DEC–B06 Final Project:
A probabilistic approach to French double negation

Jeremy Yeaton

22 June 2018

1 Introduction

In French, sequences of potentially negative expressions like *personne* and *rien*, referred to here as Negative Concord Items (NCI) (Watanabe 2004), can have two possible interpretations; a sentence like (1a) can have a Negative Concord (NC) reading as in (1b) or a Double Negation (DN) or negative discord reading as in (1c), in which negations cancel each other out to produce a positive reading.

(1) a. *Personne ne dit rien*

b. Nobody says anything = Everyone is silent

c. Nobody says nothing = Everybody talks

On a par with other Romance languages, French is generally regarded as a concord language (Zeijlstra 2004, De Swart 2002). Double negation readings of NCI sequences are possible and not uncommon, however (Déprez 2000, Corblin & Tovena 2003). Still quite unexplored, however, are the different factors that govern the choice between NC and DN interpretations for French speakers. We also know very little about how this phenomenon distributes in production.

2 Negative Concord

In languages like French where an ambiguity exists between DN and NC, there are several potential explanations for how listeners arrive at an interpretation. The first is that the morphosyntactic structure of the NCIs guides the interpretation. In a study investigating the role of syntactic structure on interpretations, Déprez et al (2015) found that the syntactic structure of the subject, but not the object, influenced the interpretation. In the case of parallel pronominal subjects and objects, participants responded at chance in a picture choice task. If instead of a pronoun, the sentence contained a DP as a subject, e.g.: *aucun enfant*, participants were more likely to interpret it as DN than NC. The results of this study are shown in figure 1.

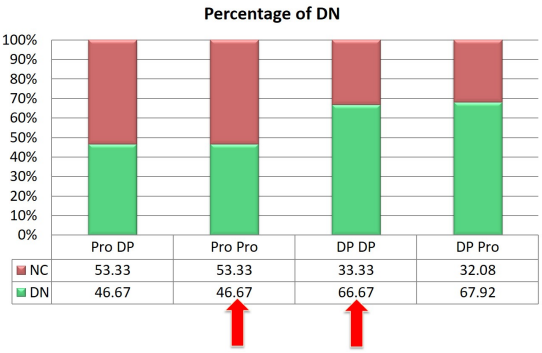

Figure 1: Distribution of NC and DN interpretations for different syntactic structures.

A follow-on study used a production task to investigate the role of context and prosody in determining the accessed interpretation. In the study Déprez & Yeaton (2018) found that context is a reliable predictor of interpretation. As shown in figure 2, participants respond according to context significantly above chance. Participants also used different prosodic contours for NC and DN productions, but as the perception study is not yet complete, we cannot report on whether or not that is a reliable marker for listeners.

As an example of NC or DN intended context, take the ambiguous sentence 2:

(2) *personne ne boit rien dans les soirées*

The experimental contexts are given in 3 (NC) and 4 (DN):

(3) *Dans notre famille, on est tous allergique à l'alcool:*

personne ne boit rien dans les soirées.

(4) *Chez les jeunes, la consommation d'alcool est effrayante:*

personne ne boit rien dans les soirées.

In 3, one understands that because all of the people in the family are allergic to alcohol, that nobody drinks, and in 4, one understands that everyone drinks something at parties because the youth are out of control.

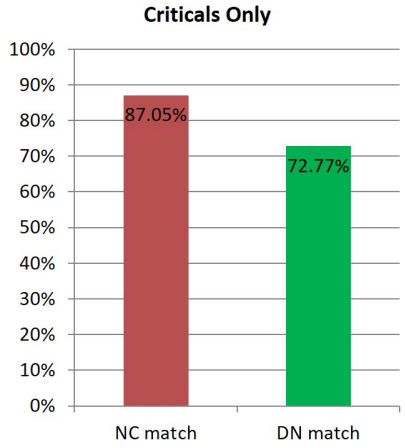

Figure 2: Percent context-matching responses for NC- and DN-intended contexts.

It was also found that speakers do not uniformly interpret ambiguous sentences in context the same way. The results of the context responses are broken down by subject in figure 3. The results are organized by their overall proportion of NC and DN responses, and there is a clear gradient from subjects who respond almost exclusively with NC, to one who responds almost exclusively with DN. It is to be noted, however, that there are more subjects with an NC preference than a DN one.

Processing cost has also been proposed as factor determining interpretations. As of yet, however, there is no unified theory that successfully accounts for all of the variability. In this paper, I would like to examine how listeners might use probability to determine the intended interpretation.

3 Probabilistic Approaches

In their textbook on Probabilistic language understanding, Scontras et al put forth a system of

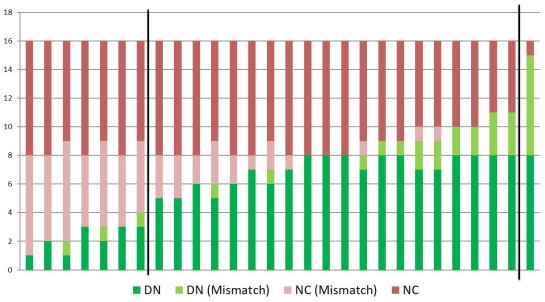

Figure 3: Percent context-matching responses for NC- and DN-intended contexts by subject.

tools to model the Rational Speech Act (RSA) framework. This system views communication as a continuous shared reasoning task between a speaker and a listener. This means that each is reasoning about the other's communicative strategy. The speaker's goal is to provide maximally clear information with the least effort, while the listener's goal is understand the meaning intended by the speaker. More formally, the listener L reasons about the speaker S, and infers the state of the world s given that the speaker chose to produce the utterance u. The speaker chooses u by maximizing the probability that a listener would correctly infer the state of the world s given the meaning of u:

$$P_L(s|u) \propto P_S(u|s) \cdot P(s) \tag{1}$$

Or, the listener L computes the probability of a state s given some utterance u by reasoning about the speaker S. L reasons that the probability of s given u is proportional to the probability that S would use u regarding the state s, times the prior probability of s itself.

As we look at a situation like the DN/NC ambiguity, we must decide which variables are in play and how to reflect those in the formalism.

4 Probabilistic Approaches applied to NC

To create a probabilistic model of the ambiguity of DN and NC, we need to model a listener who can determine if the state of the world s. For now let us continue to use 2 as our example. The listener L must determine if s is a) a *drinking* world, and in so doing, determine if s is an NC world or a DN world. As such, the following parameters would be necessary:

1. Prior probability of the world being a $s =$ *drinking* world or not: $p(s_{drinking})$
2. Prior probability of NC (out of context), given DP and pronominal forms in each position: $p(s_{NC}|u)$
3. Probability for any sequence of multiple NCI to have a NC intended meaning: $p(s_{NC})$
4. Some metric for the negative or positive guiding power of the context.

The set of possible states s of the world is:

$$s = \left\{ \begin{matrix} drinking \\ \neg drinking \end{matrix} \right\} \wedge \left\{ \begin{matrix} DN \\ NC \end{matrix} \right\}$$

And the set of possible utterances from which u can be drawn is

$$u = \left\{ \begin{matrix} \langle DP,DP \rangle \\ \langle DP,Pro \rangle \\ \langle Pro,DP \rangle \\ \langle Pro,Pro \rangle \end{matrix} \right\}$$

Based on the picture choice task results, we can assume that the first two items of u which have DPs in the subject position increase the prior probability of DN, but for the sake of simplicity we will assume that all utterances $u = \langle Pro,Pro \rangle$.

In order to see how this phenomenon distributes in the real world, I queried the ESLO corpus for each of the lexical items *personne, rien, aucun, aucune*. I then crossed these to see where more than one appeared in the same utterance. Of the 5,700 total utterances including at least one of these, only a very small fraction contained more than one within the same clause ($n = 25,\ 0.4\%$). Among the 25 exemplars found, 19 were determined to be NC by a small committee of my francophone classmates, and the remaining 6 lacked sufficient context within the given utterance to reach a conclusion. None of the utterances was perceived to be DN only from the utterance given.

This is a tiny sample, but gives leads us toward the assumption that the base rate for an NC interpretation is much higher than that of a DN interpretation. Given the results from the picture choice task, however, it would seem that listeners neglect this base rate in the absence of context. Even though NC productions are more common, they return to 50/50 when they have no additional information:

$$p(s_{NC}) = p(s_{drinking}) = \frac{1}{N}\epsilon \tag{2}$$

where N is the cardinality of the set of possible states of the world, in this case $|\{\text{NC, DN}\}| = |\{\neg drinking,\ drinking\ \}| = 2$, and ϵ is the estimation error. We assume that any given world drawn at random has an equal probability of being a *drinking* world.

When context is provided, however, it seems that more subjects have a preference toward NC than DN. It is possible that this is a result of statistical inference, and the variability is accounted for by errors in estimation. This would imply, however, that listeners were ignoring the context if they only took the base rate into account:

$$p(s_{NC}) = \frac{o_{NC}}{a}\epsilon \tag{3}$$

where a is the approximate total number of statements accessed containing multiple NCI, and o is the number of observations of NC in a. This would simply be a Bayesian accounting for the distribution of interpretations based solely on prior experience. This still does not account for the effect of context, however.

Context is a bit more complex. Psycholinguistics has shown that when presented with an uncommon word, structure, or sound following some well-formed string, that we register surprisal at the unusual or incorrect item. This means that as we receive the context for our eventual ambiguous sentence, we form predictions about the nature of the sentence that follows.

In order to account for the role of context by itself, it would be useful to conduct a picture-choice task using the contexts only in order to have experimental evidence to compare to the model. In theory, though, we can estimate the likelihood of a world being a *drinking* world or not given the information the speaker has shared. Even without the utterance *personne ne boit rien dans les soirées*, it is more likely that the world in which the context C is *on est tous allergique à l'alcool* is not a *drinking* world.

We can further formalize this as in 4, where we estimate the probability that the state of the world is $s = \{\neg drinking \wedge NC\}$, given the context C and utterance u:

$$p_L(s|u,C) \propto p(s_d) \cdot p(s_d|C) \cdot p_S(u|s_{NC}) \tag{4}$$

Such that $p(s|u,C)$ is proportional to the prior probability of the world being a *drinking*

world ($p(s_d)$) and probability that it is a drinking world given the context ($p(s_d|C)$), combined with the probability that the speaker S would use the utterance u to describe the state as NC ($p_S(u|s_{NC})$).

5 Discussion

Further experimental and corpus work is needed to flesh out a number of elements of this model:

1. Picture choice experiments to test the interpretation of the world, given only the context. This will help shed light on the size of the role that context plays in this, or the probability of being in a drinking world or non-drinking world, given the context ($p(s_d|C)$).

2. Corpus study on these constructions. In the world of the French language, how often does one encounter an ambiguous expression with multiple NCIs? Among these, how often are they NC and how often are they DN? This will shed light on the probability of a given ambiguous sentence being NC or DN ($p(s_{NC})$).

3. Follow-up corpus study, examining how often the ambiguous expressions with multiple NCIs actually are intended as DN or NC, given the various structures. This would correspond to the probability of an NC world, given the utterance ($p(s_{NC}|u)$).

Following these studies, it will theoretically be possible to test the feasibility of this model. I suspect that another parameter may need to be introduced to imitate the listener's confidence in the speaker's knowledge, or how much the listener trusts the context.

My weak WebPPL skills, in combination with the information gaps to be filled by the additional studies meant that I was unable to form a working model using WebPPL, but this has definitely been a challenging and interesting project. I hope to be able to work on this model further in the future.