also family and middle names are gender-specific (i.e. they express the gender of the name's holder).

For example, in a Bulgarian female name such as *Maria Stefanova Spassova*, the patronym *Stefanova* indicates that her father's given name is *Stefan* and the family name *Spassova* indicates that her father's family name is *Spassov* — hence, the suffix *-ova* is female-specific and indicates a female referent. If she had a

[41] It can be seen that the last two names are male-specific in francophone and catalanophone cultures, respectively, whereas they are female-specific in anglophone cultures.

brother called *Ivan*, his full name would be *Ivan Stefanov Spassov*. There are other name suffixes apart from *-ov*/*-ova* which have gender inflection, such as *-ski*/*-ska*: e.g. *Grozdarski* (m) ∼ *Grozdarska* (f). Russian names follow the same rule (*-ov*/*-ovna*: if the father's name is, say, *Ivan Krylov*, then the son's name, for example, will be *Pyotr Ivanovich Krylov*, and the daughter's name will be, say, *Varvara Ivanovna Krylovna* (Zgusta 2007).[42] The feminine inflected patronym is not always used, however, being reserved for women who have "an independent reputation under such a name," according to Jordan (1976: 181).

Turning to Greek family names, a woman's family name, e.g. Pavlidou, is the genitive singular of a related man's name (either her father's or her husband's if she chooses not to keep their own family name after marriage) e.g. Pavlidis (**?**). # IN THE ENGLISH PRESS... SEE NOTE 10

In Greek, patronyms mark gender, as well as case and number. Although there are gender-indefinite suffixes such as *-oglou* (Turkish for 'son/daughter': thus *Spyroglou*, son of *Spiros*[43]), some other suffixes are gender-specific,[44] such as *-poulos*/*-poulou*, *-atos*/*-atou*, *-as*/*-a* or *-is*/*-i*. See Table 2.3 for some examples. When referred to in anglophone media, the original inflection of the female's patronym might possibly be respected (e.g. Ms. Argyriou, Ms. Chatziantoniou and Ms. Gorouor Nana Mouskouri) but this is not always the case (e.g. Ms. Contopoulosor Christina Onassis). This feature should, therefore, cautiously be taken as a trace or indication, and not as an absolute evidence, of the referent's gender.

Suffix	**Male name**	**Female name**
-poulos/*-poulou*	*Marios Papadopoulos*	*Maria Papadopoulou*
-atos/*-atou*	*Apostolos Mourelatos*	*Aggeliki Mourelatou*
-as/*-a*	*Thomas Samaras*	*Alexandra Samara*
-is/*-i*	*Paulos Zisis*	*Anastasia Zisi*

Table 2.3: Greek gender-specific patronyms

On the other hand, we saw that Arabic names will frequently include a patronym, which can be preceded by the particles *bin*, *ben* or *ibn* 'son of' for males and *bint* 'daughter of' for females (Hedden 2007). Matronymic names (names derived from the mother or female ancestor), on the other hand, are also possible but rare. Teknonyms are another peculiarity worth mentioning: parents are often non-officially referred to as *Abu* 'father' and *Umm* 'mother' of their eldest son (or dauther if they have no sons), e.g. "Abu Hasan."[45] Arabic

[42] Check for example the Russian tennis players Dinara Safina and Marat Safin, who are siblings.

[43] These words are called *digenés monokatáliktes* 'two genders, 1 ending'.

[44] Aka *digenés dikatáliktes*.

[45] This pattern can also be used as a nickname (*laqab*) describing no parent-child relationship,

patronyms, matronyms and teknonyms are referred to collectively as the *kunya*, which is, as can be seen, indicative of the gender of the person concerned, not very differently from titles of courtesy or inflected gender-specific patronyms.

Titles Finally, the most reliable elements in a personal name, as far as the ascertainment of gender is concerned, are some kind of titles, especially courtesy titles, which apart from pragmatic status-related aspects also convey gender. Courtesy titles (such as *Ms./Mrs.* ∼ *Mr.* for English or *Daw/Mah* ∼ *U/Maung/Ko* for Burmese people)[46] are today generalized to the whole population who do not possess another title, especially in the written media, and, together with honorific titles indicating aristocracy or nobility, and to a lesser extent also eclesiastical or clerical titles, are the ones which tend to be always gender-specific, as can be seen in the Table 2.4. On the other hand, titles indicating profession or professional rang are generally gender-indefinite.

Let us sum up and sketch the general trends that English language presents regarding the expression of gender. On the one hand, the great majority of personal nouns do not mark gender (except for the ones stated above and others alike) whereas all third-person singular pronouns[47] do mark it. On the other hand, most personal names can also encode gender, either lexically or morphologically, but some are ambiguous. Check out Table 2.5 for a more structured synthesis:

2.3.4 Gender agreement

Once we have seen what gender is about, what kinds of gender there are and how gender can be expressed, we must underline that, in fact, it is not gender in itself we must be concerned with, but **gender agreement**, which is the way in which gender is realized in language use.

The standard definition of agreement is that it is a "systematic covariance between a semantic or formal property [in our case, the gender class] of one element and a formal property of another" (Steele 1978: 610), quoted in (Corbett 2006: 4). Another interesting view is that agreement has to do with the 'displacement' of grammatical meaning (Moravcsik 1988: 90): one word carries the grammatical meaning relevant to another. This means that the formal feature of one element (the controlled element, or satellite)[48] varies systematically ac-

e.g. the Palestinian President Mahmoud Abbas is sometimes referred to as "Abu Mazen".

[46] English titles can also be used for Burmese people.

[47] To the exception of *one*, see 2.4.3.

[48] Agreement satellites are also called targets but to avoid confusion we reserve this designation for the gender-indefinite personal noun we must interpret.

Title type	**Male-specific**	**Female-specific**	**Gender-indefinite**
Courtesy	*Mr., Master, Al-Sayid* *U, Maung, Ko*	*Mrs., Ms., Miss* *Daw, Mah*	
Academic			*Dr., Professor, Dean*
Kunya	*Abu, Bin, Ibn*	*Umm, Bint*	
Aristocratic	*Sir, King, Emperor, Archduke, Prince, Earl, Lord, Count, Emir, Sheik*	*Madam, Queen, Empress, Princess, Her Majesty, Lady, Sheikhah*	
Clerical	*Pope, Father,* *His Holiness*	*Mother, Sister*	*The Rev., Archbishop, Cardinal, Bishop*
Army/police			*Private, Capt., Admiral, Inspector, Lt., Maj., Sheriff, Gen., Col., Gov., Specialist*
Political			*President, Sen., Rep., Councillor, Gov., Republican, Conservative, Ambassador, Minister, Speaker*
Judicial			*Judge, Justice, Solicitor, Barrister Prosecutor, Mayor, Attorney*

Table 2.4: Gender-specificity of titles

Type of reference	**Gender-specific**	**Dual-gender**
Personal nouns	Some.	Most.
Personal names	Most. All names preceded by a title of courtesy.	Some.
Personal pronouns	All (except *one* and the interrogatives *who*, *whose* and *whom*).	Only *one* etc.

Table 2.5: Gender expression in person reference in English

cording to the semantic feature of another element (the controlling element), regardless of whether the controlling element varies formally too according to its own semantic feature. Differently phrased, the controller might have (but not necessarily) overt expression of that feature, e.g. gender, whereas the satellite has bound expression of agreement, i.e. obligatory marking — this relationship is therefore assymetric, in as much as the feature of the satellite depends on that of the controller noun).

On the other hand, we might add that the controller is always a noun (see Corbett 2006; 1991: 105), whereas the satellites can be any of a wide range of categories (determiners, modifiers, verbs, pronouns, etc.). In addition, it is worth noticing that the feature that motivates the agreement might not be semantic, as it happens with gender in English: the lexicosemantic specification of many personal nouns does not include any particular linguistic gender, but they might be gender-specific in a real context according to their referential gender. Reference can, in fact, override semantics (remember examples *zorrón* and *sentinela*, seen above). Agreement occurs, therefore, between expressions which corefer (see 2.2.2 and 2.4.2).

An alternative view, albeit not very different from the one just stated but more suited for our purposes and framework (referential personal nouns), is that agreement occurs between (1) the gender of the individual referred to and (2) the realization of all the expressions which refer to her or him, which are correferential and (3) agree among themselves too. Some of the expressions might express that linguistic gender overtly (satellites do) but others might not by themselves (for example, dual-gender personal nouns in English).

Figure 2.2 represents the triadic relationship among the referential gender, the controller and the satellite. The referent (i)'s gender determines the linguistic gender of the expressions that refer to her/him, and that fact is represented by an arrow from the former to the latter. The arrows from the referent's gender to the controller and the satellite are dashed and thick, respectively, to show that the controller noun's gender is covert and that the satellite noun's gender is overt. The bidirectional arrow ⇔ represents the agreement between controller and satellite and the blue arrow from the satellite to the controller represents what we do: project the satellite's overt gender to the controller. The controller and satellite's subindex shows the referent and their superindex shows the linguistic gender ($[\emptyset]$ = covert gender).

Potential satellites, as exposed in Corbett (1991: 106-115) according to their category, are: adjectives, demonstratives, articles, numerals, possessives, participles, verbs, pronouns, adverbs, adpositions, complementizers and, of course, other nouns. In English only nouns, possessives and pronouns can. Here, I prefer to organize those elements according to the functional level where

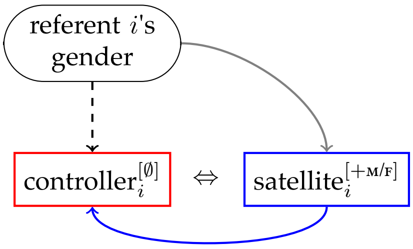

Figure 2.2: Relation between referential gender and agreement.

they can be found (emphasis is made on agreement in English:[49]

Within the noun phrase: the two elements of the agreement relationship are included in the NP. A typical example is dependent-head agreement (e.g. *eine sehr gute Anwältin* 'a (f) very good (f) lawyer (f)'). In English this level of agreement is only possible when coordination occurs within the NP, i.e. when an NP conjoins a controller and a satellite, e.g. *She and her brother*).

Within the clause: the controller is the subject or an object and the satellite an agreeing element in another phrase within the same clause. For example: subject-predicative agreement (e.g. Ca. *Joan és alt* 'Joan [m] is tall (m)' (#see example "they themselves have become significant agents of change and modernization",) but Eng. *Joan* [f] *is a good mother*); object-predicative (e.g. Gl. *Considero-a umha mui boa candidata* 'I consider her a very good candidate (f)' but Eng. **I consider him a very good actress*); verb-subject agreement (e.g. Ar. تتكلم '(she) talks (f)' vs. يتكلم '(he) talks (m)'), not present in English; subject-object (pronoun) agreement (e.g. *Lesley visited her friends*). See Corbett (1991: 106, 126) for examples in other languages.

Within the text: NP-NP agreement. This is the agreement that can occur out of the boundaries of the clause between two NP's (one of which might not necessarily be a pronoun). Example: *L'ex-candidate à la présidentielle a précisé lundi que les problèmes du couple qu'elle formait avec François Hollande* (m) *datent d'avant la présidentielle. Le patron* (m) *du PS assure que la séparation n'a «ni cause ni conséquence politiques».*

To sum up: we have seen that, in the case of human referents, there can be a correspondence between the referential gender of the person in question and the

[49] Some authors, such as Lyons (1968: 283), however, consider that in English "there is no gender-concord."

grammatical gender of the referring expression, shown (if shown!) either by inflection or by agreement. In the case of languages such as English, most nouns are gender-indefinite and never show gender inflection but gender agreement does happen between their referential gender and other coreferent NP present somewhere in the text, either within the NP, within the clause or somewhere else in the text. The form of satellites is not controlled by the noun's grammatical or lexical gender (as there is none) but by the noun's referential gender (i.e. the gender of the person that noun refers to).

2.4 Language and the text

Let us now turn to cohesion and comment on the relationship that exists between it and gender agreement. On the one hand, some cohesion devices agree with the gender class of the controller noun, if coreference is involved, or hold some sort of gender-specific semantic relationship, in the case of associative anaphora — they can, therefore, serve as cues for the gender interpretation. On the other hand, as gender agreement can be a cohesive constraint (an anaphor and its antecedent, if coreferential, must agree in gender), cohesion can be broken down in the target language if there is a gender mismatch between a personal noun and other coreferential expressions as a result of a gender-fair translation.

2.4.1 Lexical cohesion

According to Halliday & Hasan (1976: 1-2), 'texture' is what makes a passage be a text and it derives from the fact that the passage in question functions as a unity with respect to its environment. These authors explain by what means texture is created, namely, what linguistic features are present in a passage containing more than one sentence that contribute to the passage's total unity, allowing us to perceive it as a text. They describe, among other things, several semantic-discoursive relations between different parts of a passage, especially those in which one element is interpreted (or its interpretation is reinforced and/or confirmed) by reference to other expressions in the co-text. Therefore, these relations make us think that those elements form part of one and the same self-sufficient unit, i.e. they give **cohesion** to the text. Of course there might be more than two elements which maintain cohesive relations with each other, forming a **cohesive chain** or network (Halliday & Hasan 1976; Baker 1992: 180; Mitkov 2002: 4).

Summing up, for us cohesion occurs when the interpretation of some element in the discourse depends on that of another and both elements are interpretable and the reader interprets all the parts of the text as effectively belonging to the same whole.[50]

2.4.2 Endophora and coreference

# reversed agreement

As we saw, Halliday & Hasan (1976) distinguish between **endophoric** and **exophoric** reference, but only the former has cohesive force and is, indeed, one of the main devices of cohesion. Examples of endophora are anaphora, i.e. cohesion between an element (the **anaphor**) and a previous item (the **antecedent**) to which it points back,[51] and cataphora, i.e. anticipated reference to a element mentioned subsquently in the text, much less common than anaphora (Halliday & Hasan 1976; Mitkov 2002: § 1.10) Let us remind that here we will focus on cohesion among nominal expressions. On the other hand, looking at cohesion from the point of view of exophoric reference (cfr. 2.2.1), when the members of a cohesion chain have the same referent in the world we say they are **coreferential** and that form a **coreferential chain**, which is also a source of cohesion.

As we see, the two main devices of cohesion are endophora and coreference (which are not mutually exclusive). Two or more elements linked by an endophoric relationship might also be coreferential, in which case they are said to have **identity of reference** (e.g. in ex. 34, repeated here for convenience, "Jeannette Horan" and "she" are said to be linked by an identity-of-reference anaphora.

(53) Jeannette Horan, a president, says she thinks she knows why.

However, this is not always necessarily the case and thus abound the cases in which there is an anaphoric relationship between two expressions which do not corefer, either because they have different referents, as in associative anaphora (see § 2.4.3), identity-of-sense anaphora (see below) or because they are not referring expressions (Mitkov 2002: 6 and § 1.6). Mitkov (2002: 16) defines an identity-of-sense anaphor as one which "does not denote the same entity as its

[50] It might be noticed that we are overlooking some devices mentioned by Halliday & Hasan (1976), such as conjunction (i.e. the use of formal markers to relate structural units of the text to each other) or verbal anaphora (Halliday & Hasan 1976: § 3.3; Mitkov 2002: § 1.4.4), and other cohesion devices which are not discussed in the sources cited, such as continuity of tense, consistency of style, morphological structure or punctuation (see Baker 1992: 210-212), because here we would like to focus on cohesion provided by referring or applying nominal expressions (basically NPs) only.

[51] Although *anaphora* is the term used by many authors by default to refer to any kind of cohesion, we prefer to keep neutral in this regard and use the less used, but more generic and precise, term *endophora* for what we are discussing here and leave *anaphora* as the designation of just the kind of endophora in which the antecedent precedes the anaphor.

antecedent, but one of a similar description", that is, when the anaphor and its antecedent do not correspond to the same referent in the world but their referents are equivalent in some way (e.g. because they are members of the same class), such type of anaphora is called **identity of sense** anaphora (see ex. 54).[52] Expressions linked by identity of sense anaphora have, then, different referents and must therefore be handled with care, as they can agree in gender but not always will they necessarily do so. For example, the determiner *other* in ex. (54) seems to indicate that Dr. Núñez belongs to the category of 'women', but by no means does "she" in ex. (55) indicate that all the "other professors" are women too.

(54) Dr. Vilma Núñez, President of the Nicaraguan Center for Human Rights (CENIDH), begins her candidacy for President of Nicaragua. (…) Dr. Núñez, like many other women, has been a strong critic of the neoliberal policies instituted during the Chamorro Administration.

(55) Like other professors, she bristles at the idea of treating students as customers.

On the other hand, let us recall that we consider predicates do corefer with other referring expressions such as those occupying the position of subject.

Coreference and gender agreement # WHERE? When two expressions in a text corefer or co-apply, they usually display (# RETOMAR EM PROBLEMA/DIF. COM COREF. RES. "In 2.4.2 we said that...") the same grammatical properties, or at least those which are informative about the inherent nature of the referent or which are discourse independent, such gender (and number, to a lesser extent),[53] unlike definiteness[54] or case, which vary according to the position of the expression in the text.

Therefore, as we will see, we will take into account both corefering and endophoric expressions as cues for our gender interpretation process.

The available literature reflect on what constituents can be satellites, but hardly does on the relationship that must exist between a controller and its satellites. Moravcsik (1978): 'coreferentiality principle'

[52] There can also be cases of coreference without anaphora, but we will not consider them here because they must involve more than one document.

[53] Cfr. "They (pl.) are merely a drunken and surly gang (sg.) of hitchhikers."

[54] Definiteness is a category concerned with the grammaticalization of identifiability and non-identifiability of referents on the part of a speaker or addressee (Loos 2003).

2.4.3 Realizations of nominal endophora

Let us now review the different realizations of nominal endophora[55] (which can always, but not necessarily, hold among coreferential expressions),[56] because they will be the basis for our choice of the features of our machine translation experiments (see #XX). Although we will not exploit or deal with them all at this stage of the research, it is interesting to see them all now to realize what problems they might pose or what kind of help they might provide.

We could, according to the kind of relationship between the elements linked enphorically, easily organize the realizations of nominal endophora in identity of reference, identity of sense or non-identity endophora, but we prefer to do it according to the kind of anaphor, however. We rely mainly on the presentation by Halliday & Hasan (1976) but we follow our own personal organization. Thus, the main devices of nominal cohesion in English can be either grammatical, as pronominal endophora and one-endophora, or lexical, such as lexical anaphora. #VEG. TREBALL DE METODOLOGIA

Pronominal endophora As we briefly saw in § 2.2.3, third-person **pronouns** are used primarily to refer to other referring (and frequently corefering) expressions in the previous (or next) co-text, but there are other types of reference, such as demonstratives (on a scale of proximity) or comparatives (on a scale of similarity) (Halliday & Hasan 1976: § 2.4, 2.5). There are, however, also uses of personal pronouns which are neither referring nor cohesive, such as the generic use of pronouns.

One-anaphora *One*-anaphora is a cohesive relationship between a an anaphoric NP headed by the word *one* and its antecedent. The main kinds of *one*-anaphora (sometimes referred to as identity of sense anaphora)[57] are **substitution** and **fused heads**. Let us see them together with other expressions with the word *one* which must not be interpreted but which might pose difficulties as regards their distinction from gender-interpretable instances of *one*.

In substitution, (part of) an expression is referred to by means of its replacement with a sort of wildcard word: As a rule of thumb, in English the **nominal substitute *one*/*ones***, also known as 'pro-nominal *one*' (Huddleston & Pullum

[55] Nominal endophora is so called when the anaphor refers back to a noun phrase.

[56] We will talk specifically about endophora and not about cohesion in general (which we would have to if we meant coreference to be included too), regardless of whether it is concomitant with coreference or not.

[57] Although coreference is not involved *stricto sensu*, *one*-anaphora has to be interpreted too, as it the *one* anaphor is generally used to refer to an entity of the same kind (and probably of the same gender).

2002: 1513), has the same structural role as the item it substitutes (the NPs' head) and is always accompanied by some modifying element which redefines the referent, resulting in a contrastive effect with the substituted item.[58] (56) is one of the few examples in our corpus with a human referent:

(56) This vague sense of terror kept coming back all night, as if to remind the dancers, including the ones in the bleachers — that there's no such thing as innocent fun.

However, as happens with pseudo-generic pronouns (see above), we must be warned that there are occurrences of *one* which are not instances of substitution and not always need be interpreted (but which are, however, difficult to make out): the generic pronoun, the human 'pro-noun', the cardinal number and the determiner (an alternative form of the indefinite article). The example (57) below shows an instance of the **generic *one*** (which Halliday & Hasan (1976: 98) call pronoun of 'generalized person'), which is neither referential nor cohesive, and must not, therefore, neither agree nor be interpreted.

(57) One wonders how anyone could fulfill the Food and Drug Administration request for well-controlled trials to prove marijuana's benefits.

There is another non-cohesive[59] (and not always referential) form of *one*: that which Halliday & Hasan (1976: 102) call the **'pro-noun' *one*** (also called 'idiomatic *one*' (Ng et al. 2005)). It is exclusively used with human referents and means 'person' (or 'people' in the plural *ones*). If it is referential (cfr. (59)), its equivalent in the TT might be gender-specific and therefore it is necessary to interpret it in English.

(58) Losing a loved one was difficult enough but having a loved one murdered was inestimably worse.

(59) She is certainly the one to do it.

On the other hand, *one* as **numeral** and *one* as **determiner** in head position (which stand for the majority of occurrences of *one* in our corpus) are, for Halliday & Hasan (1976), examples of ellipsis,[60] but we prefer Huddleston & Pullum (2002: 1513-1515)'s explanation because we think that it fits better our data (cfr. examples (62-64) below). For these authors, these are **fused heads**: "the head is combined with a dependent function that in ordinary NPs is adjacent

[58] The pro-nominal *one* is normally an instance of identity-of-sense anaphora (see X), although it can be an identity-of-reference anaphor occasionally.

[59] According to Halliday & Hasan (1976: 105).

[60] Ellipsis is the elision of an element (the head, in this case) that is pressuposed structurally and sometimes present in, and recoverable from, the co-text.

to the head, usually determiner or internal modifier" (Huddleston & Pullum 2002: 410).

In any case, *one* as numeral or as determiner have a cohesive effect which we must take into account here. Examples (60) and (61) show instances of *one* as a numeral (in opposition to twenty-three, fifteen or two):

(60) She was one of 23 children born to former slaves in North Carolina, and one of only 15 who lived to adulthood.

(61) She outlived countless other relatives, including one of her two daughters.

The examples (62-64) show instances of the indefinite article *one*.[61] This (which Ng et al. (2005) calls 'partitive *one*') is the most frequent occurrence of *one* in our corpus, normally indicating that the elided noun belongs to a set of similar entities (by means of the cataphoric expression"one of").[62]

(62) Mr. Ayyappan said that he had discussed his mother's condition with his sister, Sridevi Ayyappan, 28, one of India's top film stars.

(63) She was one of the first people to designate Islamic studies in America as a discipline.

(64) Alta Charo, a professor of law and bioethics at the University of Wisconsin and one of the authors of the Institute of Medicine's report, described (...).

Table 2.6 summarises the variety of subjacent structures which are expressed by *one* and the gender-specificity of its equivalent in the target text according to its discoursive characteristics (provided that the target language belongs to the grammatical gender language type).

Lexical anaphora Lexical anaphora is the "cohesive effect achieved by the selection of vocabulary" (Halliday & Hasan 1976: 274, § 6). There are two kinds: reiteration and collocation#association?.

[61] The indefinite article *one* is the non-weakened (or stressed) form taken by the indefinite article *a(n)* when it is functioning as a fused head (or as head of an elliptical nominal group) with a countable referent.

[62] It is clear that the equivalent of *one* (in structures like "she is one of [A SET]") in the TT must be female-specific when the set is composed of only females (as in ex. (61)). However, as it is both anaphoric with the subject and cataphoric with the succeeding set, it is not clear with which of the two *one* must agree if there is a mismatch between the referential gender of the subject-referent and that of the set (i.e. if the set is mixed), cfr. *[Ella es] ?una de los autores* or *[Ella es] ?uno de los autores* as possible translations of (64).

Type of *one*	**Cohesive**	**Referential**	**Gender-specific**
Substitution	Yes	Yes	Yes
Generic	No	No	No
Pro-noun	No	Possibly	Perhaps
Fused head	Yes	Possibly	Perhaps

Table 2.6: Gender-specificity of *one*'s translation according to its type

a) **Reiteration:** This is a form of lexical anaphora which involves the use of an expression which brings back to mind, or refers back to, another expression which was previously uttered in the co-text, to which it is related by virtue of either a relation of a certain kind of identity (same noun, coreference, synonymy or near-synonymy) between the two or another kind of lexicosemantic relation such as hyperonymy or hyponymy. In many cases the reiteration is combined with demonstrative reference in a definite description ("the/this/that/the same + noun") and that is in fact, more than simply the reiteration, what has the cohesive effect.[63] See the following examples.

(65) The British military spokesman, Maj. Charles Burbridge, said Mr. Faruq was "a terrorist of considerable significance" who had been hiding in Basra, but declined to say whether he was the same man who had escaped from the American military detention center in Bagram, Afghanistan, in July 2005.

(66) Defense lawyers have told reporters that the second dancer at the party has contradicted the accuser. But that woman spoke with a local television station over the weekend, under conditions set by her lawyer that she could not be asked about specifics at the party, and she did not contradict the accuser.

Reiteration can be achieved through **repetition** of the antecedent expression or its head, **synonymy**, **hyperonymy/hyponymy**[64] or **generalization**. Generalization is an extreme case of hyperonymy: **general nouns** are the highest hyperonyms in a lexical set and their cohesive force derives from the fact that, as their meaning is so general, they must be interpreted (or their interpretation needs to be confirmed) by resort to another item. For example, within the lexical set of human nouns, the highest hy-

[63] Considered by Halliday & Hasan (1976) within the same category (endophoric reference) as pronominal endophora.

[64] Personal proper names can be considered as a extreme case of hyponymy.

peronyms (or the more general nouns) are *person*, *woman*, *man*, *child*, *girl*, *boy*, etc.[65]

b) **Association:** Associative anaphora, also known as collocation (Halliday & Hasan 1976: § 6.4)[66] or indirect anaphora (Mitkov 2002: 15), is a kind of cohesion achieved through the use of elements which are related in the reader's mental lexicon by virtue of their normal co-ocurrence or another kind of lexicosemantic association, e.g. meronymy (partitive or part-whole relations), set membership, complementarity, antonymy, causality, instrumentality, local or temporal sequentiality, etc. (Halliday & Hasan 1976: § 6.4; Feliu i Cortès 2004): "There is always the possibility of cohesion between any pair of lexical items which are in some way associated with each other in the language." Let us see a couple of examples:

(67) Representative Carolyn McCarthy, Democrat of Nassau and a crusader for gun control, whose husband was killed (...) in the Long Island Rail Road shootings in 1993, said …

(68) … said Gebran Bassil, a longtime member of General Aoun's Free Patriotic Movement who is also married to the general's crusader.

As we will see, the main endophoric relations that we will exploit are **coreference** or **associative anaphora** (basically a partnership relationship), that is: we obtain cues for the gender of the target referent either (1) directly by resorting to allegedly coreferential gender-specific expressions or (2) indirectly by taking into account the textual reference to a sexual/marital/etc. relationship between the target referent and her/his partner, referred to by a gender-specific expression.

2.5 Equivalence in translation

#(move to "Research problem"?)

Equivalence is a very important concept (in fact "it can be said to be the central issue" (**?**)) in translation and translation theory, but also one about which much debate has arisen, responding to many factors such as the type of text, the type of unit of translation discussed, the type of translation or the approach of each author. Let us now consider two particular notions of equivalence,

[65] Of course, to be cohesive, they must be coreferent or anaphoric with an expression used elsewhere in the text.

[66] We prefer this name so as not to lead to confusion BLA BLA... VANESA (McKeown & Radev 2000)

which can be taken (among many others) as an index of the quality of a certain translation and are directly related to the translation of referential personal nouns. They apply primarily (or were conceived in relation) to translation as performed by humans, but we think that they can also be taken as genuine and reasonable expectations as for the performance of machine translation engines.

Koller (1989) explains what types of equivalence there are for him and tries to set some clarification thereof. Among other kinds, Koller mentions what he calls **denotative equivalence** but what could also be called **referential equivalence** (and what other authors call 'invariance of content'): "the translation process [must] achieve referential identity between SL and TL units" (Koller 1989: 100-101). This kind of equivalence has to do with the extralinguistic content transmitted by the text or, as Baker (1998) puts it, with the equivalence as established "on the basis of the source language and target language words supposedly referring to the same thing" in the world.

The other kind of equivalence that is interesting to consider here is what Baker (1992: § 6) calls **textual equivalence**: this kind of equivalence covers both similarity of the information flow in the source text and the target text (which we will not consider here) and, on the other hand, the cohesion achieved in both texts. While it is true that the target text must be cohesive enough in itself and that languages differ in the strategies they use to achieve cohesion or the level of a certain cohesion device that they tolerate, it is nevertheless also reasonable, if a translation is to be considered accurate, to expect a certain parallelism in the number of cohesion chains and in which particular participants form part of each chain.

Chapter 3

Research purpose/outline/scheme/main points (NOT SURE ABOUT THIS CHAPTER HEADING, WHAT SHOULD THIS CHAPTER BE CALLED?)

As we have seen, the domain of analysis is the written language, the unit of analysis is the text (and more precisely the cohesion of the text provided by nominal expressions) and the linguistic level is the discourse, where a role is played by syntactic, semantic as well as morphosyntactic factors. The research object is the referential gender of referring singular NPs headed by a gender-indefinite personal noun, that is, instances of third-person, singular person reference (address terms and vocatives or first person pronouns are left out).

3.1 Problem statement

The general context of this piece of research is the automatic translation of gender-indefinite or dual-gender personal nouns in English as a source language into any language in which the equivalent expressions must necessarily be gender-specific most of times. The problem we try to solve has to do with the fact that machine translation engines will normally translate such gender-indefinite personal nouns into masculine[1] equivalent nouns by default, regardless of their referent's gender. We will call this **gender-biased machine translation** and we consider it a problem because the target text will be anomalous.

Given the set of expressions in the source language $A = \{\alpha_1, \alpha_2 \text{ ... } \alpha_n\}$, the set of expressions in the target language $B = \{\beta_1, \beta_2 \text{ ... } \beta_n\}$ and the set of human individuals in the referential realm $I = \{i_1, i_2 \text{ ... } i_n\}$, let us consider the following definition of gender-biased translation:

Definition: Machine translation will be gender-biased if there are two expressions $\alpha^{[\emptyset]}$ and $\beta^{[+\textsc{gen}^{-1}]}$ and one human entity $i^{\{\textsc{gen}\}}$ such that $\alpha^{[\emptyset]}$ refers to $i^{\{\textsc{gen}\}}$ and that $\beta^{[+\textsc{gen}^{-1}]}$ is the proposed translation for $\alpha^{[\emptyset]}$.

where the value within square brackets in the superscript of α represents the grammatical gender of the expressions, the value within braces in the superscript of i represents the referential gender of the individual and $\textsc{gen}^{-1}$ represents the counterpart or inverted gender of $\textsc{gen}$: e.g. masculine if $\textsc{gen}$ is feminine and feminine if $\textsc{gen}$ is masculine.

#SOCIAL-GENDER...

# MOVE TO PROBLEM: FROM HERE... The main reason for me choosing English as the object of my research is the fact that English is a natural gender language, in which covert-gender personal nouns pose a translation problem while there are numerous other elements which have overt gender (many more than in genderless languages, in which the same problem exists). Other factors are that, on the one hand, there are enough written production online and a lot of technological applications (such as automatic translation) have been developed that take it as a source language. On the other hand, we have a good command of English that allows us to analyse texts in that language, whereas our knowledge of any other language with covert-gender nouns would not be enough not even to compile a small corpus. # TILL HERE

My definition of gender-biased translation does not define the referential nature of $\beta^{[+\textsc{gen}^{-1}]}$ or the effects of gender-biased translation on the cohesion of the text, so we could add that, while $\beta^{[+\textsc{gen}^{-1}]}$ does not refer to a particular individual i because it is not a referring expression, it does **pseudo-refer** to an inexistent individual i' because it triggers in the addressee the motivation to search for a referent which matches the information given by that pseudo-referring expression. That matching might of course not happen, or it could

[1] This is a broad generalization: in fact, Google's machine translation engine will use the stereotypical gender in a number of occasions, for example with such personal nouns as *nurse*, which are stereotypically hold by women. In most other cases, however, and in all cases in Translendium, the gender used is the masculine.

happen to another individual given the appropriate conditions of the context, or the reference could be successfully established to the originally intended individual despite the contradiction between the expression's grammatical gender and the human referent's gender, by means of an extra effort of processing by the addressee.

Let us exemplify the problem by using one (simple one) of the examples we came across in the preliminary exploration #, repeated here for convenience:

(69) a. Linda Miller is a driver.

b. Linda Miller és ***un conductor***. (Tl)

3.2 The proposed task

The problem we have just described arises, on the one hand, from the normative adoption of the masculine as the default gender (generally assumed by the developers of language technologies), and mainly, on the other hand, from the fact that machine translation engines take into account only what lies within the boundaries of the sentence, ignoring virtually everything which is before and everything which is after it. In order to be more effective, machine translation engines should have a more human-like, discourse-oriented behaviour.Indeed, taking into consideration the text as a whole is essential to have a slight chance of dealing successfully with discourse-related problems as the one we face.

In that line, our main hypothesis was that cohesion, which is broken by a gender-biased translation, might in turn help let the system retrieve the referential gender of the personal noun in question. Gender agreement generally holds between two expressions if they corefer to a human entity, regardless of whether the controller personal noun expresses gender or not. That is, both gender-neutral and gender-specific personal nouns require ineluctably the choice of the satellite forms (e.g. pronouns *he* or *she*) according to their referent's gender, if known.

The task we propose to undertake here is to look at the gender-specific elements in the text which allegedly corefer with the personal noun whose referential gender we intend to infer, so as to classify that particular instance of human reference as feminine or masculine.[2] So, we approach the task as a problem of binary classification.

[2] Or something else, as we will see.

3.3 Objectives

- We aim to demonstrate that, by means of a shallow analysis of coreference or co-application relations in an English text, it is possible to determine whether a referential expression headed by a gender-indefinite personal noun in that text refers to a woman or else.
- We aim to propose a hybrid/combined, shallow, coreference-based analysis method (or ALGORITHM), integrable in the analysis module of an MT engine, which determines whether a referential expression headed by a gender-indefinite personal noun in an English text refers to a woman or else.

3.4 Hypothesis

- The grammatical information provided by elements in an English text which are expected to be discoursively related to a referential expression headed by a gender-indefinite personal noun can be used successfully to determine whether such referring expression refers to a woman or else.

3.5 Variables

# expressions which serve as cues do so on the basis of grammar (intralinguistic?) or on the basis of world knowledge (extralinguistic?)

Our dependent variable[3] is the referential gender (*RefGen*) of a gender-indefinite nominal expression $\alpha_i^{[\emptyset]}$, which we will call the **target** expression t.[4] Our independent variables are the patterns of co-occurrence between the target and a gender-specific anaphoric expression ($\alpha_i^{[+\textsc{gen}]}$ if coreference is involved or $\alpha_j^{[+\textsc{gen}]}$ if it is not). These anaphoric expressions we will call the **cues** $c_1, c_2 \dots c_n$ and the patterns of co-occurrence (and expected agreement if there is coreference) between a target and a cue will be expressed as $t \Leftrightarrow c_i$. We can express this relation as a function[5]: $\mathit{RefGen}(t) = f(t \Leftrightarrow c_1,\ t \Leftrightarrow c_2 \dots t \Leftrightarrow c_n)$.

[3] The independent variables are the factors that are controlled or varied in the experiment. On the other hand, the dependent variable is observed or measured for variation to determine if the variation of the independent variables had any effect on it.

[4] Corbett (1991: 151) advocates for differentiating between *controller genders* (e.g. the sets into which nouns are divided) from *target genders* (e.g. the genders which are marked on adjectives, verbs and all other agreeing elements, depending on the language). These are not to be confused with my designations *target* or *cue* — in fact, there is an inverse correspondence, i.e. my target would correspond to Corbett's controller and my cue would be Corbett's target.

[5] The independent variable is the input of the function whereas the dependent variable is the output of the function, that is, the dependent variable depends on the independent variable.

Therefore, our aim is to play with the patterns, adding or removing them, and observe how the gender interpretation process (in terms of the number of correctly classified personal nouns) responds to the quality and quantity of the patterns used.

As we saw (vid. 2.3.3 and 2.4.3), NPs in English can either express gender or not: most of personal nouns do not express gender and most personal pronouns do, whereas human named entities might do clearly, if they contain a title, or less clearly, if they do through a given name (which requires external knowledge to account for it) or not at all. While a target could be either a gender-indefinite personal noun, the pronoun *one* or any gender-indefinite human named entity, in this thesis project we will only deal, however, with personal nouns as targets. The cues, on the other hand, can be either gender-specific personal pronouns, personal pronouns and gender-specific human named entities. #IF A HUMAN ENTITY HAS A TITLE OR A GIVEN NAME THE GENDER IS RAISED TO THE FULL ENTITY. (SEE PROCESSING OF PROPER NAMES)

3.6 Justification

The topic of this dissertation is an anomaly or imperfection in a technological application. Therefore, it could be considered undesirable in absolute terms, whatever the means and the effort needed to deal with it. As for how worth it might be to devote energy to solving this problem, we think that it is legitimate for a PhD student to give it a try. If the solution can be found in the framework of a M.A. dissertation such as this, then we think that the energy and the resources needed are worth it.

On a more technical level, we can assess the traductological, linguistic or psycho-social consequences of allowing this anomaly to crop here and there in automatic translations. As we have seen, when translation is gender-biased, the cohesion of the text is broken and the intended reference might not be successful (Lyons 1977: 180-181). We also saw how equivalence is affected from the referential and the textual point of view. Gender agreement, as well as number agreement or semantic class agreement, is also a useful indicator of coreference in correference resolution applications, but it can only be used as a feature if it is known.

On the other hand, from a psycholinguistic point of view, personal nouns, if used appropriately, may contribute to the construction of individuals' identities in the desired direction but, if used inappropriately, for example referring to someone repeatedly "by a false name, by using derogatory or discriminatory language, or by not addressing someone at all, may cause irritation, anger of feelings of inferiority" (Hellinger & Bußmann 2001: 3). To that list, we could add, for example: by referring repeatedly (by mistake or by intended slovenliness about correcting this mistake) to a woman with a masculine personal noun.

The effects of gender-biased machine translation on different areas of life, especially women's lives, have not yet been investigated. However, some research has been carried out on the cognitive effects of different (pseudo-)generics, including (pseudo-)generic nouns, and the influence of sexist vs. gender-fair language, which suggest that language might reinforce or weaken gender stereotypes as well as affect the development and planning of women's professional lives (see the summary of studies in #SHALL I CITE THEM MYSELF HERE? Irmen 2007): i.e. the use of masculine generics may have harmful consequences for women. Even though the impact of machine translated texts on people's attitudes and values is allegedly much lower than that of other more standard texts of higher prestige and diffusion, those results contribute further to the desirability of achieving gender-unbiased machine translations.

Chapter 4

Brief state of the art

#PENDENT D'ACABAR........

Several approaches exist that study information technology and artificial intelligent from a gender perspective, both by social scientists and by IT professionals or computer scientists (see the different contributions to the International Symposium 'GIST – Gender Perspectives Increasing Diversity for Information Society Technology' that took place in Bremen in June 2004, some of which are gathered in Zorn et al. (2007)): namely the liberal tradition, the standpoint theory and post-structuralism. They try to show in different ways how gender research can bear an influence on software design, pointing out in what ways software can be gender-biased and, in turn, in what ways it contributes to the construction of genders (e.g. stereotyping), as well as how software systems can be designed in a more gender-fair manner (not based on simple stereotypes)[1] so as to "bridge the digital divide" and build an inclusive information society, i.e. facilitating participation of women (which is generally lower than men's, according to Stewart (2002), cited in Maas et al. (2007)) both in design and use of computers. "Lacking of opportunity as well as individual abstinence both have the [effect of] being a 'non-user' of computers or the Internet" (Maas et al. 2007: 10), what might lead to exclusion-related problems.

Gender & AI

(Adam 1998)

However, none of these approaches reflects on how machine translation deals with gender-related issues.

Gender & language: Cameron, Pauwels...

[1] "We hope (...) that software construction informed by gender studies will support the deconstruction and transformation of existing assumptions and structures as far as gender relations are concerned and will lead to software and software design processes that will empower men *and* women." (Zorn et al. 2007)

On the other hand, gender has been treated from the perspective of translation theory and translation studies in general due mainly to feminist criticism, as well as a deeper understanding of both gender, its relation with power, identity and language and the anthropological implications of translation both as a process and as an outcome. However, translatologists have focused so far on the representation of women only as far as human translation, especially literary translation, is concerned, or on the role of women as translators, taking translation as an indicator of women's access to mainstream channels of expression or as a field of experimentation for feminist writing (von Flotow 1997). Therefore, translation studies with a gender-oriented approach have been useful in increasing future translators' awareness about the repercussions of a gender-blind or an androcentric practice of translation (Susam-Sarajeva 2005), but have not devoted any attention to automatic translation, that we know of.

Gender-issues in machine translation have not been an important topic of research for computational linguists, either. #MT.

Frank et al. (2004) are the only ones, to the best of our knowledge, to address specifically the question of how gender is handled in machine translation. They point out the importance of gender issues in translation and how that subject is indeed a pervasive means of showing the limitations of machine translation in an explored way. Also, they draw attention to the fact that, even though human translators make use of their knowledge of many subtle aspects of culture and the world at large that machine translation engines might never have, it is nevertheless true that "even a software program can identify certain contexts that allow for the correct handling of particular grammatical features, in this case, feminine forms," even if it has to resort to textual information beyond the boundaries of the sentence.

Frank et al. (2004) carry out a comparative evaluation of commercial machine translation engines from the point of view of gender-related phenomena (e.g. person reference, anaphora, etc.) and discuss the coverage of dictionaries and rules of the systems they analyse, as well as issues of usability and user awareness or expectations as regards gender-(un)biased machine translation.

They point at the problems but do not suggest a means to solve them

WHAT MEANS TO SOLVE THEM? WSD:

Based on the similarity of both tasks (if we draw an analogy between the different sense that a polysemic word can have and the two covert genders that a personal noun), some approaches taken in the field of world sense disambiguation could be adapted or serve as inspiration for the task at hand. For example, in machine learning-based word sense disambiguation, which is simpler and sometimes more efficient than sequential application of rules (like in a decision list). the relevant linguistic features are selected first and then are encoded in a form usable for a learning algorithm.

Other: supervised (need large sense-tagged training set), bootstrapping, unsupervised, dictionary-based $\neq$ gender interpretation (the former: denotation $\neq$ latter: reference)

ML applied to NP corref SOLTAR AQUI RESUMOS DOS ARTIGOS

Chapter 5

Methodology

In this chapter I describe how I planned the work I did and what steps I took to solve the problem of gender interpretation. In the first place, I will describe my corpus, followed by an explanation of the kind of approach that I took and how it differs from other well-established NLP domains. Then, I will go on to describe the phase of preprocessing and feature extraction, as well as how the final results were obtained and how good they are. Figure 5.1 represents the processing chain I have just briefly introduced. Finally, a brief comment on the main problems I encountered will be given.

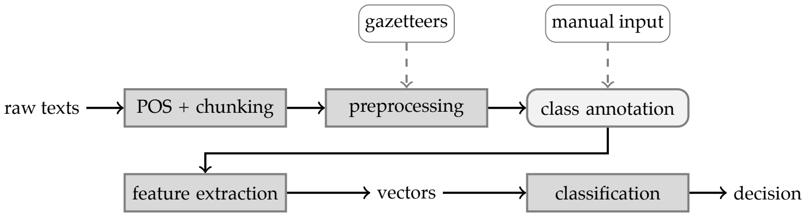

Figure 5.1: Processing pipeline.

Ours is a hybrid or combined approach because the process of interpretation consists of two clearly differentiated parts. On the one hand, human named entities / personal names are classified as feminine or masculine during, and as part of, the preprocessing of the data. The classification of some of them, considering only one expression at a time, relies exclusively in their univocally gender-specific (*Mrs.*, given names, etc.) constituents, whereas the gender of other expressions, which contain no gender-specific constituent, is inferred through a simple heuristic process (see XX#preprocessing of names). On the other hand, NPs headed by personal nouns are classified in a later phase through a somewhat more complex machine learning based process (see #XX).

5.1 Gender interpretation vs. cohesion resolution

In a first moment, as two corefering expressions would normally have the same grammatical gender, one could think that the most plausible or efficient approximation to gender interpretation could be through coreference resolution. If one succeeds in finding coreference chains, it goes without saying that all elements in the chain will share the same grammatical information, including gender. However, this approach might not necessarily be the most efficient one, due to its high complexity. Neither is anaphora resolution necessarily the best approach, because normally, given a particular pronoun, anaphora resolution tries to find out which is/are its/their antecedent/s out of all the previous NPs, whereas we need to walk exactly the opposite path: given a NP headed by a gender-indefinite personal noun, what are the elements (such as pronouns) which might be anaphors of it?

The main difference between anaphora or coreference resolution and the method for gender interpretation proposed in this dissertation is that I am not concerned directly with whether two expressions corefer or not. We are concerned neither with finding the antecedent of a pronoun, as in much of the work in anaphora resolution, nor with finding what is the individual referent of a number of coreferring elements in the text, as in coreference resolution, but instead my approach is one of binary classification in which I must assign one of two categories (feminine or else) to each expression, regardless of who precisely that entity refers to individually. Our alternative approach, then, does not consider that there are as many potential referents as expressions in the text, but instead considers only two possible classes (or collective referents) for all expressions.

The point is not, then, to check whether two expressions are coreferent so as to subsequently make the target expression subsume the gender category of the cues. Instead, there are a number of patterns which, if present and detected, contribute to raise the probability that the target has a particular grammatical gender. All of them configure a set of (perhaps contradictory) indicators the overall consideration of which is expected to determine the most likely gender. Take, for instance, the case of a text in which there are only expressions referring to female humans: trying to find out the individual referent of each realization of human reference would be much less efficient than following our approach.

It might be thought that there is an underlying implicit assumption that, in most cases, the target and the cue corefer in determining that the human entity referred to by the target expression more likely has a particular referential gender x just because it is in the company of one or more cue expressions with the corresponding grammatical gender (among many other factors). However, it might turn out that they actually do not corefer but in fact refer to two different persons of the same gender (#SEE IDENTITY OF GENDER?). In that sense we could say that it is not coreference to a single person what we are after, but coreference to a single gender, or to people of one gender.

This explanation might be better understood with an example:

(70) In 1990, just out of college, she met Jonathan Steinberg, the son of Saul Steinberg, the corporate raider. As the two began dating, it was Mr. Steinberg, who presided over a personal finance magazine and a hedge fund, who attracted the media attention. #SOURCE: (see corpus).

Reference is made three times to people named Steinberg in the previous excerpt, ex. (70), which might be a common thing when several members of a family are mentioned. Two of them include the given name, a different one in each case, so it is easy to know they refer, at least, to two people and not to the same person. A coreference resolution algorithm should try to discover whether the entity "Mr. Steinberg" could refer either to Jonathan or to Saul or to someone else not previously mentioned. In our case, we do not care about to whom it refers as long as we know he is a male. There are also another masculine positive cue ("son") and a feminine negative cue ("she"), so the gender interpretation process would not be straight-foward and, for it to be successful, more importance should be given to the four positive masculine cues at the expense of the one negative cue, but it would not need to find out whether the target "corporate raider" corefers with "Jonathan Steinberg", with the "son", with "Saul Steinberg", with "Mr. Steinberg" or with "she".

5.2 Preprocessing

5.2.1 Data collection

Our data is composed of 46 plain text articles manually selected and downloaded from the New York Times online newspaper.[1] The criteria for selecting the texts has been that they contain at least one reference to a female individual, i.e. most texts with no singular reference to at least one woman were excluded. Each text contains the full article (when they are distributed in more than one

[1] URL: http://www.nytimes.com

web page, the different parts were put together in one only file, in order to respect the text's structure as conceived by its author, regardless of how it is displayed in a particular medium). All the HTML tags and embedded code were not considered, as well as any structural feature which is not the text itself, such as the author credit, the publishing date, pictures, links, advertisements, etc. In the title, as is frequent in English and is the case with New York Times articles, all words are capitalized, what poses a great difficult for their exploitation, as proper names and titles are not distinguished from common nouns, so personal nouns in the title were not considered for interpretation but the title nevertheless was used for search of cues (especially for preprocessing personal names, see below). Data collection has accounted for a very large part of the cost and time of developing our pattern recognition system.

5.2.2 Corpus annotation

We proceeded to annotate the discourse entities or expressions we would want to handle subsequently. For that, several options existed, the most comfortable one being in principle the use of an annotation tool, such as **GATE**'s ANNIE (Cunningham et al. 2002). However, the annotation provided by GATE is not exempt of a series of problems, namely:

- It will sometimes annotate last names as `FirstName`, such as *Moore* in "Thomas A. Moore," meaning that it does not use contextual information to recognize the internal structure of human named entities.
- The orthographic coreference engine will not take into account non-Western name patterns, e.g. Chinese: *Gao Yaojie* and *Dr. Gao*.
- Such frequent job titles as *lawyer*, *professor*, *farmer*, etc. are missing from GATE's gazetteers.
- It will annotate as `JobTitle` not only nouns but also verb forms and some adjectives, such as *monitor* in "The authorities often tap phones and otherwise monitor people deemed troublemakers.", showing that it does not take into account the earlier POS-tagging at this stage.
- It will annotate as `Person` and/or `FirstName` words whose initial letter is not even capitalized, such as the underlined ones in the following examples: "tales" or "Senator Barack Obama drew cheers …"
- It will annotate as `JobTitle` words that are not even nouns: "their chief preocupation".

- It will annotate names of cities as `FirstName`: “the city of Florence”, “the university of Virginia”, etc.

- It will assign gender information incorrectly, as when it annotates *Clinton* as a male `FirstName` in “Senator Hillary Rodham Clinton” or “Mrs. Clinton”.

Mainly for all these reasons, the texts annotated with GATE had to be revised manually, with a considerable waste of time and effort, and that is why we decided to annotate the texts by other means, namely, POS-tagging and shallow parsing with **TreeTagger**[2] (Schmid 1994 and Schmid 1995), followed by an ad-hoc, semi-automatic, Perl-based XML annotation of the relevant discourse entities, with resort to built-afresh gazetteers, followed by some post-editing to remove certain easily-detectable noise.

An NP is considered to refer to a human entity (i.e. an expression referring to a person) and is annotated as such depending on its constituents, which might be either a third-person singular personal pronoun in head position, a personal name or a singular personal noun. However, different annotation strategies were followed according to the different elements to be annotated in the texts. On the one hand, **personal pronouns** are straightforward to annotate, simply by means of a list of the forms to be recognized and tagged. Indeed, this is the less problematic part of the entity annotation. As we saw, only third-person singular pronouns (either in subject or object function, either as determiner or head, ether as existencial or as possessive) are taken into consideration.

On the other hand, **personal nouns** (tagged as `<role>`), including occupational terms,[3] demonyms, kinship terms, general terms, sex-related terms, etc. (see 2.2.3), have been annotated as such when they were the last constituent of an NP (presumably the NP’s head), by resorting to three gazetteers: one for female-specific nouns, one for male-specific nouns and a third one (the longest one) for gender-indefinite nouns. Nouns listed in the first and the second gazetteers were annotated as cues and were assigned their gender class `gen=(f|m)` and nouns included in the third one were annotated as targets. Some adjectives (e.g. *lesbian*, see feature #XX) were annotated too, as cue modifiers within a NP, either referring to a person or not, either singular or plural.

A quite different, and also more complex, approach was followed to annotate **personal names**. Human entities were tracked and recognized depending on their internal structure and the gender indicated by some gender-specific constituent (if there was one) was purportedly generalized to all references to

[2] Source: `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

[3] Not to be mistaken with occupational titles.

the same person, basically based on orthographafic overlapping. Next section 5.2.3 describes in full depth the preprocessing of personal names.

5.2.3 Preprocessing personal names

Personal names can be very informative as regards their referent's gender and can thus be used as cues for gender interpretation of coreferring personal nouns, but at the same time they can be very difficult to deal with, as there is a great deal of variation depending on cultural background of the individual referred to. For our purposes, we need to identify (1) a number of elements that usually appear as part of a name (such as a given name) or adjacent to it (such as a title) and that normally agree with the referential gender if they are gender-specific (i.e. cues), and (2) other elements which do not openly agree in gender but which can be subsequently used to refer to the same individual, namely the family name. The number of elements in an personal name should not be a problem for processing it, but their order could be. It goes without saying that I will only be concerned with names and name syntax in written form.

The automatic recognition[4] of a personal name's gender must rely on world knowledge or some kind of inference because they generally have no univocal gender mark (except courtesy titles, the *kunya*, patronyms and, to a lesser extent, some Japanese given names) that indicates whether they are female-specific, male-specific or neither. On the other hand, some names (unisex names) do not provide any gender information and should be targets of gender interpretation rather than cues.

Personal names are considered as such only as long as they begin with a title or begin (for a Western name) or end (for an Eastern name) with a given name present in a gazetteer. Therefore, unless introduced by a title, a personal name's referent's gender can be deduced either by searching in a lexicon or by a morphological analysis. The former applies to given names: if the given name is in a list of male-specific or female-specific given names, it can be used as a cue. The latter applies to a few given names (*Geraldine*, etc.) but especially to patronyms which take different suffixes according to the referent's gender (see Arabic, Rus-

[4] It goes without saying that before trying to infer a personal name's gender, it must have been effectively tracked down and annotated as a human named entity. Several criteria can be followed for this: (a) they must have a syntactic status of NP (and be shallow-parsed accordingly); (b) all constituents must be capitalized (except *kunya* in Arabic names); (c) their recognition can be triggered by the presence of a title or a given name included in a gazetteer; (d) the last constituent must not be included in a stop list (*Airport, Army, Avenue, Bay, Church, City, College, County, Court, Center, Garden, Hospital, Parish, Province, Qaeda, Restaurant, River, School, Square, Street, University, Valley*, etc.), in which case it would be the head of a non-human named entity. Some *ad hoc* noise cleaning must be done too, so as to make sure that such expressions as, say, sentence titles of civil lawsuits (e.g. *Branzburg v. Hayes*) are not annotated as personal names.

sian, Greek or Slavic names). If a title of courtesy is present, the gender can be inferred from it and the rest of the constituents can be safely ignored. A personal name is, therefore, annotated as feminine (i.e. <person gen="f">) when the given name (the first element in the Western order, or the second if preceded by a title, or the last element in other orders) is in the list of female given names.

Before checking whether a given name is in a list, however, one must know in what position of the personal name the given name is, as it would of little use to look up the element, say, *Lee* in a list of given names it we are dealing with a Chinese personal name (in which the first element is not the given name but the family name). For that, some evidence of the name's structure pattern is needed or what naming tradition it comes from. Some means can be: the syllabic structure of the name, the presence or absence of a hyphen (which could be used to differentiate Korean given names, which are not hyphenated, e.g. *Il Sung*, from Chinese ones, which are), key words in the text (the name of a country or a capital or a demonym), the name of the place where the story has been written which usually introduces it etc. For example, the article from which the example (49) has been taken contains the word *Vietnam* in the title and begins as "HAIPHONG, Vietnam — The neighbors know what…". That, together with the fact that such names as *Pham Thi Hue* have three constituents, each of which is very short (typically from two to four letters), would suggest that this is a Vietnamese name. Similarly, the word *Hanoi* appears in the title of the article which is the source of (48) ("Hanoi Seeking Beverage Curbs") and the words *Vietnam* and *Vietnamese* appear in the text:

(71) The Prime Minister of Vietnam, Vo Van Kiet, wants to … the state-owned English-language newspaper Vietnam News has reported … prosperous market of 70 million Vietnamese …

Ideally all the sources of evidence mentioned should be taken into account so as to distinguish Western people referred to in an article about, say, Laos. For example, the name *Finn* ***Reske-Nielsen*** appears in a text written in "VIENTIANE, Laos," but that does not mean that it is to be considered a Laotian name straight away, as *Dr.* ***Phengta*** *Vongphrachanh* should.

No prospection has been carried out yet, however, as to the frequency with which these key words appear in articles in correlation with personal names whose structure diverge from the standard (Western) one.

5.2.3.1 Exploitation of gender-specific constituents

The gender information associated with a gender-specific given name or title can be generalized to the full personal name referring to the individual in question and the whole NP can be annotated accordingly: personal names containing a gender-specific item are annotated as cues with the prefix `c` and the appropriate gender, e.g. `<c:person gen="f">` and personal names within which no gender-specific item is found are annotated as targets, e.g. `<t:person>`. Inversely, the gender information can be extended to all the personal name's constituents, such as the last name (as we will see below).

Regarding names with the Western order, in the absence of a gender-specific title only given names at the beginning of the full name are sought for and annotated as such, all the other elements in a name which might be given names or parts of the given name being treated as something else. If more than one given name is present in an individual's name, for convenience only the first one is annotated as given name (unless the gazetteer of given names contains the matching compound name, e.g. *José María*, that is). This is so because it is not always easy to determine whether what follows a first name (and is not the last word in the name in question) is a second given name or any sort of middle name. In any case, having only one given name or a title suffices to infer the gender of the individual, so handling more than one given name is unnecessary and any additional ones can be safely ignored. Trying to deal with more than one given name per personal name might actually be misleading because of patronyms in some cultures or polysemy (a word which might be both a given name or a family name, e.g. *Mathieu* in French).

5.2.3.2 Establishing human entity chains

As we saw, only personal names which contained a recognized given name or a title are initially taken into account and exploited as regards the gender information they provide. However, as we saw, there are other occurrences of a personal name, especially on second and subsequent references, which do not contain all the constituents, consisting normally of the last name only preceded by an optional title (or, much more rarely, of the given name). No problem exists if such a last name is introduced by a gender-specific title (e.g. *Ms. Mouskouri*), but if it is introduced by a dual-gender title (e.g. 72) or by no title at all (e.g. 73), it should be linked to previous (or subsequent), gender-specific named references to the same person as a means to infer its referent's gender.

(72) Three weeks ago, Dr. Wafa Sultan was a largely unknown Syrian-American psychiatrist living outside Los Angeles, nursing a deep anger and despair about her fellow Muslims. (...) Dr. Sultan bitterly criticized the Muslim clerics.

(73) It might be a while before David Beckham steps on the field again for Los Angeles or England. Beckham, whose five-year contract guaran-

> tees an annual salary of $6.5 million, could miss the rest of the Los Angeles Galaxy's season because of a sprained right knee.

For this, my approach is simply based on anthroponymical overlapping: if two personal names share one or more constituents (namely, their family name), they are candidates for correference. My approach is inspired on Gernsbacher (1990: 221)'s 'structure building framework' (SBF),[5] according to which three structure building processes underlie comprehension phenomena: a) laying foundations of mental structures, b) mapping coherent information onto developing structures, c) shifting to initiate new structures when new and given information do hot cohere. According to this scheme "the goal of comprehension is to build cohesive, mental representations or structures." The first step in building a structure is "laying a foundation." The next process involves "developing the structure by mapping on incoming information when that information coheres with the previous information." When the incoming information is less coherent, the reader undertakes a different process: she or he shifts to initiate a new substructure.

If we apply this to the processing of personal names in a text, we can assume a substructure is built for each individual referred to (through one or more expressions) in the text. The reader "lays the foundation" of a substructure when a personal name is met for the first time as the reading progresses (e.g. *Bill Clinton*). That substructure is developed as long as the properties of second and subsequent personal names mentioned in the text are not in contradiction with the properties of the first personal name (e.g. *Mr. Clinton* or *ex-President Clinton*). If the new information is not coherent with the developing substructure, the foundation for a new substructure is formed corresponding to the new individual (e.g. *Mrs. Clinton* or *George Bush*). Therefore, unless the features of a new personal name are contradictory or incoherent with previous ones, they can be assumed to refer to the same individual and both can add to the development of the same structure.

This approach, however, can be particularly complex at times. Apparently correfering names should therefore be handled with care because in some countries, such as the US, a woman will normally adopt her husband's lastname when she marries him, or because a text can talk about two or more members of the same family (e.g. parent and child), or simply because there are references to two or more people who happen to have the same family names (although not the same family). Apart from the example of the Clintons in the previous paragraph, the example (74) is an excerpt from an article which talks about (*i*)

[5] I do not make any claim for the psycholinguistic plausibility of this framework, but I think it can be applied to human entity processing, at least from a computational point of view.

a woman who marries (j) a judge (by whose last name she is referred to on second reference) and also about (k and l) his parents (who also carry the same last name), so overall there are four people (two men and two women) referred to by the same last name.

(74) Barbara Joan Wax$_i$, a former buyer and manager of the S'fari Room at Bonwit Teller in New York, was married yesterday to Associate Justice Bentley Kassal$_j$ of the Appellate Division of the Supreme Court of the State of New York, First Department in Manhattan. (...) Mrs. Kassal$_i$ is a daughter of Ethel Wax of New York and the late David Wax. (...) Justice Kassal$_j$, a son of the late Mr. and Mrs. Hyman Kassal$_{k/l}$ of North Miami Beach, Fla., and formerly of New York, is a former Democratic member of the New York State Assembly and was a Democratic candidate for the 20th Congressional District in Manhattan.

5.2.3.3 Avoiding gazetteers

As we have seen, if relying on gazetteers, the decision as to whether a personal name should be considered as male-specific, female-specific or dual-gender depends, mainly, on whether the given name is included in a particular list of names. However, it would be practical to find an alternative method for this task, especially if the suggested gender is not in absolute but in probabilistic terms, in as much the use of gazetteers can be cumbersome to compile and use, they can hardly every be exhaustive as to include any given name that can appear in a text and, also, they can only indicate whether a name has a given gender with 0%, 100% or 50% of certainty (it seems silly to consider that *Joan* is as likely to refer to a male as to a female in an article from the *New York Times* when the probability that it is feminine is intuitively much higher than that it is the Catalan name).

With that view, in order to estimate the plausibility of determining a given name's gender taking into account only internal evidence, I have run a simple similarity-based learning experiment with TiMBL (Daelemans et al. 2007; van den Bosch 2007), for which the basis of classification is the storage of all training examples in memory. I used 2 classes (feminine and masculine) and 9 features (the number of letters, whether the initial is a vowel or a consonant, whether the last letter is a vowel or a consonant, the first letter, the second letter, the third letter, the antepenultimate letter, the penultimate letter, and the last letter). The data was composed of 18,906 given names, out of which 8,452 (44,71%) were feminine names (the data was, therefore, slightly skewedly distributed) and the algorithm had to classify each data item as masculine or feminine. The algorithm chosen was leave-one-out: with this approach, every data item in turn is selected once as a test item, and the classifier is trained on all remaining items. The accuracy of the classifier is then the number of data items correctly predicted. The results obtained are as follows: if the number of nearest neighbours or k (i.e. the most similar memory patterns on which the output class is to be based) was set to 1, the overall accuracy was 0.73, whereas if k was set to 5, the overall accuracy would rise up to 0.77.[6]

The confusion matrix (see Appendix 6.1.2) provides the necessary data to calculate precision and recall both for masculine and feminine names. Recall is the measure of how much relevant information the system has extracted from the text (that is, the number of correctly classified items over the total number of possible correctly classified items) and is thus a measure of the coverage of the system. Precision is a measure of how much of the information that the system returned is actually correct (aka accuracy), that is, the number of correctly classified items over the number of items classified (correctly or not) (Jurafsky & Martin 2000). For the feminine items, then, recall equals 6,079 / 8,006 = 0.76 and precision equals 6,079 / 8,452 = 0.72. For the masculine names, recall = 8,527 / 10,900 = 0.78 and precision = 8,527 / 10,454 = 0.82. See Table 5.2.3.3 for a synthesis.

	Feminine	Masculine
Precision	0.72	0.82
Recall	0.76	0.78

Table 5.1: Evaluation for masculine and feminine given names

It may be added that this alternative method or any other alike is not incompatible with gazetteers and could, in fact, be used hand in hand with them so as to infer the gender of dual-gender given names,[7] as well as of some sexuality-related or body-describing nouns (e.g. *brunet*, *gay*) and some gender-indefinite role nouns which can be nonetheless inflected into female-specific nouns (e.g. *actor*, *hero*, etc.). In those cases, one cannot have the absolute certainty that the word refers to a male or a female, although in some cases, however, the probability that an item refers to a member of a particular gender is overwhelmingly higher than the probability that it refers to the other. It should then be desirable and possible to find an estimation of their most likely referential gender.

[6] The similarity was computed as weighted overlap and the relevance weights were computed with gain ratio. For more details, check Appendix 6.1.2.

[7] The list of dual-gender given names has been obtained from the intersection of the two lists of gender-specific given names (for male and for female), together with the given names which the source considers unisex.

Another possible experiment (which we have not carried out) would consist of running Google searches of the given name together with some gender-specific items which would likely appear in a text in which a person is referred to: <given> she|Mrs|Ms|Miss vs. <given> he|Mr. The search with the highest number of hits would provide the value m or f, which could be used as a new feature in a target's vector.

5.2.4 External resources

A list of the words and expressions we wanted to annotate was necessary for their recognition. These word lists, also called **gazetteers**, were compiled from different sources, basically from web sites about names, and fused with GATE's own gazetteers. The given names were obtained from the sites www.sudairy.com, www.behindthename.com, www.clicfilhos.com.br, www.angelmums.com, www.20000-names.com, Bob Baldwin's collection from MIT[8] and other similar sites. Kinship terms were obtained from the Wikipedia[9] and the role nouns were obtained from the English WordNet[10] (as hyponyms of professinal titles, person, etc.), the Dictionary of Occupational Titles[11] and the job description indices from the sites CareerPlanner.com and http://www.stepfour.com/jobs/. The demonyms were obtained from the Wikipedia[12], The Geography Site[13] and Everything2.[14]

For given names, two lists were initially compiled, one for males and another one for females. However, a number of dual-gender given names were observed to be included in both lists, so they were removed from them and included in a third list of dual-gender given names which, unlike gender-specific given names, would be annotated as neither feminine nor masculine and would therefore be targets rather than cues.

As it would be ideal that our method of gender interpretation were autonomous and did not rely on external resources, we also consider the possibility of not using gazetteers and rely uniquely on the information contained in the texts themselves. Our prospective observation of the texts show that not resorting to anything but the text could sometimes provide the cue needed to interpret targets. For example, in (75) the pattern "Dr. Sinha and *his* colleagues"

[8] ftp://ftp.ox.ac.uk/pub/wordlists/names/male-names-kantr.gz and ftp.ox.ac.uk/pub/wordlists/names/female-names-kantr.gz

[9] Source: http://en.wikipedia.org/wiki/Kinship_terminology

[10] Source: http://wordnet.princeton.edu

[11] Source: http://www.occupationalinfo.org

[12] Source: http://en.wikipedia.org/wiki/List_of_adjectival_forms_of_place_names

[13] Source: http://www.geography-site.co.uk/pages/countries/demonyms.html

[14] Source: http://www.everything2.com/index.pl?node=Demonyms%20of%20the%20World

can be used to infer that a person of last name *Sinha* is a man, and subsequently that information can be used to classify the target "Pawan Sinha" as masculine and thus use it as a cue to interpret the target "scientist".

(75) Pawan Sinha, a cognitive scientist at the Massachusetts Institute of Technology, has devoted years of research to figuring out just what attributes touch off these face-specific pings. Security software that is being developed for identifying potential terrorists or detecting intruders must be able to reliably recognize faces. In teaching the software to do this, Dr. Sinha and his colleagues have arrived at unexpected insights into the question of why we sometimes see a cinnamon bun as a cinnamon bun, and other times as the earthly incarnation of a sainted nun.

But unfortunately it seems that in the overwhelmingly majority of the cases such information only would not be enough. In any case, we will leave the door open to further research in this regard in the future.

5.2.5 The tagset

Table 5.2 shows the tags used to annotate the texts. Some tags might take a prefix (`c` or `t`), according to whether the element they embrace is a cue or a target, respectively (elements which are embedded in other elements do not have a prefix, for reasons that we will see next). It can be noticed too that the some tags are embedded in others or, inversely, that some entities might contain other tagged elements, namely, a `<person>` tag, which embraces a named entity, might contain the tags `<title>`, `<given>` and/or `<last>`, which embrace a title, a given name and a family name, respectively.

The annotated elements have attributes. Cues have the attribute `gen=(f|m)` and targets (personal nouns) have the manually assigned gender class (`myg=(f|m)`) manually assigned after the semi-automatic annotation process, whose values can be `f` for feminine, `m` for masculine (see 5.2.6).[15] Given names identified as gender-specific have their gender information generalized to the personal name (the named entity tagged as `<person>` which contains it, which takes then the prefix `c` and the attribute gender `gen=` with either `m` for masculine and `f` for feminine as possible values. The personal name, and not the given name, is the one taken into account as cue later on on the feature extraction stage (the elements it contains can be safely ignored as soon as the full expression has been assigned the gender attribute of one of its parts).

[15] Remember that the class `m` is

Element	Tag	Embeded in:	Attributes
NAMED ENTITY	<(c|t):person>		gen=(f|m)
GIVEN NAME	<given>	<person>	gen=(f|m)
FAMILY NAME	<last>	<person>	gen=(f|m)
PRONOUN	<c:pron>		gen=(f|m)
PERSONAL NOUN	<(t|c):role>		myg=(f|m)
TITLE	<title>	<person>	gen=(f|m)
MODIFER	<c:mod>		gen=(f|m)

Table 5.2: Tagset

5.2.6 Classes

Initially we intended to classify all targets among masculine (m), i.e. referring to a male human, feminine (f), i.e. referring to a female human, and else. "Else" included unknown (u), i.e., reference to a person whose gender cannot be determined, neither automatically nor humanly; non-referring (n), i.e, expressions which do not refer (see 2.2.2); and noise (o), i.e. words which are not personal nouns, as happens with polysemous nouns, such as *head, chair, authority, relative, relative,* etc. The idea was to let the machine translation engine (or any other) know which targets have an unrecoverable gender or for which of them gender interpretation does not apply, so that the appropriate decision can be taken (inform the user about that ambiguity, use the masculine as generic, etc.).

Unfortunately, that approache would seemingly increase the complexity of the task, making it much less feasible. Another, more practical approach, and the one we followed eventually, was to care not about classifying a target as feminine, masculine or whatever else but about classifying it as either feminine or non-feminine, including all classes which do not correspond to the feminine gender in one only class, called m for convenience. That is, thus conceived, the task is about to determine which targets are feminine out of the total input.

5.3 Feature extraction

We will present now the patterns in the input data (i.e. discourse patterns) that were selected as the most distinguishing and represented abstractly in the feature vectors that feed the classifier. Some domain-specific knowledge is necessary for a good performance on pattern recognition, so this stage was preceded by a preliminary study of the data (the news articles). For each target, each feature captures a discourse pattern which corresponds to an anaphoric relationship between a cue, which is a gender-specific element, and the target, which is gender-indefinite. These patterns where searched for in the *target context*, and not in the whole text: that is, the context is not a particular number or sentences but instead it consists of all the left co-text up down the previous role and all the following co-text up to the next role.

#AQUI or in CLASSIFICATION?: Of course, the fact that a, say, feminine pronoun precedes a target personal noun, or that the text seems to indicate that a person is married to a woman, is not direct evidence that the person in question is a man (or vice versa for a woman). There is a risk that an individual feature suggests a referential gender which contradicts the actual correct gender but our expectation is that it is probabilistically plausible that the algorithm's overall joint consideration of all the cues in the patterns taken into account leads to the correct gender of the person referred to (in a majority of cases, that is). The actual incidence of each feature in the overall results has not been estimated yet.

A total of twenty features were used in the experiment we are describing. The discourse patterns extracted as features were captured by means of regular expressions — the precise fragments of text matched by the regular expressions are shown in green in the examples. The character in parenthesis next to the feature's name is the indicator of that feature in the vector (#SEE X).

5.3.1 Feature +: supplementary apposition

Apposition, or more precisely supplementary appositives (Huddleston & Pullum 2002: 1357), is an expression (normally an NP but also an AdjP) adjacent to another NP (called the anchor) in the clause, which might function as the subject or as an object. It is semantically related to the anchor but does not depend syntactically on it, so it might be considered a sort of loose attachment but not a syntactic constituent. This is the most obvious and reliable feature we can think of, as both expressions (the appositive and the anchor) co-apply (and not necessarily corefer) and this is a common device of displaying information in newspaper articles in English. The appositive can be a proper personal name adjacent to a personal NP (that is, an NP headed by a personal noun), or vice versa, or even a personal NP adjacent to a personal pronoun. Their place in the sentence seems to be interchangeable (so the appositive can appear immediately before the anchor or immediately after it).

See examples of appositives following the anchor, in the form of proper name followed by a personal NP (see (76-77)):

(76) Ms. Ramirez, the new curator, said her first priority was getting a working collections policy in place.

(77) “A new generation of Muslims is coming into its own,” said

Yvonne Haddad, a professor of history who specializes in women and Islam at Georgetown University.

in the form of a personal NP followed by a proper name (see (78-79)):

(78) The reporter, Judith Miller, published no articles about the agent, Valerie Plame.

(79) The judge, Chiara Nobili of Milan, signed the arrest warrants on Wednesday for 13 C.I.A. operatives who are suspected of seizing an imam named Hassan Mustafa Osama Nasr , also known as Abu Omar, as he walked to his mosque here for noon prayers on Feb. 17, 2003.

or an example of appositive preceding the anchor, in the form of a personal NP followed by a pronoun (80), in the form of an AdjP followed by a pronoun (81) and in the form of an AdjP followed by a proper name (82):

(80) A member of the Russian national team, she was considered a favorite for the European championships Aug. 27-30 in Milan, Italy.

(81) Born in 1930 in Tokyo, she was married to the actor Noboru Nakaya from 1954 until their divorce in 1978."

(82) Born in 1950 in Hindiya, between the cities of Karbala and Hilla, Mr. Maliki was educated in Iraqi Kurdistan and was exiled to Iran after Mr. Hussein vowed to eliminate leaders of the Islamic Dawa Party.

5.3.2 Feature 8: marriage

Among the several relations realizing associative anaphora can be included those which express marital or sexual relations. The expressions *to marry sb* or *to be married to sb* have been used as indicative of the referential gender of the married person, as long as the expression referring to his/her spouse is gender-specific. Although this feature could be based on the assumption that the majority of marriages are not gay marriages, for maximum certainty care is taken to detect a possible reference to homosexuality in the target context. Therefore, gender assigned to the target is f if her spouse is a man and m if his spouse is a woman, unless there is some evidence in the co-text that this is a gay relationship, in which case it is the other way round. See examples (83-84):

(83) Barbara Joan Wax, a former buyer and manager of the S'fari Room at Bonwit Teller in New York, was married yesterday to Associate Justice Bentley Kassal of the Appellate Division of the

Supreme Court of the State of New York, First Department in Manhattan.

(84) She is now married to Frederick Thaufeer al-Deen, a former federal prison chaplain.

(85) “Born in 1930 in Tokyo, she was married to the actor Noboru Nakaya from 1954 until their divorce in 1978.”

5.3.3 Feature B: copula

Under *copula* we include not only *be* clauses, but also other complex-intransitive clauses (Huddleston & Pullum 2002: 218), such as those headed by the verbs *to remain*, *become*, *seem* and constructions as *to grow up into*, *come to be*, *develop into*, *turn into* (in whatever tense).

On the other hand, it is not important here whether the predicative complement is specifying and, therefore, coreferring with the subject, or whether it is ascriptive, i.e. it rather just denotes a property (either realized by a AdjP or an NP) that is predicated of the subject’s referent, and therefore non-referential: see all examples provided (86-89)) (Huddleston & Pullum 2002: 271). The important thing is that, in complex-intransitive clauses, the predicative complement always and the subject always co-apply:

(86) Her father was a vice president of Cohen Goldman, a men ’s clothing manufacturer in New York.

(87) Justice Kassal, a son of the late Mr. and Mrs. Hyman Kassal of North Miami Beach, Fla., and formerly of New York, is a former Democratic member of the New York State Assembly and was a Democratic candidate for the 20th Congressional District in Manhattan.

(88) As obvious from her performance on Sunday afternoon with Mr. Gordeyev ’s group in an excerpt from La Sylphide, she has grown up into a happily exquisite dancer.

(89) Ms. Lewis became chief engineer last June when her boss, a man, went into private business, giving her the lead just as the project reached its hardest phase, building the tiny craft and finding ways to make nebulous plans come to life.

5.3.4 Feature j: poss. + relational nouns denoting human occupation + as + role

A role noun (# EXPLICAR ROLE NOUNS) can be the dependent of a predicative PP headed by the preposition *as* which functions as adjunct of a relational noun[16] which denotes human function: the relation is between an occupation (such as a particular job) and the occupation's holder. RelN(x,y) : predicand(x) & occupation(y) & ISA(x,y). SOMETHING LIKE THAT? The *as*-PP's dependent is interpreted predicatively, the preposition *as* functioning similarly to the verb *be*: in (90), she was an officer and director. The adjunct and the predicand co-apply, so the role noun's referential gender should correspond to the predicand's grammatical gender (if this is gender-specific). This pattern seems quite reliable and has been observed to occur with relative frequency in the texts of our corpus. The present feature captures it when preceded by a possessive pronoun (which is the cue indicating the occupation's holder's referential gender). See examples (90-93):

(90) Her anticipated absence at the helm of her company, where she gave up her roles as an officer and director after her conviction, has already had an impact.

(91) Miss Miller is acting in good faith, doing her duty as a respected and established reporter who believes reporters have a First Amendment privilege that trumps the right of the government to inquire into her sources.

(92)

(93) On Dec. 11, after the appointment of Robert Druskin as chief operating officer of Citigroup, Ms. Bartiromo and Charles Gasparino, a CNBC on-air editor, had a brief on-air clash when Ms. Bartiromo remarked that an earlier report by Mr. Gasparino that Sallie L. Krawcheck would leave her job as chief financial officer did not pan out.

5.3.5 Feature n: name + *as* phrases as transitive predicative adjuncts

We saw in the description of the previous feature (j) that predicatives marked by *as* can function as adjunct. They can also function as the predicative complement of a verb, which might be intransitive (e.g. *She worked as director*) or transitive (e.g. *I appointed her as director*). This feature captures the second case when the object is realized in form of personal proper name, as well as the case

[16] A relational noun is a noun whose meaning involves a relatinship between two entities.

of deverbal nouns derived from a transitive verb which could have the *as* phrase in question as complement (cfr. example (95) with *They appointed Robert Druskin as chief operating officer* or *Robert Druskin was appointed chief operating officer*). In these both cases the role noun's referential gender should correspond to the object's grammatical gender (if it is gender-specific) because both co-apply. See examples (94 and 95):

(94) While supporters saw Justice Priscilla R. Owen as a jurist of high intellect and integrity, opponents say that she was too far right ideologically.

(95) On Dec. 11, after the appointment of Robert Druskin as chief operating officer of Citigroup, Ms. Bartiromo and Charles Gasparino, a CNBC on-air editor, had a brief on-air clash when Ms. Bartiromo remarked that an earlier report by Mr. Gasparino that Sallie L. Krawcheck would leave her job as chief financial officer did not pan out.

5.3.6 Feature 1: rough *as*

This feature captures any pattern in which any kind of cue is linked by the preposition *as* to a personal noun within the same orthographic sentence. We saw why copulas are relevant patterns and we saw that the preposition *as* functions as an analogue of the verb *be*, so this feature deserves no further justification. The adjective *rough* refers to the fact that this pattern is very broad in matching.

(96) nytimes_19870220_mjudge.txt.xml_wi_classes.xml:
Mrs. Whitehead took the stand after Mrs. Hergenhan, who described herself as Mrs. Whitehead's best friend.

(97) And now Judge Campbell has appointed Karen Tomberlin as Ms. Kowalski 's guardian.

5.3.7 Feature 2: possessive + partner

This feature captures the pattern existing between a possessive pronoun (the cue) and a gender-indefinite relational personal noun which denotes a sentimental, marital or sexual relationship, such as *partner*, *mate*, *lover*, *beloved*, *spouse*, *companion*, etc. Some noise might be captured by the pattern, as some of these nouns are polysemous, e.g. mate or companion, so further experimentation is needed. Here, too, the referential gender assigned to the target is f if the possessive is masculine and m if the possessive is feminine, unless there is some evidence in the co-text that this is a gay relationship.

(98) Her partner and husband, the 22-year-old Andrei Ryabov, is a real find: very tall, very blond, a true danseur noble but with a new contemporary boldness that avoids the old-fashioned mien still found among some Soviet male dancers.

(99) After a five-year legal battle in Minnesota between the parents of a paralyzed, brain-damaged woman and her lesbian lover, a judge has chosen a third party to act as the disabled woman's legal guardian.

5.3.8 Feature x: gender-specific modifier

This feature captures the pattern existing between a gender-specific modifier (the cue), such as sex-related nouns and adjectives (e.g. *gay*, *lesbian*, *female*, etc.), and a gender-indefinite relational personal noun which denotes a sentimental, marital or sexual relationship, as in the previous feature (2). IS THIS ASSOCIATIVE ANAPHORA? IN WHAT WAY DOES THIS PATTERN CONTRIBUTE TO COHESION? See examples (100), from our corpus, and (101), from a blog:

(100) After a five-year legal battle in Minnesota between the parents of a paralyzed, brain-damaged woman and her lesbian lover, a judge has chosen a third party to act as the disabled woman's legal guardian.

(101) The former Hewlett-Packard chief executive Carleton S. Fiorina, once the most prominent female executive in the United States, ordered the first of a series of leak investigations into contacts by board members with journalists in January 2005, she says in a long-anticipated memoir.

Some further research is needed here to as to exploit the connotational meaning of adjectives such as *beautiful*, *pretty*, *lovely*, *cute*, *delicate*, *dashing*, *charismatic*, *charming*, etc. which might be gender-indefinite denotationally, or other such as *boyish*, *virile*, *stud*, etc. which might denote gender-specific properties but which can be applied to referents of any gender.

5.3.9 Feature O: orthographic overlapping

The feature accounts for the fact that a particular personal noun can appear as target in the target context (see XX: role_cntxt) in which it also appears as part of a cue, typically as a PROFESSIONAL TITLE in a named entity (preceding the person's name) which has been previously interpreted (see XX). If this pattern occurs, the gender of the cue expression containing the title is assigned to the target.

(102) After a five-year legal battle in Minnesota between the parents of a paralyzed, brain-damaged woman and her lesbian lover, a judge has chosen a third party to act as the disabled woman 's legal guardian. In his ruling in the case Judge Robert Campbell of the St. Louis County District Court in Duluth, Minn., likened the paralyzed woman, Sharon Kowalski, now 34 years old, to a child over whom divorcing parents do battle.

(103) A federal district judge has ruled that Monica Lewinsky did not have an agreement with the independent counsel that would give her immunity from prosecution, lawyers involved in the inquiry said on Wednesday. The ruling, filed under seal by Judge Norma Holloway Johnson , is a setback for Ms. Lewinsky, who denied under oath that she had a sexual relationship with President Clinton but talked about having such a relationship and efforts to conceal it in conversations recorded by a former colleague, according to people who have heard the tapes .

5.3.10 Feature J: justice: in his ruling

This feature accounts for a very specific associative anaphora pattern: the relation between a representative of a court and the court order she or he makes, the *ruling*. This pattern would of course be generalized to other similar relations from other professional fields. There were no occurrences of this pattern with nouns, so this feature is always empty in the vectors obtained, but some occurrences were found with named entities, which could be used to interpret them or confirm their earlier interpretation.

(104) In his ruling in the case Judge Robert Campbell of the St. Louis County District Court in Duluth, Minn., likened the paralyzed woman, Sharon Kowalski, now 34 years old, to a child over whom divorcing parents do battle.

#JUDGE ROBERT CAMPBELL: AN EXPRESSION CAN BE TARGET FIRST AND CUE AFTERWARDS. EXPLAIN

5.3.11 Feature ʃ: reported speech

This feature is based on the not tested assumption, or intuition from observing the texts, that people, in their statements, refer to themselves more often than to other people. So, the pattern captured in this feature is …*somebody (target) said she/he (cue)* … and it is expected that the pronoun after the reported speech verb corefers with the previous personal noun. One drawback of this feature is that it is perfectly plausible that one person refers to someone else when speaking, e.g. "Her son said she left for New York …," so there remains the need to see how effective this feature is in contributing towards an informative vector.

(105) Speaking from Atlanta, where she is attending a national lesbian rights conference, Ms. Thompson, an associate professor of physical education, recreation and sports science at St. Cloud State University, said she would appeal Judge Campbell's decision.

(106) The judge said she first learned that Clinton had given false testimony in the Jones suit when Clinton went on television that night to admit to having given "misleading" answers to Ms. Jones's lawyers.

(107) Mrs. Rabichow, a Bill Bradley supporter, said she did not know Mr. Gore's views on her critical issues, Medicare and prescription drugs.

5.3.12 Feature r: next adjoining (#CONTIGUO) pronoun

This and the next seven features aim to capture the pattern containing a cue and a target, in that order of viceversa. In this case, the pattern including the target and a succeeding pronoun within the target context is only matched as long as there is no other cue or target between it and the target. This feature (and the next seven) could be broken down in two subfeatures: one for nominative pronouns (e.g. *she*) and one for accusative pronouns (e.g. *her*).

(108) And as a former director for museum programs at the United States Holocaust Memorial Museum in Washington, she understands the complexities involved in presenting indescribably horrific events in recent history.

(109) In the interview, which was also shown Monday on MSNBC, the second dancer called the defense lawyers' comments about her testimony " out-and-out lies.

5.3.13 Feature R: previous adjoining (#CONTIGUO) pronoun

This feature captures the pattern containing a target and a preceeding pronoun within the target context as long as there is no other cue or target between it and the target.

(110) Her father was a vice president of Cohen Goldman, a men's clothing manufacturer in New York. #CONTRADICTORY

(111) He was designated an associate justice of the New York Court of Appeals for April 1985.

(112) The disclosure appeared to stun the judge, Harvey R. Sorkow, and sparked a series of revelations, including Mrs. Whitehead's acknowledging that she had been aware that such a letter was being written in her name and the friend's admission that she had lied in testimony earlier today. The friend, Susan Hergenhan, made the admission in response to a question from Judge Sorkow.

(113) There is more than curiosity value here, although viewers who saw the American film about the Kirov Ballet school entitled "The Children of Theater Street" (1978) will be pleased to learn that its real-life little heroine, Anzhelina Armeyskaya, has lost none of the charm she had as a child.

(114) nytimes_19901024_armeyskaya.txt.xml_wi_classes.xml: Her partner and husband, the 22-year-old Andrei Ryabov, is a real find: very tall, very blond, a true danseur noble but with a new contemporary boldness that avoids the old-fashioned mien still found among some Soviet male dancers. CONTRADICTORY WITH FEATURE (2)

5.3.14 Feature s: next rough pronoun

This feature captures the pattern, within the target context, containing a target and a following pronoun, whatever there is between the two. By rough we mean that this pattern is searched without any attempt at precision. This kind of "rough" features, as well as the "adjoining" ones although to a lesser extent, do not aim at any particular target and do not rely on any particular assumption, other than the chance of coincidence that such pattern includes predominantly expressions which apply to same-gender referents.

BUSCAR EXEMPLOS COM OUTRAS ENTIDADES POLO MEDIO

(115) nytimes_20060428_marijuan_21marijuana.txt.xml_wi_classes.xml: Lyle E. Craker, a professor in the division of plant and soil sciences at the University of Massachusetts, said he submitted an application to the D.E.A. in 2001 to grow a small patch of marijuana to be used for research because government-approved marijuana, grown in Mississippi, was of poor quality. BUSCAR EXEMPLOS FEMININOS:

(116) nytimes_20060427_grandjur_27inquire.txt.xml_wi_classes.xml: According to court documents filed by the prosecutor, Mr. Libby has testified that Mr. Cheney told him the president had authorized the release of some information contained in a classified National Intelligence Estimate chronicling the efforts of Saddam Hussein to purchase uranium from Niger.

(117) nytimes_19910426_mjudge-lesbians.txt.xml_wi_classes.xml: After a five-year legal battle in Minnesota between the parents of a paralyzed, brain-damaged woman and her lesbian lover, a judge has chosen a third party to act as the disabled woman's legal guardian. In his ruling in the case Judge Robert Campbell of the St. Louis County District Court in Duluth, Minn., likened the paralyzed woman, Sharon Kowalski, now 34 years old, to a child over whom divorcing parents do battle. CONTRADICTORY WITH (SEE EXAMPLE WHICH SAYS (LESBIAN) LOVER IS FEM)

5.3.15 Feature S: previous rough pronoun

This feature captures the pattern, within the target context, containing a pronoun which precedes the target, whatever there is between the two.

DUDA: PERGUNTAR: NURIA

```
<t:role[^>]*>$roles[$i]<\/t:role>.*?<c:pron gen="(.)".+?>.+?>
```

(118) nytimes_20060422_shiites_m_22iraq.txt.xml_wi_classes.xml: He said he told Mr. Maliki on Friday of the importance of independent ministers, and of the need to eliminate militias from the Iraqi security forces and to scale back the Shiite-backed program that bars former Baathists from many jobs in the Iraqi government and institutions. "I met with him for a long time today and he was very positive," Mr. Khalilzad said. Born in 1950 in Hindiya, between the cities of Karbala and Hilla, Mr. Maliki was educated in Iraqi Kurdistan and was exiled to Iran after Mr. Hussein vowed to eliminate leaders of the Islamic Dawa Party. Mr. Maliki later traveled to Syria and returned to Iraq after the American invasion three years ago. "He is very close to Jaafari," said the acting Assembly speaker, Adnan Pachachi.

(119) nytimes_20060314 -moussaoui.txt.xml.xml: "It's not the Justice Department against Mr. Moussaoui," the judge said, "it's the United States" that is trying to have him executed.

Ms. Martin, appearing stricken and dressed in black, looked like a mourner at a funeral having trouble keeping her composure. When Judge Brinkema said she faced possible civil or criminal contempt charges Ms. Martin said she had not yet been able to contact her lawyer.

5.3.16 Feature p: next adjoining (#CONTIGUO) name or role

This feature captures the pattern containing a target and a following non-pronominal cue as long as there is no other cue or target between it and the target.

(120) nytimes_20060422_kiss_22sat4.txt.xml_wi_classes.xml:
Doctors, nurses, therapists and sex educators sued, and the federal judge, Thomas Marten, held that Mr. Kline's opinion violated the actual language of the underlying state statute, which gives those treating adolescents discretion to decide whether illegal sexual activity amounts to actual child abuse. Kansas law prohibits intercourse, oral sex and lewd touching by anyone under 16.

(121) nytimes_20060314_fjudge-moussaoui.txt.xml_wi_classes.xml:
Mr. Korman, a security official with the Transportation Security Administration, was listed to be called as a hostile witness for the defense. Judge Brinkema also appeared nonplussed when learning that two other witnesses had spoken to a prosecutor on the case simultaneously.

(122) nytimes_20060313_fjudge.txt.xml_wi_classes.xml:
But the agent, Michael Anticev, acknowledged under cross-examination that the bureau had indeed known of earlier Al Qaeda plans to fly planes into the C.I.A. headquarters in Langley, Va., and into the Eiffel Tower.

5.3.17 Feature P: previous adjoining (#CONTIGUO) name or role

This feature captures the pattern containing a target and a preceeding non-pronominal cue as long as there is no other cue or target between it and the target.

(123) nytimes_20060418 _18duke.txt.xml.xml:
The woman, a 27-year-old single mother of two and a student at a nearby university, is black. COORDINATION

(124) nytimes_20050929 -fNOTjudge.txt.xml.xml:
Ms. Miers, 60, was the first woman to become a partner at a major Texas law firm and the first woman to be pre sident of the State Bar of Texas. At one point, Ms. Miers was Mr. Bush's personal lawyer.

(125) nytimes_20040601.txt.xml.xml:
The governing People's Action Party ratified the appointment of Lee Kuan Yew's son as the next prime minister, cementing the family's role in shaping the city-state's destiny.

5.3.18 Feature q: next rough name or role

This feature captures the pattern containing a target and a non-pronominal cue, in that particular order, regardless of what there is between the two.

```
<t:role[^>]*>$roles[$i]<\/t:role>.*?<c:(person|role) gen="(.)"
```

(126) nytimes_20060428 _21marijuana.txt.xml.xml:
Lyle E. Craker, a professor in the division of plant and soil sciences at the University of Massachusetts, said he submitted an application to the D.E.A. in 2001 to grow a small patch of marijuana to be used for research because government-approved marijuana, grown in Mississippi, was of poor quality. In 2004, the drug enforcement agency turned Dr. Craker down.

(127) nytimes_20060427 _27inquire.txt.xml.xml: To date, the only criminal charges in the case, which involves the exposure of a C.I.A. operative's identity, have been brought against I. Lewis Libby Jr., the former chief of staff to Vice President Dick Cheney, who was charged with lying and obstruction last November and is preparing for trial.

(128) nytimes_19920821.txt.xml.xml: Natalya Ivanova, a Russian champion water skier, was stabbed to death during a competition near Moscow, a news agency reported today. Ivanova, 21 years old, had just finished her program Tuesday in trick riding when she was attacked by a man with whom she apparently had argued earlier, the ITAR-Tass news agency said. The man was later arrested, it said.

5.3.19 Feature Q: previous rough name or role

This feature captures the pattern containing a non-pronominal cue and a target, in that particular order, regardless of what there is between the two.

(129) nytimes_20060303 -nurse_03cullen.txt.xml.xml:

"We are no longer the Shanaghers. We are that poor family where that nurse killed their father." "There is part of me that would like to see you put to death," Mr. Shanagher continued. "A part that would appreciate the irony of you dying by lethal injection." Mr. Cullen, 46, was a career nurse with an icy bedside manner and a history of mental illness who worked at nine hospitals and one nursing home in New Jersey and Pennsylvania over the course of 16 years.

(130) nytimes_20050929 -fNOTjudge.txt.xml.xml:
In 1995, Mr. Bush, then governor of Texas, named her chairwoman of the Texas Lottery Commission and gave her the task of cleaning up that scandal-plagued agency. Ms. Miers has never been a judge, although that is not a requirement for a Supreme Court justice.

(131) nytimes_20050929 -fNOTjudge.txt.xml.xml:
Influential Republicans said there was a serious possibility that Mr. Bush would name a woman or a minority candidate to succeed Justice O'Connor, particularly after the president said Monday, in response to a question about how close he was to choosing a successor, that "diversity is one of the strengths of the country."

5.3.20 Feature &: coordination

Coordination is a relation between two or more syntactically equivalent expressions. This feature aims to account for the conjunctive binary coordination between two personal nouns, one of which is gender-specific, and functions as a cue that can be exploited to the interpret the other one, the target.

(132) nytimes_19901024.txt.xml.xml:
Her partner and husband, the 22-year-old Andrei Ryabov, is a real find: very tall, very blond, a true danseur noble but with a new contemporary boldness that avoids the old-fashioned mien still found among some Soviet male dancers.

(133) In other developments on Wednesday in the wide-ranging investigation of the Whitewater independent counsel, a lawyer representing the president and first lady confirmed that Hillary Rodham Clinton had rebuffed two questions posed to her by Whitewater prosecutors in a five-hour session at the White House on Saturday.

(134) Ms. Miers, 60, was the first woman to become a partner at a major Texas law firm and the first woman to be president of the State Bar of

Texas. At one point, Ms. Miers was Mr. Bush's personal lawyer.

5.4 Classification

PENDENT D'ACABAR......................

The task of the classifier is to use the feature vector provided in the previous stage to assign a personal noun to a class (feminine) or not. The first experiment was run when the number of features reached 20 and the search patterns had been tested and superficially refined.

to avoid noise: binary classification (nuo -> m) . CLASSIFICATION: CLASSES? initially: f, m, u, o, n. But being practical the number of classes is reduced to 2: f and !f / m (everything which is not f).

The vectors are processed with Weka... The algorith was...
The ratio of correctly classified instances was 73.35% Precision and recall... TO BE CALCULATED

Dispersion: most of the features are not present in all vectors corresponding to a same entity. ??? POR VER.

Results with DT, with NaiveBayes and with BayesADOBE?

Results with nou removed or not (f+monu vs. f+m).

Analysis with SPSS: principal components analysis ?

Ensinar curva de aprendizagem (margin curve?) para demostrar que com mais textos a accuracy (EQUIVALENTES DE TODOS OS TERMOS PARA A APRESENTAÇOM!) subiria mui pouco, ou seja, que nom fam falta muitos mais textos. É melhor jogar com features (traços): acrescentar, tirar uns, tirar outros, etc. Ir jogando… e ver que resultados dá. Tirar todos os da direita e compara com tirando todos os da esquerda.

Analisar erros clicando nas cruzezinhas que aparecen nos 4 cantos.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     gender-weka.filters.unsupervised.attribute.Remove-R1
Instances:    274
Attributes:   21 (+ 8 B j n 1 2 x O J ʃ r R s S p P q Q & gender)
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------
```

```
Q = f
|   + = f: f (57.97/16.91)
|   + = m
|   |   q = f: f (15.7/6.89)
|   |   q = m: m (28.92/10.86)
|   |   q = ?: m (0.0)
|   + = ?: f (0.0)
Q = m
|   R = f
|   |   q = f: f (30.07/10.04)
|   |   q = m: m (36.79/11.45)
|   |   q = ?: m (0.0)
|   R = m: m (104.55/12.79)
|   R = ?: m (0.0)
Q = ?: m (0.0)

Number of Leaves  :         11

Size of the tree :         16


Time taken to build model: 0.11 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         201               73.3577 %
Incorrectly Classified Instances        73               26.6423 %
Kappa statistic                          0.4053
Mean absolute error                      0.378 
Root mean squared error                  0.4223
Relative absolute error                 79.9246 %
Root relative squared error             86.857  %
Total Number of Instances              274     

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.514     0.13       0.711     0.514     0.597    f
  0.87      0.486      0.742     0.87      0.801    m

=== Confusion Matrix ===

   a   b   <-- classified as
  54  51 |   a = f
  22 147 |   b = m
```

5.5 Evaluation OR Main problems

PENDENTÍSSIM D'ACABAR. AQUESTES NOTES SÓN DE L'ANY DE LA POLKA. PUEDE QUE HOY EN DÍA ALGUNOS YA NO TENGAN SENTIDO...

1. First problem: Annotation of personal names (recognize named entities and correctly classify them as representing a person) and personal nouns -> gazetteers 2. Dificuldade ou impossibilidade de distinguir umha referring expression from u / n / o. 3. There's not always a pattern that can be used to infer gender (EMPTY VECTOR). examples: one word alone in the desert… a full vector with no feature. -> examples 4. When there is one: it is not always easy to capture it somehow through concordance features (compute how many vectors are empty whatsoever out of the total) 5. The correct class must be assigned manually so as to have a training corpus = tedious task

WHY It is difficult to produce patterns that match the contexts which have been represented by empty vectors without being so general that it might match (many) other things. For example, it is difficult to avoid appositions in order to link a subject and its copula:

> Justice Kassal, a son of the late Mr. and Mrs. Hyman Kassal of North Miami Beach, Fla., and formerly of New York, is a former Democratic member of …

A strict, precise expression that only matches a specific pattern is likely to have a high precision but a low coverage, whereas a more relaxed expression might have a higher coverage but also might match a lot of noise. This problem might be avoided by using some functional analysis and annotation, which marks what is the subject and what is the object or copula.

HOW TO FACE On the other hand, this problem CAN BE / HAS BEEN handled in a second turn, using the formal coincidence as another feature: if the personal noun /role uninterpreted is the same as another which has been interpreted, the former will subsume the latter's gender information. For example: text nytimes_19870220_mjudge.txt.xml_wi_classes.xml:
the first, third and forth occurrences of the personal noun "judge" remain uninterpreted (with an empty or nearly empty vector) after the first round, but the second occurence does have some gender information that can be extrapolated to all homonymic occurrences. This is not 100% certain because there might very well be the case that not all occurrences of a personal noun refer to the same person or to people with the same gender, but IT'S WORTH TRYING IT.

Another example of the same thing: the first and the second occurrences of "friend" in the article are not interpretable in themselves but a third occurrence happens later which is easily interpretable by apposition:

> **TRIAL HEARS WHITEHEAD DID NOT WRITE TO JUDGE**
>
> An emotional 11-page plea for understanding signed with Mary Beth Whitehead's name and sent last May to the judge in the Baby M case was not written by Mrs. Whitehead, but by a close friend, testimony revealed today.
>
> The disclosure appeared to stun the judge, Harvey R. Sorkow, and sparked a series of revelations, including Mrs. Whitehead's acknowledging that she had been aware that such a letter was being written in her name and the friend's admission that she had lied in testimony earlier today.
>
> The friend, Susan Hergenhan, made the admission in response to a question from Judge Sorkow.

(Second round: first round, application of results, and second round) ANALYSIS OF THE ERRORS THAT I'VE GOT (NURIA DIXIT).

Chapter 6

Conclusions

PENDENT D'ACABAR........

The problem stated in Section #X has been solved: as shown in Sections XX to XX, an algorithm capable of handling gender interpretation in machine translation has been developed. The principal mechanism needed in the algorithm is machine learning bla bla.

6.1 Proposal of future research for the thesis

Plural targets referring to several known individuals (all of them)

Enlarge the targets to substantivized adjectives (fused heads?): The youngest is my daughter.http://www2.parl.gc.ca/HousePublications/Publication.aspx?DocId=1979399&Language=E&Mode=1&Parl=38&Ses=1

Provar-lo amb textos d'altres dominis? (weakness)

FUTURE WORK: refinar patterns to reduce noise and increase precision

FUTURE WORK: run experiments to see whether performance increases or decreases without gazeteers, with different optimizations, etc..

6.1.1 Recurrent people's names database

It would be too silly not to assign a particular grammatical gender to the NP "the British Prime Minister" when referring to a person called Blair just because there is no formal mentioning of his forename, his title ('Mr') or any gendered pronouns.

A database storing the names of known people who are recurringly and frequently mentioned or quoted in the media could contribute to solve this problem. Unlike the personal names database, this database should store information about the gender of the actual person normally associated with a particular full name (forename + pa¨tronimic + surname, etc.).

6.1.2 Exploitation of images in web pages

Diego Alberto.

Patter classification techniques could be succesfully applied to the recognition of traits typically associated with a particular biological sex or referential gender. These tecniques could be helpful in resolving the gender of a human entity invoked in the text when (a) no other explicit reference is made to its referential gender and (b) when there is a caption which names explicitly the person appearing in the photograph. e.g http://news.bbc.co.uk/sport2/hi/cricket/4474340.stm

```
<div>
        <img alt="Raed Juhi on 13 June 2005"
                src="http://newsimg.bbc.co.uk/media/images/41062000/jpg/
                _41062674_juhi203bafp.jpg" />
        <div class="cap">
                There is tight protection for the officials involved
                in the case
        </div>
</div>
```

Spot possible problems of overfitting by processing a corpus from another source or of a different nature.

Appendices

Appendix A

```
TiMBL 5.1.0 (release) (c) ILK 1998 - 2004.
Tilburg Memory Based Learner
Induction of Linguistic Knowledge Research Group
Tilburg University / University of Antwerp
Wed Sep 26 16:03:51 2007

Examine datafile 'names.test+train.data' gave the following results:
Number of Features: 9
InputFormat       : C4.5

Phase 1: Reading Datafile: names.test+train.data
Start:          0 @ Wed Sep 26 16:03:51 2007
Finished:   18906 @ Wed Sep 26 16:03:51 2007
Calculating Entropy         Wed Sep 26 16:03:51 2007
Lines of data     : 18906
DB Entropy        : 0.99189621
Number of Classes : 2

Feats   Vals    InfoGain        GainRatio
    1     17    0.0024961989    0.00088632920
    2      2    8.2332444e-06   1.0753919e-05
    3      2    0.11702617      0.11709155
    4     28    0.0098049550    0.0022193321
    5     34    0.0070420229    0.0019288355
    6     43    0.0088213329    0.0020599866
    7     36    0.023593689     0.0056457213
    8     40    0.085982816     0.021102564
    9     33    0.23606775      0.062543550

Feature Permutation based on GainRatio/Values :
< 3, 9, 8, 7, 4, 5, 1, 6, 2 >
Phase 2: Learning from Datafile: names.test+train.data
Start:          0 @ Wed Sep 26 16:03:51 2007
Finished:   18906 @ Wed Sep 26 16:03:52 2007
```

```
Size of InstanceBase = 101408 Nodes, (2028160 bytes), 44.10 % compression


Starting to test using Leave One Out
Writing output in:          names.test+train.data.LOO.O.gr.k5.out
Algorithm     : LOO
Global metric : Overlap
Deviant Feature Metrics:(none)
Weighting     : GainRatio
Feature 1        : 0.000886329201614
Feature 2        : 0.000010753919090
Feature 3        : 0.117091554135355
Feature 4        : 0.002219332086882
Feature 5        : 0.001928835475732
Feature 6        : 0.002059986637441
Feature 7        : 0.005645721281619
Feature 8        : 0.021102563900776
Feature 9        : 0.062543549558873

Tested:      1 @ Wed Sep 26 16:03:52 2007
Tested:      2 @ Wed Sep 26 16:03:52 2007
Tested:      3 @ Wed Sep 26 16:03:52 2007
Tested:      4 @ Wed Sep 26 16:03:52 2007
Tested:      5 @ Wed Sep 26 16:03:52 2007
Tested:      6 @ Wed Sep 26 16:03:52 2007
Tested:      7 @ Wed Sep 26 16:03:52 2007
Tested:      8 @ Wed Sep 26 16:03:52 2007
Tested:      9 @ Wed Sep 26 16:03:52 2007
Tested:     10 @ Wed Sep 26 16:03:52 2007
Tested:    100 @ Wed Sep 26 16:03:52 2007
Tested:   1000 @ Wed Sep 26 16:03:52 2007
Tested:  10000 @ Wed Sep 26 16:03:55 2007
Ready:   18906 @ Wed Sep 26 16:03:57 2007
Seconds taken: 5 (3781.20 p/s)

F-Score beta=1, microav: 0.771852
F-Score beta=1, macroav: 0.768681
AUC, microav:            0.767453
AUC, macroav:            0.767453
overall accuracy:        0.772559  (14606/18906), of which 1532 exact matches

Confusion Matrix:
             m      f
        --------------
     m |   8527   1927
     f |   2373   6079
   -*- |      0      0


There were 655 ties of which 363 (55.42%) were correctly resolved
```

Bibliography

Abe, N. (2007). Japanese Baby Names — About.com: Japanese Language http://japanese.about.com/library/weekly/aa042901a.htm. [Online; accessed 24 July 2007].

Adam, A. (1998). *Artificial Knowing. Gender and the Thinking Machine*. London and New York: Routledge.

Anderson, J. M. (2007). *The Grammar of Names*. Oxford: Oxford University Press.

Bailey, R. (2004). Talking about Words: Presidentess. *Michigan Today - University of Michigan News Service* http://www.umich.edu/news/MT/NewsE/01_04/words.html. [Online; accessed 2-Sep-2007].

Baker, M. (1992). *In Other Words*. New York: Routledge.

Baker, M. (ed.) (1998). *Routledge Encyclopedia of Translation Studies*. London and New York: Routledge.

Braun, F. (2001). The communication of gender in Turkish. *The linguistic representation of women and men* **I**. 283--310.

Brown, G. & Yule, G. (1983). *Discourse analysis*. Cambridge: Cambridge Textbooks in Linguistics.

Campbell, M. (2007). Behind the name. http://www.behindthename.com. [Online; accessed 20-Jul-2007].

Corbett, G. G. (1991). *Gender*. Cambridge: Cambridge University Press.

Corbett, G. G. (2006). *Agreement*. Cambridge: Cambridge University Press.

Craig, C. G. (1994). Classifier languages. In Asher, R. E. (ed.), *The Encyclopedia of Language and Linguistics*, Oxford: Pergamon. vol. 2, 565--569.

Cunningham, H., Maynard, D., Bontcheva, K. & Tablan, V. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th anniversary meeting of the association for computational linguistics*.

Curzan, A. (2003). *Gender Shifts in the History of English*. Cambridge: Cambridge University Press. `http://assets.cambridge.org/97805218/20073/sample/9780521820073ws.pdf`.

Daelemans, W., Zavrel, J., Van der Sloot, K. & Van den Bosch, A. (2007). TiMBL: Tilburg Memory Based Learner, version 6.0, Reference Guide. Technical Report Technical Report Series no. 07-03, ILK Research Group.

Dari, L. (2006). Chinese personal names. *The Indexer* **25**. C1--C7.

Feliu i Cortès, J. (2004). *Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica*. Ph.D. dissertation, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona. `http://www.tdx.cesca.es/TESIS_UPF/AVAILABLE/TDX-0520104-111213//tjfc1de1.pdf`.

Foley, W. A. & van Valin, R. D., Jr (1984). *Functional Syntax and Universal Grammar*. Cambridge: Cambridge University Press.

Frank, A., Hoffmann, C. & Strobel, M. (2004). Gender issues in machine translation `http://yltis.com/Publikationen/GIST.pdf`. Univ. Bremen.

Geotravel (2007). *Culture Briefing: Vietnam*. Kissimmee, Florida: Geotravel Research Center.

Gernsbacher, M. A. (1990). *Language Comprehension As Structure Building*. Hillsdale, New Yersey: Lawrence Erlbaum Associates.

Goldstein, N. (ed.) (2000). *The Associated Press Stylebook and Briefing on Media Law*. Cambridge, Massachusetts: Perseus Publishing.

Grice, H. P. (1975). Logic and conversation. In Cole, P. & Morgan, J. L. (eds.), *Syntax and Semantics, 3: Speech Acts*, New York: Academic Press. 41--58.

Hall, K. & Buchotz, M. (1995). *Gender articulated: Language and the socially constructed self*. New York: Routledge.

Halliday, M. & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hamilton, M. C. (1991). Masculine bias in the attribution of personhood. people = male, male = people. *Psychology of Women Quaterly* **15**. 393--402.

Hedden, H. (2007). Arabic names. *The Indexer* **25**. C9--C16.

Hellinger, M. & Bußmann, H. (2001). Gender Across Languages. *The linguistic representation of women and men* **I**. 1--25.

Hirschman, L. & Chinchor, N. (1997). MUC-7 Coreference Task Definition. Version 3.0. Technical report, National Institute of Standards and Technology. `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/co_task.html`.

Huddleston, R. & Pullum, G. K. (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Ibrahim, M. H. (1973). *Grammatical Gender. Its Origins and Development*. The Hague: Mounton.

Irmen, L. (2007). What's in a (Role) Name? Formal and Conceptual Aspects of Comprehending Personal Nouns. *Journal of Psycholinguist Research* **Online first**. `http://www.springerlink.com/content/d74p72671m6x34q1/`.

Jeff (2003). Chinese surnames. `http://www.chinaculture.org/gb/en_chinaway/2006-03/24/content_80519.htm`. [Online: accessed 18-july-2007].

Jordan, L. (ed.) (1976). *The New York Times Manual of Style and Usage: a Desk Book of Guidelines for Writers and Editors*. New York: The New York Times Company.

Jurafsky, D. & Martin, J. H. (2000). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. New Jersey: Prentice Hall.

Karlsson, F. (1999). *Finnish: An Essential Grammar*. London and New York: Routledge.

Kessler, S. J. & McKenna, W. (1978). *Gender: An Ethnometodological Approach*. The University of Chicago Press.

Koller, W. (1989). Equivalence in translation theory. In Chesterman, A. (ed.), *Readings in Translation Theory*, Helsinki: Loimaan Kirjapaino Oy.

Loos, E. E. (2003). Glossary of linguistic terms. `http://www.sil.org/linguistics/glossaryOflinguisticTerms`.

Lyons, J. (1968). *Introduction to Theoretical Linguistics*. Cambridge: Cambridge University Press.

Lyons, J. (1977). *Semantics*, vol. 1. Cambridge: Cambridge University Press.

Maas, S., Rommes, E., Schirmer, C. & Zorn, I. (2007). Gender Research and IT Contruction: Concepts for a Challenging Partnership. In *Gender Designs IT. Construction and Deconstruction of Information Society Technology*, Wiesbaden: VS Verlag für Sozialwissenschaften.

McKeown, K. R. & Radev, D. R. (2000). Collocations. In Date, R., Moisi, H. & Somers, H. (eds.), *Handbook of Natural Language Processing*, New York: Marcel Dekker. 507--523.

McShane, M., Zacharski, R., Nirenburg, S. & Beale, S. (2005). The Boas II Named Entity Elicitation System. http://ilit.umbc.edu/ILIT_Working_Papers/ILIT_WP_08-05_Boas_II.pdf. [Online; accessed 12-Dec-2006].

Miller, G. A. (2007). patronymic. *WordNet* **3.0**. http://dictionary.reference.com/browse/patronymic. [Online; accessed: July 26, 2007].

Mitkov, R. (2002). *Anaphora Resolution*. London: Pearson Education.

Moore, A. W. (ed.) (1993). *Meaning and Reference*. Oxford: Oxford University Press.

Moravcsik, E. A. (1978). Agreement. In Greenberg, J. H., Ferguson, C. A. & Moravcsik, E. A. (eds.), *Universals of Human Language: IV: Syntax*, Stanford: Stanford University Press. 331--374.

Moravcsik, E. A. (1988). Agreement and markedness. In Barlow, M. & Ferguson, C. A. (eds.), *Agreement in Natural Language: Approaches, Theories, Descriptions*, Stanford: CSLI. 89--106.

Mulkern, A. E. (1996). The Game of the Name. In Fretheim, T. & Gundel, J. K. (eds.), *Reference and Reference Accessibility*, Amsterdam and Philadelphia: John Benjamins Publishing. 235--250.

Nelson, R. J. (1992). *Naming and Reference*. London: Routledge.

Ng, H. T., Zhou, Y., Dale, R. & Gardiner, M. (2005). A machine learning approach to identification and resolution of one-anaphora. In *Proceedings of IJCAI'2005*. 1105--1110. http://ijcai.org/papers/1582.pdf.

Omar, M. K. (1975). *Saudi Arabic. Urban Hijazi Dialect*. Washington, D. C.: Foreign Service Institute, Department of State (US). http://fsi-language-courses.com/Courses/Arabic/Saudi%20Arabic%20Basic%20(Urban%20Hijazi%20Dialect)/FSI%20-%20Saudi%20Arabic%20Basic%20Course%20(Urban%20Hijazi%20Dialect)%20-%20Student%20Text.pdf.

Real Academia Española (2001). obispo. *Diccionario de la lengua española 21^{a} ed.* http://buscon.rae.es/draeI/SrvltConsulta?TIPO_BUS=3&LEMA=obispo.

Real Academia Española (2005). Género2. *Diccionario panhispánico de dudas* http://buscon.rae.es/dpdI/SrvltGUIBusDPD?lema=g%E9nero2.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees http://citeseer.ist.psu.edu/schmid94probabilistic.html.

Schmid, H. (1995). Improvements in Part-of-Speech Taggin With an Application to German http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger2.pdf.

Siewierska, A. (2004). *Person*. Cambridge: Cambridge University Press.

Silveira, J. (1980). Generic masculine words and thinking. In Kramarae, C. (ed.), *The voices and words of women and men*, Oxford: Pergamon. 165--178.

Stahlberg, D., Braun, F., Irmen, L. & Sczesny, S. (2007). Representation of the sexes in language. In Fiedler, K. (ed.), *Social communication. a volume in the series frontiers of social psychology*, New York: Psychology Press. 163--187.

Steele, S. (1978). Word order variation: a typological study. In Greenberg, J. H., Ferguson, C. A. & Moravcsik, E. A. (eds.), *Universals of Human Language: IV: Syntax*, Stanford: Stanford University Press. 585--623.

Stewart, J. (2002). Information Society, the Internet and Gender: A Summary of Pan-European Statistical Data. Technical Report D02_Part2. http://www.rcss.ed.ac.uk/sigis/public/D02/D02_Part2.pdf. [Online: accessed 29-07-2006].

Stryker, S. (2004). Transgender http://www.glbtq.com/social-sciences/transgender.html.

Susam-Sarajeva, î (2005). A Course on 'Gender and Translation'. As an Indicator of Certain Gaps in the Research on the Topic. In Santaemilia, J. (ed.), *Gender, Sex and Translation. The Manipulation of Identities*, Manchester: St. Jerome.

ThingsAsian.com (2007). Vietnamese Names — ThingsAsian.com http://www.thingsasian.com/stories-photos/1044. [Online; accessed 24 July 2007].

van den Bosch, A. (2007). TiMBL: Tilburg Memory Based Learner, version 6.0, API Guide. Technical Report Technical Report Series no. 07-05, ILK Research Group.

van Nijmegen, W. (2002). Hungarian names 101. http://www.geocities.com/Athens/1336/magyarnames101.html. [Online; accessed 22-July-2007].

von Flotow, L. (1997). *Translation and Gender. Translating in the 'Era of Feminism'*. Manchester: St. Jerome.

Zgusta, L. (2007). Names. *Encyclopædia Britannica* **24**. 728--733. `http://www.britannica.com/eb/article-9108749`. [Online; accessed 17-Aug-2007].

Zorn, I., Maas, S., Rommes, E., Schirmer, C. & Schelhowe, H. (eds.) (2007). *Gender Designs IT. Construction and Deconstruction of Information Society Technology*. Wiesbaden: VS Verlag für Sozialwissenschaften.