A French Version of the FraCaS Test Suite

Maxime Amblard, Clement Beysson, Philippe de Groote, Bruno Guillaume, Sylvain Pogodalla

A French Version of the FraCaS Test Suite

Maxime Amblard, Clément Beysson, Philippe de Groote, Bruno Guillaume, Sylvain Pogodalla
Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
{maxime.amblard, clement.beysson}@univ-lorraine.fr
{philippe.degroote, bruno.guillaume, sylvain.pogodalla}@inria.fr

Abstract

This paper presents a French version of the FraCaS test suite. This test suite, originally written in English, contains problems illustrating semantic inference in natural language. We describe linguistic choices we had to make when translating the FraCaS test suite in French, and discuss some of the issues that were raised by the translation. We also report an experiment we ran in order to test both the translation and the logical semantics underlying the problems of the test suite. This provides a way of checking formal semanticists' hypotheses against actual semantic capacity of speakers (in the present case, French speakers), and allow us to compare the results we obtained with the ones of similar experiments that have been conducted for other languages.

Keywords: semantics, inference, French

1. Introduction

FraCaS (A Framework for Computational Semantics) was a European project on language research and engineering that ran from 1994 to 1996. It identified semantic inference as a significant part of computational semantic processing and developed an inference test suite to provide a way to evaluate and compare inferential competence of natural language processing systems and semantic theories, and proposed to test NLP system's semantic capacity against inferencing tasks. The proposed tests take the form that was popularized by textual entailment recognition (TER) tasks (Dagan et al., 2006; Korman et al., 2018): the system is given an input text T together with a claim (a.k.a., hypothesis) H, and is "asked to determine whether [H] follows from T" (Cooper et al., 1996) (or, in Dagan et al. (2006)'s words, "if, typically, a human reading T would infer that H is most likely true").

Cooper et al. (1996, chap. 3) provided 346 problems expressed in English and that are grouped by linguistic and semantic phenomena, e.g., monotonicity of quantifiers, collective and distributives plurals, ellipsis and anphora, etc. Figure 1. shows an example of the provided inferences: sentences above the line are the text T, the hypothesis H to be tested is worded as the question under the line, and the expected answer to the question is stated at the very bottom. Section 2. more precisely describes the range of problems and, in particular, the range of the possible expected answer. Contrary to more recent TER tasks, language data were artificially build, in order to precisely illustrate semantic phenomena and to cover a wide range of natural language inference occurrences.

(T)	Every Italian man wants to be a great tenor.
(T)	Some Italian men are great tenors.
(H)	Are there Italian men who want to be a great tenor? [Yes]

Figure 1: Problem 2 as presented in (Cooper et al., 1996)

The FraCaS test suite has indeed been used by several understanding systems with different underlying logics, for instance systems implementing natural logic (MacCartney and Manning, 2007; MacCartney and Manning, 2008), or using the Coq proof assistant (Bernardy and Chatzikyriakidis, 2017; Chatzikyriakidis and Bernardy, 2019), exemplifying that "[t]he major added value of logic as a representational framework in computational linguistics is its suitability for the development of provably correct inference procedures." (Pinkal and Koller, 2012).

For these experiments, MacCartney provided an XML version of the FraCaS test suite[1]. This version has been used in different projects, in particular to create a bilingual (English and Swedish) treebank of parsing trees (Ljunglöf and Siverbo, 2011; Ljunglöf and Siverbo, 2012) using Grammatical Framework (GF, Ranta (2011)). It also has been used in the MultiFraCaS project[2] to provide Farsi, German, Greek, and Mandarin translations. The test suite also influenced building a problem set for Japanese (Kawazoe et al., 2017). Interestingly, this latter project did not stick to a one-to-one correspondence with the FraCaS test suite and are not literal translations because of language specificities. This highlights several important points about cross-lingual comparisons of natural language inference:

- some phenomena may be triggered by quite different linguistic constructions;
- linguistic constructions expressing some specific semantic phenomenon may be lacking in some languages
- natural language polysemy has to be taken into account, and the concepts related to a word in a language may not completely fit the concepts of its translation in another language. For instance, an adjective might be more easily considered as non-intersective in a language than in another one.

Studies have been performed in order to assess the extent to which intuitions of inference of ordinary speakers corre-

[1] https://nlp.stanford.edu/~wcmac/downloads/

[2] https://gu-clasp.github.io/multifracas/

spond to the theoretical notions of inference that the authors of the FraCaS test suite intended to illustrate. They demonstrate the degree to which the intuitions of the speakers may vary for some of the problems of the test suite (Cooper et al., 2016; Chatzikyriakidis et al., 2017).
In this article, we introduce a French version of the FraCaS test suite.
The resource is freely available[3] under Creative Commons BY-NC-SA license. To the best of our knowledge, beside the Cross-lingual Natural Language Inference corpus (XNLI[4], Conneau et al. (2018)), it is the only available corpus for testing natural language inference for French. We also ran an experiment to assess both the idiosyncrasy of the translation and the extent to which it fits the expectation in terms of the resulting inference.
Section 2. introduces more precisely the FraCaS test suite and the data it contains. Section 3. describes our French version of the the FraCaS test suite and the design choices we made. Section 4. describes an experiment that we ran, and compare its results with the results that were obtained for other similar experiments.

2. Description of the Original Resources

The English original resources was presented in (Cooper et al., 1996) as chapter 3, entitled "A Semantic Test Suite". Examples in this chapter are numbered from (3.1) to (3.346), and these numbers (1 to 346) were kept in the further works on the test suite and are now considered as identifiers for these problems, even if 4 of the 346 examples in the original document are not stated as problem but only as linguistic examples (examples 276, 305, 309 and 310).
As a starting point, in particular with respect to the XML formatting, we used the XML conversion provided by MacCartney. In this XML version, all examples are encoded (even the 4 non problems). We stuck to this structure, the XML file we provide contains a French version of the 346 original statements of the initial document, even if the experiment presented in Sect. 4. was only based on the 342 examples that are actually presented as inference problems.
As shown in Fig. 1., each problem contains a few premises, from 1 to 5 premises, and one question. Among the problems, 55.5% have only one premise, 35.3% have two premises, 8.4% have three premises, and only three problems (0.9%) have more than four premises. In the XML version by MacCartney, as is now standard in TER, each question is reformulated as an hypothesis which is the declarative counterpart of the question. Finally, the expected answer is also given. An answer can be *Yes* or *No*, stating whether the entailment holds or not, but it can also be *Don't know*, *Unknown*, or also contain some explanation (e.g., *No, if both commissioners are female; otherwise there are more than two commissioners* in problem 62, or *Yes, on one reading* in problem 87). Table 1 shows an example of how the problems are rendered in the XML resource. In the following, we only give the hypothesis H and not the question Q, as the latter can easily be reconstructed.

P_1	Every Italian man wants to be a great tenor.
P_2	Some Italian men are great tenors.
Q	Are there Italian men who want to be a great tenor?
H	There are Italian men who want to be a great tenor.
A	Yes

Table 1: Problem 2 in the English resource

Expected answers are distributed among classes as follows: 52% are *Yes* answers, 27% are *Don't know* answers, 9% are *No* answers, and the remaining 12% are more detailed answers.
The *Don't know* answer may be quite confusing and is usually to be understood as expressing that premises do not give enough information to decide whether the hypothesis follows or not from the premises. Table 2 illustrates such a problem: H could be true together with P_1 although it does not follow from the latter.

P_1	Neither commissioner spends a lot of time at home.
H	Neither commissioner spends time at home.
A	Don't know

Table 2: Example of a *Don't know* answer (problem 30)

In few cases, more detailed information is given in answers. For instance, problems 256 and 257 share the same premises and hypotheses, and differ only by the answer (Table 3). In these examples, the lost may have occurred from 1982, in which case the answer is we do not know. But the hypothesis can also be understood as a loss for ITEL every year since 1982, i.e. in 1983.

P_1	ITEL has made a loss since 1992.
P_2	It is now 1996.
H	ITEL made a loss in 1993.
A (pb 256)	Don't know, on one reading of the premise
A (pb 257)	Yes, on one reading of the premise

Table 3: Identical problems with different answers (problems 256 & 257)

The problem set is divided into nine sections, corresponding to different semantic phenomena. They are summed up in Table 4. Two phenomena are tested in details (Generalized quantifiers and Temporal reference) and two sections are really small (Verbs and Attitudes). The other sections are balanced.

3. A French Version of the FraCaS Test Suite

Semantic resources for French are scarce, especially in the case of text entailment. Accordingly, we decided to build a French version of the FraCaS test suite. The goal is threefold:

[3] https://gitlab.inria.fr/semagramme-public-projects/resources/french-fracas
[4] http://www.nyu.edu/projects/bowman/xnli/

Section name	Subsection number	Number of problems		Number of LQ/HC problems	
		Absolute	% of the whole	Absolute	% of the section
Generalized quantifiers	5	80	23%	5	6%
Plurals	6	33	10%	2	6%
(Nominal) anaphora	6	28	8%	3	10.7%
Ellipsis	9	55	16%	16	29.1%
Adjectives	6	23	7%	2	8.7%
Comparatives	6	31	9%	2	6.4%
Temporal reference	5	75	22%	10	13.33%
Verbs	2	8	2%	0	0%
Attitudes	6	13	4%	0	0%

Table 4: Problem and LQ/HC problems breakdown by topic (LQ/HC problems are low quality or high complexity problems. See Sect. 4.4.)

- to have a test bed for precisely delimited logical semantics phenomena;
- to take advantage from the different versions that already exist to make cross-lingual comparisons;
- to provide a starting point for further extension with corpus data.

The current version is 1.1 and is the first one that is made publicly available. It includes only minor changes to version 1.0 on which the experiment described in Sect. 4. was run (typo corrections).

3.1. Methodology

Previous translation experiments showed that some phenomena that are easily rendered in the English test suite may not be as apparent in literal translations (and *vice versa*). This can be measured, as done by Cooper et al. (2016) and Chatzikyriakidis et al. (2017), and as shown in Sect. 4.. Moreover, because we also aim at providing an automatic form-to-meaning translation through syntactic analyses, and have ways to compare such a process for different languages, we tried to stick as much as possible to the original English syntax. As a result, the main constraint to build the v1.0 version of the French FraCaS test suite was to be as much as possible lexically and syntactically faithful, as in the MultiFraCaS resource, even though it sometimes led to awkward French wordings.

There were 5 translators, with formal semantics skills, and a three-stage procedure was set up. During the first stage, 10% of the whole set of problems from all the sections, i.e., 34 problems, have been translated by each of the translator (for every $1 \leq n \leq 34$, problem number $10n$ was translated). Then all translations have been discussed at the same time by all the translators to decide a gold translation and to set guidelines for translating the remaining problems (90% of the whole set).

During the second stage, the remaining 312 problems were split into 10 subsets of comparable size so that each of these subsets was translated by a different pair of translators. The work is done by 5 translators, thus, we have $C_5^2 = 10$ pairs of translators. For each subset, the two participants first translated all the problems on their own and then decide together of the final translation. They also spotted the difficult problems. The later were eventually discussed and provided with a translation by all the translators during the last stage.

3.2. Translation Principles

In order to enforce coherent translations through the problem set, some translation rules have been set and constantly referred to during the process. Some of them have or may have a strong impact on the underlying semantic phenomena. For instance:

- Tense and mode: there are differences between French and English in that respect. We chose to keep the same tense when available. Preterit was translated using French *passé composé*, as the French *passé simple* is nowadays usually literary style.
- Number distinction between *there is* and *there are*: constantly translated by *il y a* with a number distinction on the following noun phrase.
- Questions: use inversion for questions (standard for written but not for spoken French) and avoid the *est-ce que*[5] construction.
- Determiners: determiners strongly influence the semantics of sentences, and all the possible translations do not necessarily have the same semantic effects. Moreover, there are no bare plurals in French. Table 5 sums up the determiner translations that were used.
- Proper names: proper names were translated, with a special care to cases to gender marking in case of referring expressions. For instance, in French, possessive determiners show gender agreement with the noun it determines, not with the possessor. But names of named entities such as companies have remained unchanged.

In the experiment we ran, we also asked the participants for feedback about the overall quality of the French text (see Sect. 4.1.).[6]

[5] It is more or less equivalent to *is it the case that*.

[6] This information, possibly consolidated by experiments with more participants, might be included to the resource.

each, *every*	*tout* (sometimes *tous les*)
most	*la plupart de*
bare plurals	*les*
few	*peu de*
neither	*aucun des deux*
many	*beaucoup*
no	*aucun*
some + singular	*un*

Table 5: Determiner translation

3.3. Examples of Translation Issues

The results of the design choices (the same as the ones of MultiFraCaS) on the semantic phenomena, as witnessed by the actual natural language inferences that can be drawn, have been evaluated through an experiment with French native speakers. Section. 4. describe this experiment and provides quantitative results. In this section, we focus on a qualitative analysis of some issues.

Table 6 illustrates the kind of issues related to the translation of determiners, where two different problems in English end up exactly the same in French.

P_1	Just one accountant attended the meeting.
Q	Did **any** accountant attend the meeting?
H	Some accountant attended the meeting.
A	Yes

P_1	Just one accountant attended the meeting.
Q	Did **some** accountant attend the meeting?
H	Some accountant attended the meeting.
A	Yes

Table 6: Different problems (108 and 110) that show no difference in their translation

Table 7 illustrates another difference between French and English. While the English problems seem to show that the position of the adverbial phrase influences the possible inferences, it does not seem to be the case in French, as confirmed by the experiment: both positions give rise to ambiguity.

P_1	**Since 1992** ITEL has made a loss.
P_2	It is now 1996.
Q	Did ITEL make a loss in 1993?
H	ITEL made a loss in 1993.
A	Yes

P_1	ITEL has made a loss **since 1992**.
P_2	It is now 1996.
Q	Did ITEL make a loss in 1993?
H	ITEL made a loss in 1993.
A	Don't know, on one reading of the premise

Table 7: Effects of adverbial phrases position that do not show up in French (problems 255 and 256)

Some problems do not even make sense in their French version. For instance, problem 116 in Table 8 shows an inference that relies on inferring the gender of an entity based on a possessive adjectives. However, in French, gender inflection of possessive adjectives depends on the gender of the possessee instead of the gender of the possessor. So, in French, the answer to problem 116 only depends on whether the first name used is more generally a first name for women or for men. Fully epicene fist names such as "Claude" or "Dominique" would resolve in a completely underspecified problem.

P_1	Mary used her workstation.
Q	Is Mary female?
H	Mary is female.
A	Yes

Table 8: Agreement of possessive adjectives (problem 116)

Finally, bare plurals do not exist in French and noun phrases have to have a determiner. Translating a bare plural requires adding a determiner that restrict the possible behaviors that bare plurals can exhibit in English. Table 9 shows an example where the translation of the bare plural *clients* requires using a determiner in French. This can be *les* (definite determiner) or *des* (indefinite determiner). With a definite determiner, the quasi-universal behavior of the bare plural is rendered (the expected answer is *Yes* as well), but it is definitely not the case with the indefinite determiner (*Yes* is not a possible answer).

P_1	Clients at the demonstration were all impressed by the system's performance.
P_2	Smith was a client at the demonstration.
Q	Was Smith impressed by the system's performance?
H	Smith was impressed by the system's performance.
A	Yes

Table 9: Bare plurals (problem 99)

Next versions of the resource will add French specific phenomena on adverb, tense and aspect, negation and specific determiners (*Tout*, *Quelque*, *Quel*, *Différents*, *etc.*) (Corblin and De Swart, 2004).

4. Experiment with French Native Speakers on the Test Suite

4.1. Conditions of the Experiment

The goal of this experiment was twofold. On the one hand, the goal was to test semantic intuitions of French native speakers about the inference tasks of the FraCaS test suite under conditions that enable a comparison with similar results for other languages (MultiFraCaS). On the other hand, we aimed at getting feedback from the participants about the quality of the translation and the complexity of the problems.

To this end, we set up an online survey (Fig. 2). Participants to this survey were presented problems in random order. Premises were introduced by *Sachant que* (*given that*), and the hypothesis was introduced by *je dirais de* (*I would say from this*). Participants then had to provide 3 answers.

The first one expressed the extent to which the participant agrees with the proposed inference, and three possible choices were given: *Vrai* (*True*) when the participant considered the hypothesis to be true when assuming the premises, *Faux* (*False*) when the participant considered the hypothesis to be false when assuming the premises, and *Pas assez d'information* (*Not enough information*) when the participant was not able to decide.

There might be different reasons why deciding is not possible, in particular because of scoping ambiguities. At this stage, we decided not to test the possible reasons, as the experiment design would have been much more complex.

The second one (section *Difficulté du problème* of Fig. 2) expressed the complexity of the inference, as perceived by the participant, on a four-level scale:

0 *Très facile* (*Very easy*);

1 *Facile* (*Easy*);

2 *Difficile* (*Difficult*);

3 *Très difficile* (*Very difficult*).

The third one (section *Qualité du français* of Fig. 2) expressed the extent to which the problem was stated in regular French. A four-level scale was used as well:

0 *Très mauvais* (*Very bad*);

1 *Mauvais* (*Bad*);

2 *Pas très naturel* (*Not very natural*);

3 *Tout à fait naturel* (*Completely natural*).

Participants were also provided the problem number (number 135 in Fig. 2) and the number of problems that they had already answered (6 in Fig. 2).

4.2. Participants of the experiment

There were 18 participants, with a relatively high level of education. Half of them were trained in logic and linguistics, and the other half did not have any specific related background. The results regarding answers to the inference tests (Sect. 4.3.) consist only of the answers from the 7 participants who answered all the 342 problems. The results about the complexity of the task and the quality of the French texts consist of the answers from the 12 participants who answered more than 10% of the problems. The total number of considered answers is 2,847 answers, with an average of 8.32 answers per problem. The selected annotators are not specialist of the task. They have different level of education.

4.3. Analysis of the Data about Inference

We computed the inter-annotator agreement between the 7 participants who answered all the problems, so that coefficients that are easier to interpret. We consider here the task as a 3-class classification problem and we observe the agreement among the 7 participants. We computed the coefficients with the NLTK library `nltk.metrics` package[7]. Krippendorff's α coefficient was computed in all settings, but it never differed significantly from Cohen's κ, so we only report κ below.

The observed agreement is $A_o = 0.734$, Scott's π is $\pi = 0.552$ and Cohen's κ is $\kappa = 0.553$. These values are quite low and does not show a high agreement. This was however expected, as the task of textual entailment is a difficult one in general, and maybe even more difficult with constructed data. It is also interesting to note that, among the 21 pairs of annotators, pairwise Cohen's κ varies in a somewhat wide range, from 0.463 to 0.642.

We also evaluated the maximum ratio of common answers for each of the problems. As Fig. 3 shows, the answers of 149 problems are all the same (100%), 90% of the answers are the same for 2 problems, and only 1 problem shows 33% of common answer. The latter is of course the lowest one as there are exactly three possible answers for each problem (*True*, *False*, and *Don't know*). This shows there is a maximal agreement for 43.5% of the problems, and 69% of problems have at least 75% of common answers. So despite the complexity of the task, there is a strong agreement on the answers for a large part of the problems.

4.4. Analysis of the Data about Complexity of the Task and Quality of the Texts

Quality was measured on a scale from 0 to 3 (0 is the lowest quality, and 3 is the highest one), and complexity as well (0 for easy tasks, and 3 for very difficult ones). The average for quality is 2.8 and the average for complexity is 0.53. This shows that the overall feeling of the participants about the problems is that they are expressed in a quite natural way, and rather easy to process.

Figure 4 presents the correlation between quality and complexity. Each point represents the average (complexity, quality) pair for a problem. The number of problems that get a given value is represented by the color of the point: the more problems have this value, the darker the point. 36 problems have been given a 3 quality (highest quality) and 0 complexity (lowest complexity) average.

Most problems stand in the upper left square of Fig. 4 which is our target square. More precisely, 302 problems, i.e., 88.3% of the problems, have a complexity less than 1 and a quality better than 2. We now focus on the remaining 11.7% of the problems (i.e., 40 problems) that we call LQ/HC (low quality of high complexity) problems.

We have looked in more detail at the distribution of these 40 LQ/HC problems over the 9 topics of the test suite, see table 4. Three sections have more than 10% of the LQ/HC problems: the *Anaphora* section (which is a small one), the *Temporal Reference* section, and the *Ellipsis* section. We note that for *Temporal Reference*, the last subsection consists of 13 complex problems, 6 of which are considered LQ/HC, which is not surprising. For the section *Ellipsis*, 12 over the LQ/HC problems belong to the *Ellipsis and Anaphora* subsection. Further analysis is required, in particular to check the extent to which these results are a con-

[7] https://www.nltk.org/api/nltk.metrics.html

Figure 2: Screenshot of the survey for one problem. English translations with orange background are not part of the interface but were added to the picture for explanatory purposes.

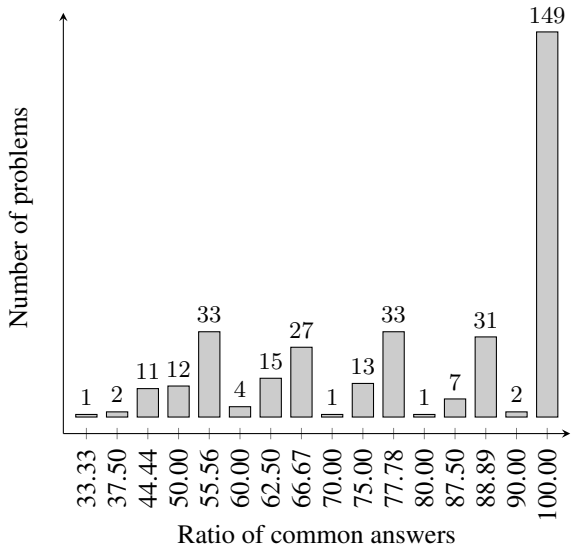


Figure 3: Number of problems for each ratio of common answers

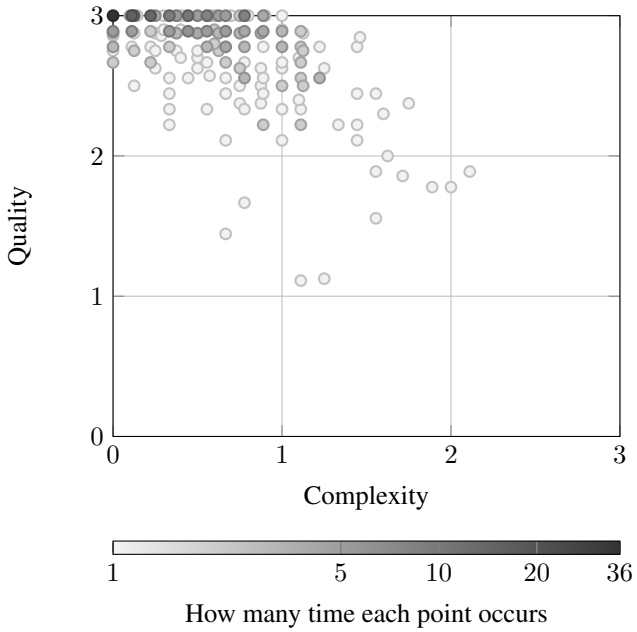


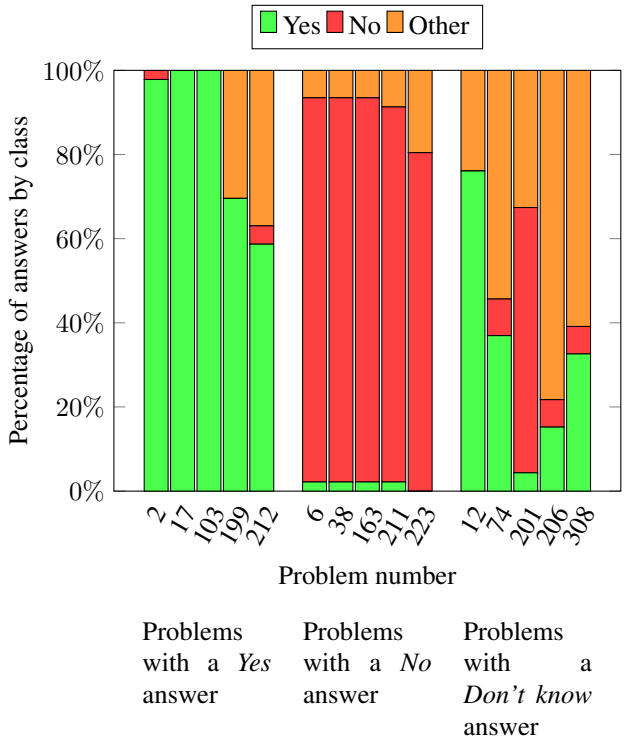
Figure 4: Correlation between complexity and quality

sequence of the differences between English and French to express of ellipsis and anaphora.

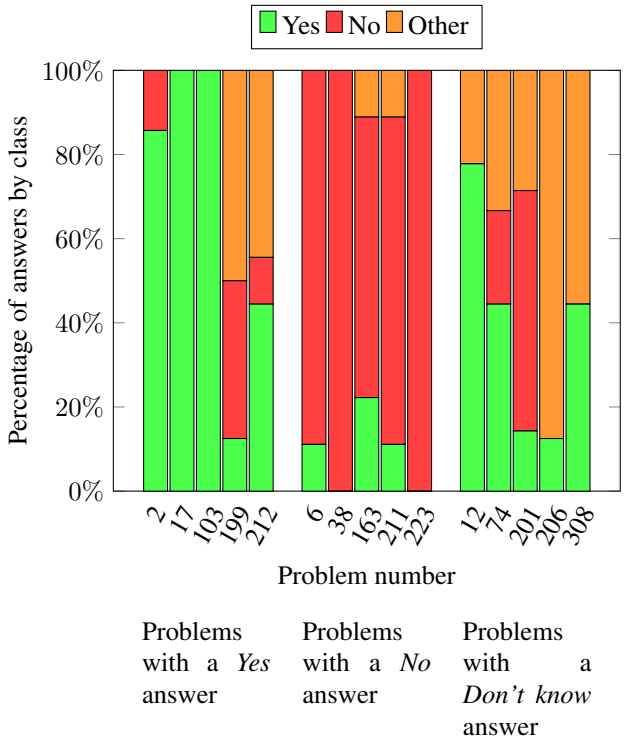
4.5. Cross-Lingual Evaluation of the Natural Language Inference Tests

Cooper et al. (2016; Chatzikyriakidis et al. (2017) describe a similar experiment that was run for the English, Greek and Slovenian test suites. They focus on 15 problems: 5 with a *Yes* expected answer, 5 with a *No* expected answer, and 5 with a *Don't know* expected answer. Table 10(a) shows some of their results, while Table 10(b) shows the result from the similar experiment we ran with the French test suite on the very same problems. According to the expected answers of the test suite, the left-hand part of both table should be green (100% of the answers are *Yes*), the middle part should be red (100% of the answers are *No*), and the right-hand part should be orange (100% of the answers are *Don't know*).

For problems expecting a *Yes* answer (problems 2, 17, 103, 199, and 212), both experiments show results that per-



(a) Experiment with MultiFraCaS (data courtesy of Chatzikyriakidis et al. (2017))



(b) Answers of participants on 15 selected problems

Table 10: Natural language inference by speakers

fectly fit the expectations for problems 17 and 103. For the other problems, despite some similarities, our results show a larger deviation from the expectation than the results of Chatzikyriakidis et al. (2017). In particular, problem 212, presented in Table 11, is a LQ/HC problem (average complexity is 1.44 and average quality is 2.77). It contains 4 premises and, beside possible cultural differences, highlights a tension between purely linguistic inferences and inferences build using background knowledge.

P_1	All mice are small animals.
P_2	All elephants are large animals.
P_3	Mickey is a large mouse.
P_4	Dumbo is a small elephant.
Q	Is Dumbo larger than Mickey?
H	Dumbo is larger than Mickey.
A	Yes

Table 11: Example on adjectives (Extensional Comparison Classes) problem 212

For problems expecting a *No* answer, results fit the expectation rather well in each experiments. For problems expecting a *Don't know* answer, both experiments show large differences with the expectation. Problem 201, presented in Table 12, illustrates both difficulties due to scope variations for adjectives and for translation (at least in French: *successful* has no obvious direct translation in this syntactic construction).

P_1	John is a former successful university student.
Q	Is John a university student?
H	John is a university student.
A	Don't know

Table 12: Example on adjectives (Affirmative and Non-Affirmative) problem 201

5. Conclusion and Future Work

This article presents the results of translating the FraCaS test suite into French. This resource gathers problems to test inferences carried by language. It is based on problem composed of hypothesis and conclusion. The translation was carried out in a systematic way, in order to obtain a result which is both homogeneous and faithful to the original data. The work was done by 5 translators expert of formal semantics, with cross validation. A lot of translating issues raised in the the process and were discussed. Such data are important because there exists so few ressources of that type. They are useful to entailment tasks, for example theorem-prover used with natural language.

The test suite has been evaluated with respect to the natural language inference expectations, to the complexity of the inferences, and to the quality of the translation, thanks to an experiment with 7 French speakers. In order to be able to make cross-lingual comparison, experimental conditions were similar to the ones described by Chatzikyriakidis et al. (2017). The goal is validate the answer propose in the ressource. This clearly highlights different levels of accept-

ability depending on the language, which highlights specificities for each of these languages. While problems with don't know answers have relatively similar understandings, some problems are interpreted with a completely different manner.

Futur works follow three different perspectives. First to pursue the tagging of the ressource by native speakers and propose a more fine grain analysis. The second perspective is to extend the annotation of the ressource by adding multilingual informations (like the acceptance), morphological tags, syntactic analysis and (logical) semantic representation. We also plan to relate and to extend the constructed data with real ones, and to develop evaluation of natural language inference tasks through gamification. Finally, we plan to extend the resource with more natural translations, in particular using a less specific vocabulary. We shall also include additional phenomena with French specific constructions and define aspects of French that have not been covered by this test or have not come to the fore. This is an important challenge for such a ressource.

6. Acknowledgments

We wish to thank Karën Fort, Manon Boudet, the participants of the experiment for their help on the survey; Robin Cooper, Stergios Chatzikyriakidis, Simon Dobnik for important discussions on the ressource. This work was supported partly by the french PIA project "Lorraine Université d'Excellence", reference ANR-15-IDEX-04-LUE. The survey was hosted in the CPER LCHN (Contrat de Plan État-Région - Langues, Connaissances et Humanités Numériques) infrastructure.

7. Bibliographical References

Bernardy, J.-P. and Chatzikyriakidis, S. (2017). A type-theoretical system for the FraCaS test suite: Grammatical Framework meets Coq. In Claire Gardent et al., editors, *IWCS 2017 - 12th International Conference on Computational Semantics - Long papers*. ACL anthology: W17-6801.

Chatzikyriakidis, S. and Bernardy, J.-P. (2019). A wide-coverage symbolic natural language inference system. In Mareike Hartmann et al., editors, *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 298–303. Linköping University Electronic Press. ACL anthology: W19-6131.

Chatzikyriakidis, S., Cooper, R., Dobnik, S., and Larsson, S. (2017). An overview of natural language inference data collection: The way forward? In *Proceedings of the Computing Natural Language Inference Workshop*. ACL anthology: W17-7203.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. (2018). XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485. Association for Computational Linguistics. ACL anthology: D18-1269. DOI: 10.18653/v1/D18-1269.

Cooper, R., Crouch, R., van Eijck, J., Fox, C., van Genabith, J., Jaspars, J., Kamp, H., Pinkal, M., Milward, D., Poesio, M., Pulman, S., Briscoe, T., Maier, H., and Konrad, K. (1996). Using the framework. Technical report, FraCaS: A Framework for Computational Semantics. ftp://ftp.cogsci.ed.ac.uk/pub/FRACAS/del16.ps.gz.

Cooper, R., Chatzikyriakidis, S., and Dobnik, S. (2016). Testing the FraCas test suite. Presentation at the Unshared Task "Theory and System analysis with FraCaS, MultiFraCaS and JSeM Test Suites" of Logical Engineering of Natural Language Semantics 13 (LENLS 13).

Corblin, F. and De Swart, H. (2004). *Handbook of French semantics*. CSLI publications.

Dagan, I., Glickman, O., and Magnini, B. (2006). The PASCAL recognising textual entailment challenge. In Joaquin Quiñonero Candela, et al., editors, *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190. Springer. DOI: 10.1007/11736790_9.

Kawazoe, A., Tanaka, R., Mineshima, K., and Bekki, D. (2017). An inference problem set for evaluating semantic theories and semantic processing systems for japanese. In Mihoko Otake, et al., editors, *New Frontiers in Artificial Intelligence*, number 10091 in Lecture Notes in Computer Science, pages 58–65. Springer. DOI: 10.1007/978-3-319-50953-2_5.

Korman, D. Z., Mack, E., Jett, J., and Renear, A. H. (2018). Defining textual entailment. *Journal of the Association for Information Science and Technology*, 69(6):763–772. DOI: 10.1002/asi.24007.

Ljunglöf, P. and Siverbo, M. (2011). A bilingual treebank for the FraCaS test suite. Technical report, Centre for Language Technology, University of Gothenburg. https://github.com/heatherleaf/FraCaS-treebank/blob/master/doc/FraCaSBank.pdf.

Ljunglöf, P. and Siverbo, M. (2012). A bilingual treebank for the FraCaS test suite. In Pierre Nugue, editor, *Proceedings of the Fourth Swedish Language Technology Conference (SLTC 2012)*. http://fileadmin.cs.lth.se/nlp/sltc2012/proceedings/sltc2012proceedings.pdf#page=62.

MacCartney, B. and Manning, C. D. (2007). Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 193–200, Prague, 06. Association for Computational Linguistics. ACL anthology: W07-1431.

MacCartney, B. and Manning, C. D. (2008). Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528, Manchester, UK, 08. Coling 2008 Organizing Committee. ACL anthology: C08-1066.

Pinkal, M. and Koller, A. (2012). Semantic research in computational linguistics. In Claudia Maienborn, et al., editors, *Semantics. An International Handbook of Natural Language Meaning*, number 33 in Handbooks of Linguistics and Communication Science, chapter 108, pages 2825–2859. Mouton De Gruyter. http://www.coli.uni-saarland.

de/~koller/papers/sem-handbook.pdf.
DOI: 10.1515/9783110253382.

Ranta, A. (2011). *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Studies in Computational Linguistics. CSLI Publications.