

Grammars of Coercion

Towards a cross-corpora annotation model

Working group 1 (“WORCK” COST-Action)

Working paper (Last revised 22-08-2020)

By Juliane Schiel (Vienna University), Johan Lund Heinsen (Aalborg University), and Claude Chevalyere (CNRS)

This document describes the first digital steps of Working Group 1 “Grammars of Coercion” (hereafter WG1) towards a common and multilingual cross-corpora annotation model for the semantic analysis of historical sources on coercion in labor. It is the result of four rounds of experiments with Catma ([a semantic annotation and analysis tool](#)) held online during the first semester of 2020. The purposes of this working paper are threefold: 1. to (help) keep track of the discussions and decisions that led to designing our initial annotation model; 2. to serve as guidelines for members of WG1 committed to contribute to the project by annotating texts; 3. to serve as a basis for a future methodological article.

WG1 has adopted a bottom-up methodology which consists in collecting structured information on coercion in labor from various historical contexts and with a primary focus on “texts” as a means to reconceptualize “coercion”. To avoid Eurocentric biases that would consist in transposing “universal” modern categories onto pre- and early-modern contexts, WG1 tries to avoid analysis based on labels and nouns and, instead, focuses on the actions describing coercion in labor activities and relations. Inspired by [Maria Ågren \(et alii\)](#)’s “verb-oriented” method, we propose an annotation model that takes “action-phrases” and their “actors” as the main focus of annotation, combined with the “entry-extraction-exit” phases of labor relations designed by [Marcel van der Linden](#). As a whole, the annotation model is designed to isolate “action-phrases” from texts without losing related information as to the actors involved in the actions. The texts chosen for this digital experiment are highly heterogenous. They are provided by active members of the working group, based on the sources they know best. Such a disparity is not without causing issues, in particular as to the comparability of the (con)texts, and as to variations in “genres” and modes

of speech. However, facing this crucial problem at an early stage of the project shall also provide an opportunity to address it along the road.

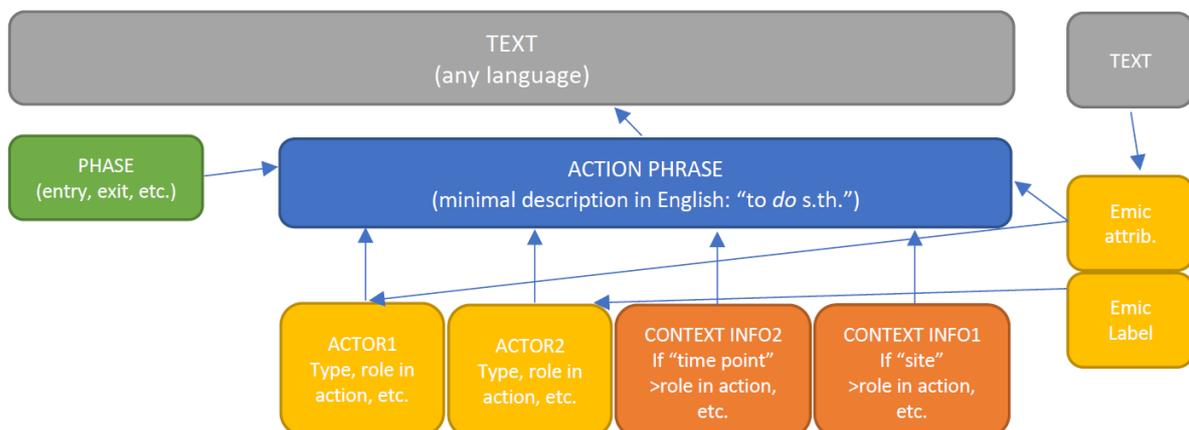
In our annotation model, an “action phrase” is conceived as a coherent continuous grammatical compound describing an “action”, whatever the grammatical structure of the language of the text. It can be a whole sentence, but also a segment of a sentence (when, for instance, long sentences describe several actions). In this model, “action phrases” are not dissociated from their “actors”. The term “actors” is understood in a broad sense. “Actors” can be individuals, groups, institutions, administrations, etc. who take part in or are concerned by the action, but also the related contextual information (such as time and space) that play a role in the action. The relationship between the “action-phrase” and its “actors” is maintained by assigning “roles” to the actors (e.g. a person who does an action is the “actor” of the action, a group who is done something through an action is the target of the action, a place has a role in an action, etc.).

The annotation model is (at this stage of our experiment) composed of 5 different tagsets, that might undergo changes as we refine the methodology and experiment more with the practice of annotation:

1. “action phrases”,
2. “actors”,
3. “contextual information”,
4. “phase”,
5. “semasiological naming”.

Besides “action phrases”, “actors and “contextual information”, the “phase” tagset helps situate an “action phrase” in a chain of actions and within the life cycle of a labor relation (e.g. the action corresponds to the “entry” into a labor relation, takes place after the “exit” phase, or between an “exit” phase and a new “entry” phase, etc.). The “semasiological naming” tagset is still at a very early stage of conceptualization and elaboration. Its *raison d’être* is that at early stages of our annotation attempts we felt the need to keep track of context-specific concepts, and of the “emic labels” and “emic attributes” separated from the “action phrase” itself, but implicitly related to it or to its actors. For example, an action can be judged “shameful” farther in the text; an actor can be assigned some qualification in another portion of the text where no action phrase related to labor and/coercion is to be found; or an actor can be related to an “emic label” in the action phrase, while being related to another label elsewhere in the text.

As for the organization of the tagsets, we decided to keep the five tagsets separated. Catma provides many ways to structure an annotation model, in particular thanks to the possibility to add



as many “subtag” levels as the user wishes to. As the annotation model is still under development and might undergo major changes, it seems relevant to compartmentalize the tagsets so as to permit focused changes within one tagset without altering the others.

Experimenting with Catma has been a decisive phase for the team to familiarize with semantic annotation methodology. It has been a crucial step toward building an annotation model and elaborating a methodological approach. Yet, the decision to carry on with the tool depends on a few important factors.

- First, having access to a more stable Catma instance would be an improvement. All of us have experienced bugs and other annoying small technical issues. On the “user-friendly” side, being able to create tags and subtags with pre-defined mandatory properties and values fields would be highly time-saving: We expect a very large collection of “action-phrases” subtags. Creating each of them will take time, all the more if, for each new subtag, we have to manually add the same properties and value-lists again (notwithstanding the risk of human error when doing so, which might undermine computing results in the end).
- Second, it requires to clearly assess whether we are expecting too much from the tool, and whether it can achieve what we expect it to do. Our annotation methodology is based on one important assumption, which requires further discussion: When comes the time for analysis, can we maintain the links between the tagged portions of text and the tags from different tagsets (which we implicitly build by structuring the tagsets the way we do)?
- Third, even though Catma is able to maintain links between the tags and the segments of texts they appear in, another problem will arise when an implicit link exists between an “action phrase” and information found outside of the action phrase itself. This can appear in the following situations:
 - In the case of “semasiological” tagging, mentioned above;
 - In the case where an actor is implicitly linked to an action phrase (e.g. when the identity of the actor of an action phrase is only implicit; or when the subject of a long sentence is only present once but assigned to two or more actions, like in the fictitious sentence “the slave ran away, disappeared, and hired himself out as a mariner”).
 - Our solution to this issue would be to automatically assign unique identifiers to actors and action phrases. This would make possible, for instance, to tag an actor as being assigned a role in an action described in another part of the text through its unique identifier, not by its position in the text; it would also allow to relate an actor to its emic attribute present in another part of the text. However, this not only seems a lot more work, but also not something Catma was designed for (having Catma assign unique identifiers automatically would, again, be time saving, on the condition that two tags can have the same identifier, like in the case of the exact same action being described more than once in a text). Unique identifiers employed to relate tags within a same text shall also be differentiated from identifiers we assign to external reference tables (such as one recording all the biographical information on the individual actors mentioned in the texts).
 - Finally, we would have to maintain a relation between the work we do with Catma and external tables or tools we use for different purposes. Developing a general “biographical information” table of all the actors mentioned in the texts might serve many purposes, one of which could be to assess the representativity of the “work” situations explored in the project. We also need a complete list of the “action phrase”

tags and all their declinations, properties and values for the sake of curating this central part of our project.

1. Action Phrases

An “action-phrase” is any sentence (or segment of a sentence) expressing views on, or describing actions related to labor, to labor relations, and to coercion. Each “action-phrase” tag shall be reduced to a minimal expression using the infinitive form of an English verb (e.g. “to do”) chosen as the closest “translation” of the action described and corresponding as closely as possible to what the text actually says (e.g., “to perform labor” shall be used when the text says an actor performs a kind of labor, not to interpret an action as a labor performance). The first-level subtag (the infinitive form of an English verb) is not assigned any object (like “s.th.”, or “labor” as in the above-mentioned example). Further characterizations of the action will be provided through additional levels of subtags.

In all cases, the “action phrases” tagset will require regular curating activities, either through small discussions or workshops, based on an updated list of the currently used verb-forms.

Structure of the “action phrases” Tagset

- **First level: “action phrase” (one TAG)**
 - The reason why we chose to use “action phrase” as a first level tag of the “action phrase” Tagset is that decisions made during the annotation activity require intimacy with the text and reflection. Deciding which English word corresponds best to the action, and setting properties and values might take time and multiple passages. As a first step, we still want to be able to simply identify a segment of text as an “action phrase”, and refine its properties and values later on.
 - Eventually, if participants in the project feel the need to annotate specific “grammatical words” (in particular verbs, adjectives or adverbs, depending on the language), a second tag can be used for this purpose.
- **Second level: English infinitive verb forms (many SUBTAGS)**
 - This level will result in an extended collection of minimal action phrases. As underlined by Johan Heinsen, the principal challenge of the project will be to come up with a curated list of potential action phrases that would allow for some uniformity in the way of tagging the infinity of possible “actions”. Normalizing this level of subtags will be a major and potentially daunting task. To some respects, this step amounts to parsing the texts and to creating normalized tokens of “action phrases”.
- **Third level: declinations of second-level verb forms (many SUBTAGS)**
 - This level is used to restrict/refine the meaning of the action conveyed by the verb form used in the second-level tag. E.g., a third level tag for action phrase “to ask” might be “to interrogate”.
- **Fourth and Fifth levels: PROPERTIES of the actions and corresponding VALUES**
 - The properties of the action phrases are of two kinds: 1. Common to all action phrases; 2. Contextual properties, depending on the action-phrase itself (and on which values can be appended).
 - If the action cannot be assigned a property, N/A (not available/applicable) shall always be recorded.
 - Mandatory properties with their corresponding values (to be improved and refined):

- “**id**”
 - *Unique identifier of the action*
- “**lawfulness**” of the action (is the action lawful?)
 - *Lawful*
 - *Unlawful*
 - *Uncertain* (when one cannot ascertain whether the action is lawful or not in the context)
- “**outcome**” (did the action succeed?)
 - *Failure*
 - *Success*
- “**type**” (what are the attributes assigned to the action?)
 - *Values are contextual and defined according to the type of action.* As examples:
 - “To answer” may have “under the threat of detrimental consequences” as a “type” value
 - “To ask” may have “s.o. to do s.th.” as a “type” value
- “**actor intention**” (did the actor engage in the action willingly?)
 - *Willingly*
 - *Unwillingly*
- “**incentive**”
 - *Any short information regarding the incentive of the actor to engage in the action*
- “**occurrence**” (did the action already occur? one or many times?)
 - *Unique*
 - *Repeated*
 - *Prospective*
- “**justification**”
 - *Any short information regarding the justifications for the actor to engage in the action*
- “**target compliance**” (was the compliance of the action target required?)
 - *Required (willingly)*
 - *Required (unwillingly)*
 - *Unrequired (willingly)*
 - *Unrequired (unwillingly)*
- Contextual properties and value: When the meaning of a verb form can be refined by the use of a function word + a complement, the function word is used as a property, which can in turn be assigned the specific value of the complement:
 - As an example, the meaning of the verb “to prepare” can be refined by adding “for” as a property and “entry”/“exit”/“extraction” as values.

2. Actors

The actors are all the persons, groups of persons and entities who participate, are involved or concerned with an action, whatever their role, and who are not “contextual actors” (such as time and space, for which there is a separate tagset). So far, we have not much elaborated other tags than the “persons” tag.

Structure of the “actors” Tagset

- **First level: “actors” (several TAGS)**
 - At this level, the user simply assigns a type to the segment of text tagged as an “actor”:
 - Person (the actor is an individual, named or not);
 - Group (the actor is a group of individuals, named or not);
 - Institution (the actor is an institution)
 - Administration (the actor is an administration)
 - Hypothetical actor (the actor is a hypothetical actor, for instance the hypothetical “slave” for whom a normative text such as a regulation establishes penalties, duties, etc.)
- **Second and third level: PROPERTIES and VALUES for the “persons” (many SUBTAGS)**
 - **Id**
 - **Gender**
 - *Female*
 - *Male*
 - *N/A*
 - *Non-binary*
 - **Role in action**
 - *Actor (the person who actually does the action)*
 - *Target (the actual object of the action, on whom the action is done)*
 - *Object (the person about whom the action is done, but who is not the target/direct object of the action)*
 - *Non-partisan (the person is mentioned, but not concerned by the action)*
 - *Beneficiary (the person who benefits from the action, but is not its direct target)*
 - *Mediator (the person who acts as a mediator but is neither the main actor nor the direct object of the action)*
 - *Cause (the person who is the cause of the action)*
 - *Associate (the person who is an associate of the actor)*
 - **Employer/Mediator/Worker**
 - *N/A*
 - *Worker*
 - *Mediator*
 - *Employer*

With regards to the “persons” properties and values, we still wonder whether it is necessary to have a standardized (or canonical) “name” property, and if we can do more to assess intersectionality through specific properties and values. Another issue still under discussion is whether we want to pre-impose such categories as “worker” and “employer” to the actors. Coercion related to work might not be so clear in many contexts. For instance, being assigned a position/status in society would imply that one is expected, among other things, to perform labor, without being primarily (or at all) identified in the said context as a “worker”. However, in cases where we have a “traditional” labor relation, such roles are useful, because it creates a heuristic in which we can ask, what do “employers” at entry, or what are shared emic labels of workers.

3. Phase

The “Entry-Extraction-Exit” model designed by Marcel van der Linden requires additional liminal and post/pre phases that might be mentioned in an action. As a first possibility, some of us considered that the “phase” annotation could have been a property of the action itself. However, we came to consider that an “action-phrase” can involve more than one “phase”. For the moment, we only annotate “action-phrases” or segments thereof as “entry”, “exit”, “extraction”, etc. Properties and values might offer possibilities of refining the “phase” annotation, with regards either of the conditions of “entry”, “exit”, etc., or of sequences.

Structure of the “phases” Tagset

- **First level: several TAGS corresponding to the identified phase**
 - N/A
 - Entry
 - Extraction
 - Exit
 - Liminal phase
 - Pre-entry
 - Post-exit

4. Contextual Information

All information on the time, space, and other context of the actions, including objects and prices. We see Space and Time as “actors” of an “action-phrase” (although in a different way than “persons” and “groups”) as they also play a “role” in an action.

Structure of the “contextual information” Tagset

- **First level: several TAGS corresponding to several types of contextual information**
 - Date
 - Duration
 - Space
 - Object
- **Second level: properties and values of the above-mentioned tags**
 - Date
 - **Date**
 - *N/A*
 - *Date in ISO format (ex. 2020-05-23, 2020-05, or 2020)*
 - **Role in action**
 - *N/A*
 - *Date of action*
 - *Date mentioned in action*
 - *Deadline of action*
 - *Reference to past action/event*
 - *Reference to future action/event*
 - **Date relation to text**
 - *N/A*

- *Current*
- *Past*
- Duration
 - To be developed
- Object
 - **Type** (types can be many and will be contextual, like clothes, papers, food, etc.)
 - **Role in action**
 - *Object in action*
 - *Outcome of action*
 - *Object of action*
- Space
 - **Type**
 - *N/A*
 - *Site of an action*
 - *Site mentioned in action*
 - *Place of action*
 - *Place mentioned in action*
 - **Location ID**
 - *id*

5. Semasiological naming

To be developed...

6. Examples

Example 1

A sentence from a *Xing'an huilan* case stating (in a raw English translation followed by the original text):

Because her female servant Wu Yunzhu had an illicit sexual intercourse with Liu Chun, who cultivated her land, Mrs. Niu Hulu expelled Liu Chun, and exchanged Wu Yunzhu against a female servant of Tuo Xinbao's.

[鈕祜祿氏][因(使女)烏雲珠與租種伊地之劉春通姦]。[將劉春攆逐]。[并將烏雲珠與(托新保)(使女)互換]。

This sentence:

- describes **two actions** (in blue, between brackets and underlined):
 - to expel the male farmer
 - to exchange the female servant
- both actions have the **same main named actor** (in green between brackets):
 - Mrs. Niu “does” the two above-mentioned “actions (ID 0001 in the table below)

- both actions have the **same justification** (in orange between brackets):
 - Because of an illicit sexual intercourse
- each action has **a different named target** (in red):
 - the named male farmer (who is expelled, **ID 0002** in the table below)
 - the named female servant (who is exchanged **ID 0003** in the table below)
- the second action has **two “secondary actors”** (in yellow between parenthesis):
 - The named person with who the exchange is done (**ID 0004**)
 - The unnamed female servant against whom the first female servant is exchanged (**ID 0005**)
- There is one “emic label” (the term **使女**, in maroon between parenthesis), used both to qualify the named female servant who is exchanged because she had a relation with the farmer, and who is used to name the unnamed female servant against the other is exchanged.
 - *One issue here, is that the “emic label” present in the action, which is not a legal term but a household designation, is also implicitly related to a legal category mentioned earlier in the text (how to reconnect establish this link?)*

Annotations:

ACTION 1					
Tagset	Tags	Subtags 1	Subtags 2	Properties	Values
Action phrase	“action phrase”	“to exchange”		<i>type</i>	s.o. against s.o.
				<i>lawfulness</i>	lawful
				<i>outcome</i>	success
				<i>intention</i>	willingly
				<i>justification</i>	offence committed by target
				<i>occurrence</i>	unique
				<i>target compliance</i>	unrequired (n/a)
				<i>incentive</i>	n/a
				<i>ID</i>	...
Actors	“person”			<i>ID</i>	0001
				<i>role</i>	actor
				<i>Employer/ med...</i>	n/a
				<i>gender</i>	female
	“person”			<i>ID</i>	0002
				<i>role</i>	cause of action
				<i>Employer/ med...</i>	n/a
				<i>gender</i>	male
	“person”			<i>ID</i>	0003
				<i>role</i>	target
				<i>Employer/ med...</i>	n/a
				<i>gender</i>	female
	“person”			<i>ID</i>	0004
				<i>role</i>	beneficiary
				<i>Employer/ med...</i>	n/a
				<i>gender</i>	male
	“person”			<i>ID</i>	0005
				<i>role</i>	target

				<i>Employer/ med...</i>	n/a
				<i>gender</i>	female
Contextual info	<i>none directly mentioned in the action, but we know the region where the case took place from elsewhere in the text</i>				
Phase	<i>“exit” although the exchange of servants is, for each of them an exit and a new entry at the same time</i>				
Semasiological	<i>Emic label</i>			type	household role
				label	<i>使女</i>
ACTION 2					
Action phrase	<i>“action phrase”</i>	<i>“to discharge”</i>		<i>type</i>	<i>s.o. from work</i>
				<i>lawfulness</i>	lawful
				<i>outcome</i>	success
				<i>intention</i>	willingly
				<i>justification</i>	offence committed by target
				<i>occurrence</i>	unique
				<i>target compliance</i>	unrequired (n/a)
				<i>incentive</i>	n/a
				<i>ID</i>	...
Actors	<i>“person”</i>			<i>ID</i>	0001
				<i>role</i>	actor
				<i>Employer/ med...</i>	n/a
				<i>gender</i>	female
	<i>“person”</i>			<i>ID</i>	0002
				<i>role</i>	target
				<i>Employer/ med...</i>	n/a
				<i>gender</i>	male
	<i>“person”</i>			<i>ID</i>	0003
				<i>role</i>	Cause of action
				<i>Employer/ med...</i>	n/a
				<i>gender</i>	female
Contextual info	<i>none directly mentioned in the action, but we know the region where the case took place from elsewhere in the text</i>				
Phase	<i>“exit”</i>				
Semasiological naming	<i>None</i>				

Example 2

A short letter written from Margherita Datini to her husband Francesco Datini, Prato to Florence, on 22 March 1395 (English translation by Carolyn James/Antonio Pagliaro).

Schiatta and his wife were here on Sunday. He came to get a baby [for his wife] to wet nurse, but he didn't come to an agreement with anyone. He left word that they should come to speak to me for further information because he is one of our workers, but no one came to give me a reply. He has left it to me to search for a baby but I haven't done so; and I won't, because it would be wrong and shameful. The woman is old and the milk too abundant, despite the fact that she says she has very little. [...]

[iSchiatta] ci fu qui domenica, egli e lla donna sua: [veniva per tôte uno fancullo a balia], non si achordò cho' (niuno); [lasc(i)ò che venisono a favelare a mme. a 'formarsi mecho], [perch'è nostro lavoratore]; (no') ci sono mai venuti a me per darmi la risposta, lascomi ch'io ne gli facessi cerchare d'uno; [io non n'ò fatto cerchare e non ne farò cerchare], [perch'è cosa d'averne pechato e vergogna], perché la donna (è vecchia e 'l late abundato, bene ch'ella dica che gl'abia pocho); [...]

This letter:

- describes **three actions** (in blue, between brackets and underlined):
 - to search employment for so.
 - to ask so for help to find an employment for so.
 - to refuse to help
- the actions have **two main named actors** (in green between brackets):
 - Schiatta, Francesco's worker, "does" the first two "actions" (ID 0001 in the table below)
 - Margherita, Francesco's wife, "does" the third "action" (ID 0002 in the table below)
- the second and third action are followed by a **justification** (in orange between brackets):
 - Schiatta asks for help "because he was Francesco's worker"
 - Margherita refuses to help "because it would be wrong and shameful"
- all actions have **the same named target** (in red):
 - Schiatta's wife [nameless] (whose services as wet nurse are offered and refused, ID 0003 in the table below)
- all actions have **"secondary actors"** (in yellow between parenthesis):
 - no one wants to employ Schiatta's wife (unspecified, no ID)
 - no one contacts Margherita to employ Schiatta's wife (unspecified, no ID)
 - Margherita reports all three actions to her husband, Francesco Datini (ID 0004)
- the third action is followed by **emic labels and attributes** (in maroon between brackets) used to qualify Schiatta's wife (ID 0003)

Annotations:

ACTION 1+2					
Tagset	Tags	Subtags 1	Subtags 2	Properties	Values
Action phrase	"action phrase"	"to search employment"	"to ask for help to find employment"	<i>type</i>	for s.o. else
				<i>lawfulness</i>	lawful
				<i>outcome</i>	unsuccessful
				<i>intention</i>	willingly
				<i>justification</i>	because he was our worker

				<i>occurrence</i>	repeated, in variations
				<i>target compliance</i>	unrequired
				<i>incentive</i>	unknown
				ID	0001
Actors	“person”			ID	0001
				Role	actor
				<i>Employer/ med...</i>	mediator
				<i>gender</i>	male
	“person”			ID	0002
				<i>role</i>	addressee
				<i>Employer/ med...</i>	mediator
				<i>gender</i>	female
	“person”			ID	0003
				<i>role</i>	target
				<i>Employer/ med...</i>	worker
				<i>gender</i>	female
	“person”			ID	0004
				<i>role</i>	receiver
				<i>Employer/ med...</i>	employer
				<i>gender</i>	male
Contextual info	<i>Schiatta di Niccolò, or Tantera, was an agricultural labourer and served the Datini household for many years.</i>				
Phase	“pre-entry”				
Semasiological naming	“emic label”	“balia”	“la sua donna”	Type	Wet nurse
ACTION 2					
Action phrase	“action phrase”	“to refuse”		Type	to help
				<i>lawfulness</i>	lawful
				<i>outcome</i>	success
				<i>intention</i>	willingly
				<i>justification</i>	because it would be wrong and shameful
				<i>occurrence</i>	Potentially repeated, in variations
				<i>target compliance</i>	unrequired (n/a)
				<i>incentive</i>	Bad reputation
				ID	0002
Actors	“person”			ID	0002
				<i>role</i>	actor
				<i>Employer/ med...</i>	mediator
				<i>gender</i>	female

	“person”			ID	0001
				role	Target
				Employer/ med...	mediator
				gender	male
	“person”			ID	0003
				role	Cause of action
				Employer/ med...	worker
				gender	female
Contextual info	<i>We know from other letters that (Margherita fears that) mediating a bad wet nurse brings bad reputation for the Datini couple.</i>				
Phase	“pre-entry”				
Semasiological naming	“emic attributes”	“è vecchia”	“late abundato, bene ch'ella dica che gl'abia pocho”	Type	Wet nurse