Predicting Linguistic Difficulty by Means of a Morpho-Syntactic Probabilistic Model

Philippe Blache and Stéphane Rauzy

Aix-Marseille Univ & CNRS
LPL, 5 Avenue Pasteur, 13100 Aix-en-Provence, France
{blache;rauzy}@lpl-aix.fr

Abstract. We propose in this paper a new contribution to the evaluation of linguistic difficulty. At the opposite of classical approaches relying on syntax, we show that a probabilistic morpho-syntactic analysis provides information enough to calculate different parameters, including *surprisal*. This method constitutes an original and robust model to linguistic complexity. It constitutes a solution towards complexity experiments using spoken languages.

Keywords: Linguistic complexity, difficulty model, computational psycholinguistics, sentence processing.

1 Introduction

Linguistic complexity is an important problem both from linguistic, computational and cognitive perspectives: it is important to know for a given language what are the phenomena or the constructions that can lead to a processing problem for a human subject. Such knowledge is the basis of a better understanding of the cerebral mechanisms of language processing. It is also necessary in the persepctive of the development of a cognitively-plausible automatic parser. Today, several computational models have been developed in this perspective. However, most of them are partial, only take into account local phenomena, and are experimented on highly controlled written text (cf. (Gibson, 2000), (Vasishth, 2003) or (Lewis and Vasishth, 2005) for example). More recent works propose to go one step forward in experimenting models on larger corpora (see in particular (Demberg and Keller, 2008), making it possible to validate complexity models. One of the conclusion of these works is that generic models have to bring together many different complexity parameters in order to take into account a large variety of phenomena. Moreover, it still remains necessary to develop and validate generic computational models dealing with unrestricted linguistic material, in particular spoken languages.

Among the different parameters involved in its definition, the evaluation of the difficulty human subjects encounter in language production or perception plays a major role.

In the literature, one of the most frequently used approaches relies on surprisal effects that can occur when integrating a new word to the structure (cf. (Hale, 2001)): roughly speaking, surprisal is high when the probability for a word to occur in a specific context is low. This parameter is evaluated as the negative logarithm of the probability of a word w_t to appear in its context, taking into account the sequence $w_1...w_{t-1}$.

Surprisal is evaluated as a function of the probabilities of the possible parses when integrating the word w_t. When using a probabilistic grammar, evaluating surprisal consists in measuring the difference between the probability of all parses spanning the sequence $w_1...w_{t-1}$ and those spanning $w_1...w_t$:

$$SI_t = -\log \frac{P(w_1...w_t)}{P(w_1...w_{t-1})} \quad (1)$$

Surprisal has been shown by different studies as being correlated with difficulty. This has been demonstrated experimentally, in particular by several works relying on eye-tracking in reading tasks (cf. (Boston *et al.*, 2011), (Demberg and Keller, 2008) or (Mitchell *et al.*, 2010)). These last studies use results from the Dundee corpus (see (Kennedy *et al.*, 2003)) from which some short extracts have been chosen (20 texts read by 10 subjects). This corpus has been analyzed, surprisal values calculated and put in perspective with eye-tracking data. Results show a correlation between difficulty parameters (in particular reading time) and surprisal indexes.

One of the problems is that surprisal evaluation is sensible to the syntactic formalism. Moreover, from a computational point of view, surprisal also depends on the characteristics of the parser used in the evaluation. Surprisal being measured in terms of the probabilities of the possible parses, it is necessary first to build a probabilized grammar (which can be problematic) and second to correctly generate all candidate parses. We know that when parsing unrestricted linguistic material, it is not always possible to build parse trees. Robustness and precision are then pre-requisite properties of the parsers when evaluating surprisal. This also means that, because of the limitation of parsing technologies, it is not possible to measure difficulty for non-canonical or ill-formed construction (not to speak of spoken language).

We propose in this paper to explore a new direction for difficulty evaluation: syntax being problematic when dealing with unrestricted material, we propose to analyze precisely the contribution of morpho-syntactic information in surprisal evaluation. Our hypothesis is that surprisal effects appear at the lexical level, as corroborated by works on the impact of the lexicon on ambiguity or reanalysis phenomena (cf. (Fodor and Ferreira, 1998)). We propose to measure surprisal effects only by means of a probabilistic categorization task, without using parsing information. We use in this perspective a probabilistic morpho-syntactic tagger, taking into account at each step a variable-size context.

In the first section, we present the probabilistic method thanks to which we evaluate the different parameters entering in the difficulty model. In the second section, we describe how to calculate the parameters. Finally, the last section will present some results.

Before entering into the description of our approach, we want to underline the fact that this paper presents a theoretical proposal which still has to be validated experimentally. However, we propose in the last section preliminary results calculated from a large corpus showing a positive tendency.

2 The statistical model

The mechanism consists in evaluating a probability to any sequence of *tokens* (or part-of-speech). Each token is associated, thanks to a lexicon, to its corresponding lexical tags distribution, where tags take their values in the set of morphosyntactic categories $\mathcal{C}$. An example of tags distribution is illustrated table 1.

Form	Lemma	Sampa	Tag	Proba.
est	être	E	`Auxiliary`	0.39
est	être	E	`Verb`	0.52
est	est	Est	`Noun`	0.09

Table 1: Tags distribution for the french lexical form *"est"* (meaning *"is"* in its verb or auxiliary use, or *"east"* in its noun use).

The tagging process is usually very ambiguous. This ambiguity generates an exploration tree space of potential solutions $\mathcal{S}ol$. Its cardinality card($\mathcal{S}ol$) is a function of the combinatory associated with the potential choices for each token of the sentence. The tokens being integrated step by step to the sentence, the cardinality card($\mathcal{S}ol(t)$) at position t in the sentence increases (or remains stable for tokens with one-to-one tag correspondence). For a given sentence, we denote hereafter by $S_i(t)$ the i^{th} solution at position t ($i \in [1, \texttt{card}(\mathcal{S}ol)]$),

$$S_i(t) = c_{1,i}...c_{t,i} \tag{2}$$

where $c_{t,i}$ is the morphosyntactic category associated to the token at position t for solution $S_i(t)$. The probability of the solution $S_i(t)$ is obtained recursively by Bayes formulae :

$$P(S_i(t)) = P(c_{t,i}|S_i(t-1)) \times P(S_i(t-1)) \tag{3}$$

where $P(S_i(t-1))$ is the probability of the solution i at position $t-1$ and $P(c_{t,i}|S_i(t-1))$ is the conditional probability of category $c_{t,i}$ given the right context $S_i(t-1) = c_{1,i}...c_{t-1,i}$.

In order to calculate the probability of each solution of the space of solutions $\mathcal{S}ol$, we need a statistical model predicting the probability of each category c_t conditioned by each right context $c_1...c_{t-1}$, for all the combinatory cases found for each solution of the exploration tree :

$$\texttt{Model} \Rightarrow P(c_t|c_1...c_{t-1}) \quad , \quad c_1, ..., c_t \in \mathcal{C} \tag{4}$$

The probabilistic model may be limited to the sole use of morphosyntactic information (that is the strategy experimented in this paper), or makes use of higher syntactical information such like structure constituents and relations describing the context.

Given a model, the conditional probabilities of equation 4 allow to calculate for each added token the relative contribution of each solution $S_i(t)$ belonging to the space of solutions. The density function $\rho(t)$ associates to each solution at position t its relative probability $\rho_i(t)$:

$$\rho_i(t) = \frac{P(S_i(t))}{A_t}, \texttt{ with } A_t = \sum_{\mathcal{S}ol} P(S_i(t)) \tag{5}$$

where $P(S_i(t))$ is computed from the model given equation 4 and the normalization factor A_t warrants the equality $\sum_{\mathcal{S}ol} \rho_i(t) = 1$. The evolution of the space of solutions and of its density is investigated in the following section. It provides us with the different parameters proposed for measuring the processing difficulty of a sentence at the morphosyntactic level.

The probabilities mentioned equation 4 are herein obtained using the *Pattern Model* (Blache and Rauzy, 2006), a Hidden Markov Model (HMM) more efficient than the standard *N-grams* models. For *N-grams*, the states of the automaton are characterized by sequence of categories of equal size $N-1$. The pattern model relaxes this constraint by accepting states identified by sequences of varying size. This property allows in practice to learn from the training corpus a set a states, the *patterns* of the model, which capture in an optimal way the morphosyntactic regularities found in the training corpus. The patterns model is seen as an approximation of the real statistical model describing the data, the approximation becoming more and more accurate as the training corpus size and coverage become larger.

The model is trained on the *GraceLPL* corpus, a version of the *Grace/Multitag* corpus (see (Paroubek and Rajman, 2000)) corrected by us. It is a French corpus containing about 700 000 words with morphosyntactic annotation following the tagset features Multext (Ide and Véronis, 1994). GraceLPL is regularly corrected and enriched in order to improve its tagging. The morphosyntactic information has been organized in an ad-hoc way in 50 tags (4 types of categories for punctuation marks, 1 for interjections, 2 for adjectives, 2 for conjunctions, 1 for determiners, 3 for

nouns, 8 for auxiliaries, 4 for verbs, 5 for prepositions, 3 for adverbs and 15 for pronouns). In its current form, the tagset does not include gender, number and person features distinction nor tense and mood information for verbs.

The pattern model describing our tagger is composed of 2 841 patterns of varying size (the largest right context in the list of patterns counts 8 categories). The evaluation of the model is performed by comparing the tagged output with the reference. For the selected tagset of 50 categories mentionned above, the quality of the tagger (version 2011) reaches a score of 0.975 (F-measure).

3 The morpho-syntactic difficulty model

Surprisal is usually calculated thanks to a probabilistic grammar in building the set of possible parses (the number of parses being possibly limited to a given number). We propose to replace parses by the study of part-of-speech sequences, without any other information on structure nor relations.

The evolution of the surprisal with position t is herein defined for each solution $S_i(t)$ presented equation 2 as follows:

$$SI_i(t) = -\log \frac{P(S_i(t))}{P(S_i(t-1))} + \log(P(c_{t,i}|.)) \tag{6}$$

where the correction term $P(c_{t,i}|.)$ stands for the non-contextual probability of the morphosyntactic category $c_{t,i}$ associated to token at position t for solution $S_i(t)$. Following equation 3, the equation rewrites :

$$SI_i(t) = -\log \frac{P(c_{t,i}|c_{1,i}...c_{t-1,i})}{P(c_{t,i}|.)} \tag{7}$$

which is the negative logarithm of the ratio between contextual and non-contextual probabilities for category $c_{t,i}$. The higher this ratio, i.e. the category has a strong probability to occur within this context, the more negative is the surprisal index. This term aims at correcting asymmetry in category distribution. Some of them are densely populated (e.g. common noun) whereas some other are not (e.g. present participle of the auxiliary) without any consequence on human processing difficulty. With this correction, the overall surprisal $SI(t)$ is the same as proposed in the literature (see for example (Demberg and Keller, 2008)):

$$SI(t) = \sum_{\mathcal{S}ol} \rho_i(t)\, SI_i(t), \tag{8}$$

which is the weighted average of individual surprisal values for each solution, weighted by their relative contribution $\rho_i(t)$ (see equation 5). Surprisal measures a function of these two difficulty parameters: one related to sparseness of the POS sequences, the other coming from the number of solutions to take into account at the same time (consequence of the lexical forms ambiguity).

This observation leads us to propose a second parameter making it possible to separate these two effects. The consequence of morpho-syntacic ambiguity on processing difficulty can be directly measured. A good indication is for example entropy variation of the probability distribution $\rho_i(t)$ described equation 5:

$$AH(t) = -\sum_{\mathcal{S}ol} \rho_i(t)\ \log \rho_i(t) \tag{9}$$

$$AI(t) = AH(t) - AH(t-1) \tag{10}$$

The value $AH(t)$ measures evolution of the effective size of solution space. Its derivative across time, the ambiguity index $AI(t)$, shows that integrating the lexical form at a position t locally contribute to solve ambiguity ($AI(t) < 0$) or to increase it ($AI(t) > 0$).

4 Experimentation

As mentioned in the introduction, this paper present a preliminary study and many work is still to be done in order to validate our hypothesis. First, we have to verify that our model is a good predictor for difficulty. In this perspective, corpora with eye-tracking data have to be collected (for French in particular). Moreover, a systematic comparison with results obtained with syntactically-based surprisal values has to be done. Finally, new kinds of data have to be created in the perspective of validating our model against spoken language corpora.

However, we propose in this section a preliminary study on real data, giving some indications on the efficiency of the model.

Our study relies on a French corpus, made of 317,258 tokens (14,545 sentences), collected from the GRACE corpus mentioned before (automatically generated and manually corrected). The sentence selected from the initial corpus are those which have no difference between the reference categorization (that of the corpus) and that of our tagger. For each token, *surprisal* and *ambiguity* values have been calculated. Moreover, each sentence is associated to a revision rate, which relies on the number of revisions (or backtrackings) occurring during the tagging process. More precisely, the *revision rate* is the ratio between the number of solutions with a maximal probability that have been abandoned and the number of tokens in the sentence.

4.1 Model Validation

We propose in this section to verify the validity of our proposal in applying *surprisal* and *ambiguity* parameters to two phenomena known to be complex: relatives (subjects vs. objects) and clitics (nominative vs. accusative).

Relatives Several psycholinguistic studies (cf. (Gibson, 1998; Gibson, 2000), (Just and Carpenter, 1992)) have shown the that object relatives are more complex in terms of processing than subject relatives. Different analysis have been given for this: subject relatives follow the canonical SVO order, object relatives require to memorize more information, etc. The complexity differences have been shown experimentally by means of reading times (see for example (Lewis and Vasishth, 2005), (Demberg and Keller, 2008)).

We obtain the same prediction with our morpho-syntactic model: object relative pronoun has a surprisal value significantly higher than subject. Moreover, we also observe a relation between surprisal and ambiguity: the object pronoun is more ambiguous than the subject and causes a higher surprisal.

POS	Occur.	Surprisal	Ambiguity
Subj	2507	-1,438	-0,076
Obj	1228	-0,953	0,202

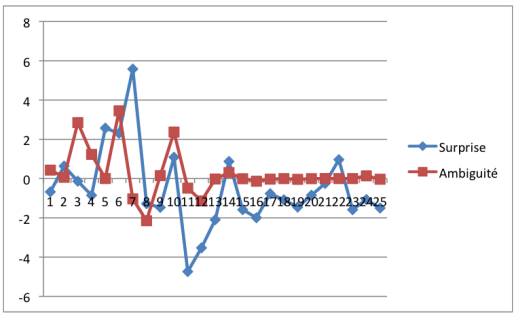
Clitics Language acquisition studies show a later acquisition of accusative clitics than nominative ones (see (Jakubowicz, 2003)), underlying higher complexity for the accusative case. This difference is predicted by our model: accusative clitics have surprisal values twice higher than nominative ones.

POS	Occur.	Surprisal	Ambiguity
Nomin.	9682	-1,586	0,003
Acc.	1857	-0,701	0,261

These results tend to confirm our hypothesis stipulating that surprisal evaluation on a morpho-syntactic basis is a good predictor of processing difficulty, as the syntactic-based surprisal model.

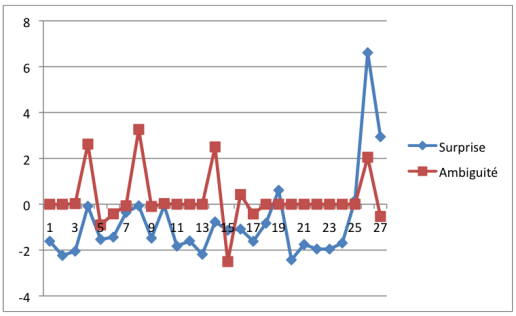
4.2 Surprisal

The following examples underline the effects of surprisal and ambiguity values, obtained on a morpho-syntactic basis. The figures show the evolution by words of the two parameters. In the first example, the maximal surprisal value is reached with the preposition "*de*", located between two determiners (a cardinal and a demonstrative). In this case, category sequence sparseness is the cause of the surprisal, more than ambiguity. This last parameter show a peak on the previous token, whereas the token associated to the surprisal peak corresponds to a negative ambiguity value.



"Ainsi pour Le Seuil, quarante-sept de ces droits ont t acquis par des diteurs de langue allemande."

"So, for Le Seuil, 47 of these rights have been acquired by German-speaking editors."



"On parle aussi d'étoiles de l'Opéra : Noella Pontois, qui est à la retraite, Michael Denard, qui va bientôt l'être ..."

"We also talk about Opera stars: Noella Pontois, who is retired, Michal Denard, who will be soon too ..."

Figure 1: Surprisal examples

In the second example, the surprisal peak is located on the infinitive verb "*être*". This value is related to the categorization of the preceding word "*l'*" as a determiner. It reveals a syntactic difficulty related to the elliptic construction. Moreover, at the difference with the previous example, surprisal comes with high ambiguity value.

These two examples show that the contribution of the ambiguity parameter to surprisal is not direct. Nevertheless, we have observed on the entire corpus a correlation between surprisal and ambiguity (Pearson correlation coefficient of 0.4).

Ambiguity, parallelism: A frequently discussed question in the literature concerning human processor is to know whether it works on a sequential or parallel manner (cf. (Gibson, 1991; Boston *et al.*, 2011)). This question is important and has consequences on complexity study, and more particularly on memorization costs. A parallel parser requires to store all possible parses at the same time whereas a sequential parser memorizes less states, but relies on backtracking. This question is still open, even with the massive use of probabilistic approaches in language processing that could militate in favor of a parallel processing.

(Brants and Crocker, 2000) have shown that the choice of an adequate hierarchization factor renders the parallelization process almost useless. The study of revision rates obtained in the corpus tends to confirm this result: surprisingly, the categorization process is almost deterministic. More precisely, we measured on the corpus 7,133 sentences (49% of the total) that do not present backtrack: the choice of the most probable solution at each step leads to the right solution. The overall revision rate is 4.3%.

In other words, the tagger behavior confirms the hypothesis of a very limited parallelism in language processing. It is an interesting result not only from a cognitive perspective, but also for

automatic language processing: parsing techniques relying on a "*beam search*" strategy seems to constitute an adapted answer.

5 Conclusion

The analysis of language processing difficulty by human subjects necessarily has to take into account language in *natural situation*, which means the necessity of treating spoken language. The methods and models used in such studies must then deal with different uses in order to be general and reusable, which is not the case with current difficulty models. In particular, evaluating automatically difficulty parameters means tools able to process non canonical or even ill-formed sentences.

The solution we propose in this paper consists in using low-level information: morpho-syntax features can been calculated with a good performance, even for spoken language transcriptions. This means that automatic tools can always provide some information, whatever the structure (which is not the case with syntactic models). Our approach constitutes a first answer towards natural language difficulty modeling for mainly for two reasons:

- The approach is formalism-independent and theoretical-neutral: evaluating difficulty on the basis of morpho-syntactic information avoids the use of syntactically-annotated resources as well as parsers. It becomes the possible to adapt this technique to other languages, even those without syntactic resources.
- Robustness: the efficiency of modern taggers makes it possible to process all kinds of linguistic material, including spoken language.

The results we obtained on corpus analysis have to be generalized by means of experimentations on human subjects leading to reading-time, eye-tracking and evoked potentials data.

References

Blache, P. and S. Rauzy. 2006. Control mechanisms for parsing in property grammars. In *Proceedings of TALN*.

Boston, M. F., J. T. Hale, S. Vasishth, and R. Kliegl. 2011. Parallelism and syntactic processes in reading difficulty. *Language and Cognitive Processes*, 26, 301–349.

Brants, T. and M. Crocker. 2000. Probabilistic parsing and psychological plausibility. In *Proceedings of the 18th International Conference on Computational Linguistic*, Saarbrcken/Luxembourg/Nancy.

Demberg, V. and F. Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. In *Cognition*, volume 109, Issue 2, pp. 193–210.

Fodor, J. D. and F. Ferreira. 1998. *Reanalysis in sentence processing*. London: Kluwer Academic Publishers.

Gibson, E. 1991. A computational theory of human linguistic processing: Memory limitations and processing breakdown. In *Doctoral dissertation*. Carnegie Mellon University.

Gibson, E. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68, 1–76.

Gibson, E. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In *Image*. A. Marantz, Y. Miyashita, W. O'Neil (Edts).

Hale, J. 2001. A probabilistic earley parser as a psycholinfuistic model. In *Proceeding of 2nd Conference of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, PA.

Ide, N. and J. Véronis. 1994. MULTEXT: Multilingual text tools and corpora. In *Proceedings of the 15th. International Conference on Computational Linguistics (*COLING 94*)*, volume I, pp. 588–592, Kyoto, Japan.

Jakubowicz, C. 2003. Computational complexity and the acquisition of functional categories by french-speaking children with SLI. *Linguistics*, 41(2).

Just, M. A. and P. A. Carpenter. 1992. A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99(1), 122–149.

Kennedy, A., R. Hill, and J. Pynte. 2003. The dundee corpus. In *Proceedings of the 12th European Conference on Eye Movements*.

Lewis, R. L. and S. Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29, 375–419.

Mitchell, J., M. Lapata, V. Demberg, and F. Keller. 2010. Syntactic and semantic factors in processing difficulty: An integrated measure. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 196–206.

Paroubek, P. and M. Rajman. 2000. Multitag, une ressource linguistique produit du paradigme d'évaluation. In *Actes de Traitement Automatique des Langues Naturelles*, pp. 297–306, Lausanne, Suisse, 16-18 octobre.

Vasishth, S. 2003. Quantifying processing difficulty in human sentence parsing: The role of decay, activation, and similarity-based interference. In *Proceedings of the European Cognitive Science Conference 2003*.