QÜESTIIÓ, vol. 21, 1 i 2, p. 241-266, 1997

REDACTION AUTOMATIQUE DE COMMENTAIRES A PARTIR DE BASES DE DONNÉES STATISTIQUES

Dr. JEAN LOUIS ROOS*

Institut National de la Statistique et des Etudes Economiques

ALIEN is a knowledge based system that modelizes the economist's behaviour It is used at the French Statistical Institute and at Eurostat. It can be seen as an assistant that help economists to write economic diagnosis in natural language as well as a human economist could do.

Alien is very useful for analysis in very large data base and/or for regular diagnosis. It study the shape of all series in an application, the instantaneous relationship and determine what is important and what is not.

ALIEN is used under a user friendly software (DARWIN) developed for Windows 95. DARWIN build text generation with ALIEN, but also construct graphs and table, help the user to analyse the data base and the semantic of an application. Last, DARWIN build a nice layout of the mixed comment, graph and table with WINWORD or an equivalent software.

Text generation from statistical data

Keywords: Text Generation, data Base, Artificial Intelligence, Automatic economic diagnosis

*Dr. Jean Louis Roos. Institut National de la Statistique et des Etudes Economiques. Avenue Albert Einstein BP 26000, 13791 Aix En Provence Cedex 3. Tel 42 16 29 59. Fax 42 16 29 00. roos@cil3a.insee.atlas.fr

– Article rebut l'octubre de 1996.

– Acceptat el febrer de 1997.

Les bases de données statistiques sont de plus en plus fréquemment de très grandes tailles. Lorsqu'un statisticien doit rédiger des commentaires réguliers sur les données contenues dans de telles bases, son travail peut être particulièrement long et difficile car il devra examiner une masse considérable de chiffres. Qui plus est, les données possèdent une sémantique qui n'apparaît jamais (ou que très rarement) dans les bases, donc le commentaire que le statisticien fera dépendra en grande partie de ses propres connaissances. Dès lors, tout changement du responsable chargé d'une étude statistique régulière aura des conséquences parfois dramatiques sur la qualité des textes d'analyse. Curieusement, les fabriquants de logiciels ont privilégié le développement d'outils puissants d'interrogation, mais rien n'a été fait en ce qui concerne l'interprétation des données contenues dans de telles bases.

A notre avis, et dès maintenant, les statisticiens ont besoin d'un outil qui doit répondre à trois préoccupation:

- **Disposer d'un assistant**

  Régulièrement, un statisticien doit pouvoir demander à son assistant un texte correctement rédigé qui analyse l'information dans une base de données. Ce texte doit pouvoir s'accompagner de tableaux et de graphiques. Il doit être aisément modifiable à travers un traitement de texte. L'assistant doit être suffisamment efficace pour ne rien oublier d'important et, évidement, ses analyses doivent être «stables» dans le temps, et totalement indépendantes de facteurs subjectifs extérieurs.

- **Pouvoir traiter des commentaires de masse**

  Il est fréquent de ne pas réaliser des commentaires à partir de la totalité de l'information dont on dispose. Ceci simplement parce que la masse d'information à traiter est trop important; c'est le cas lors d'analyses sectorielles, par communes, par pays, etc. Un être humain ne peut faire face —parfois simplement pour des raisons de temps— à la rédaction de dizaines, voire de centaines de diagnostics. Les difficultés seront encore plus lourdes si les rédactions doivent s'accompagner de graphiques et être mises en page.

- **Disposer de références**

  On appelle références (guidline) la définition, dans des domaines divers, de recommandations de «bonnes pratiques». Une bonne pratique est ce qu'il faut faire, dire, écrire face à une situation donnée. Ces «guidlines» ont été surtout développés en médecine, à partir de travaux de formalisation de diagnostics ou de traitements sur lesquels des consensus existaient. En simplifiant, une référence permet de décrire une situation et un traitement de celle-ci formalisés et acceptés par une majorité des praticiens. Ce consensus, certes très relatif, fournit une réponse efficace et argumentée face à une maladie donnée.

Pour un statisticien, il est tout aussi important de disposer de telles références associées à chaque base sur laquelle il travaille. Sinon, les risques d'erreurs dans l'interprétation des données peuvent être importants. Dans le cadre de travaux spécifiques de formalisation de l'analyse conjoncturelle, l'**Insee** a commencé à développer un logiciel répondant à ces préoccupations. **Eurostat** s'y est très vite associé et a assuré une grande partie du financement. Des utilisations sont maintenant en cours tant à l'**Insee** qu'à **Eurostat**.

Nous allons présenter ici successivement le logiciel de génération de texte: **Alien**, et son environnement de fonctionnement sous Windows: **Darwin**. L'ensemble se présente comme un assistant «informatique» pour le statisticien. L

1. LA GENERATION DE TEXTE: ALIEN

Alien est un logiciel qui formalise —en réalité qui modélise —les mécanismes qui amènent un statisticien à «rédiger» un diagnostic à partir d'un ensemble de données quantitatives. Cette formalisation repose sur l'observation suivante: la plupart des textes d'analyse simples ont un aspect «stéréotypés» —et ce quelle que soit la langue dans laquelle ils ont été écrits— il existe donc des schémas d'analyse très répétitifs, et il est alors possible de les modéliser.

1.1. Les bases theoriques

Les fondements d'**Alien** sont assez simples: le statisticien ne manipule plus des séries mais des indicateurs; ceux-ci sont des objets incorporant une sémantique. Deuxièmement, le statisticien utilise des formes de discours stéréotypés. Au sein de ces formes, des dictionnaires permettront la génération de texte et le changement de langues.

1.1.1. Les indicateurs

Alien repose sur l'idée que l'économiste ne manipule pas des séries statistiques, mais plutôt un ensemble de connaissances sur celles-ci, y compris de connaissances sur le vocabulaire associé à chaque série. Ces connaissances ont deux origines: au départ, et pour une application donnée, le système possède une connaissance minimale standard; cette connaissance est ensuite développée par un processus d'apprentissage qui repose sur un mécanisme d'inférence.

Alien range la connaissance utile dans des «objets» que l'on nomme les «indicateurs», et qui englobent, après une étape d'apprentissage, toute l'information indispensable pour la rédaction d'un diagnostic.

Les connaissances «ex ante» seront, par exemple, le nom des séries à analyser, différents codages sur la sémantique possible de la série (par exemple: la série peut-elle être interprété en variation? est-ce un flux ou est-ce un stock?,...), mais aussi le vocabulaire, parfois le «jargon», le plus courant associé à l'application.

Les connaissances «ex post» résultent d'opérations de calculs et permettent d'interpréter les valeurs en niveau et en variation pour les périodes à étudier. Ces «jugements» sont liés aux valeurs passées des séries.

Le plus important, pour obtenir des commentaires de qualité, est de bien coder certaines valeurs ex-ante. C'est un travail délicat, mais identique à celui que fournit l'être humain lorsqu'il rédige un commentaire; bien souvent nous manipulons des séries sans connaître toute leur signification, et toute erreur sur la sémantique de départ peut être ensuite une cause d'erreur dans l'analyse que l'on en fait.

1.1.2. Les discours stéréotypés

La richesse d'une langue est telle qu'il est illusoire, sans moyens importants, de définir une méthode qui permettrait d'écrire tous les discours possibles pour un diagnostic donné. Cependant l'usage montre que, lors de la rédaction de textes de commentaires statistiques, une certaine paresse, ou tout au moins l'habitude, fait que les discours utilisés sont fortement stéréotypés. En voici deux exemples choisis dans la presse:

> *Sales of new cars in western Europe increased by 1.2 per cent last month, with higher demand in Germany and France compensating for a heavy decline in sales in Italy. Sales in the first 11 months of the year, at 12.53m, were 1.3 per cent lower than in the corresponding period a year earlier.*
>
> *Consumer optimism about the economy rose slightly in August from July. But the gain was small: from 77 in July to 77.3 in August.*

Ceci étant, d'une part il existe de nombreuses variantes à de semblables discours, d'autre part il est possible de rédiger certains discours plus spécifiques dans des situations particulières. Qui plus est, le discours stéréotypé reste insuffisant pour faire une rédaction de qualité; dans un domaine particulier, par exemple la finance, la conjoncture, le commerce,... nous pouvons observer des «dialectes», c'est-à-dire un vocabulaire, parfois une forme lexicale particulière, qui est spécifique au sujet traité. **Alien** prendra en compte cette préoccupation.

1.1.3. Les dictionnaires

Les dictionnaires jouent plusieurs rôles dans la rédaction d'un texte: évidement, ils fournissent avant tout le vocabulaire, mais ils indiquent aussi la plupart des règles

grammaticales courantes: les formes de pluriels par exemple ou encore les conjugaisons. C'est ce qui se passe avec les dictionnaires d'**Alien**. Et tout comme un être humain, **Alien** utilisera aussi des dictionnaires de synonymes, ou encore des dictionnaires de dialecte dont nous venons de parler.

1.2. Le systeme formel: une double arborescence

L'écriture est une activité humaine particulière qui implique de fréquent «retours en arrière», c'est aussi le cas d'une analyse statistique. De ce fait, il est difficile de construire un logiciel totalement déterministe. Alien se présente en pratique comme un système organisé autour d'une double arborescence d'informations, la rédaction étant traitée par une série de petits traitements organisés sur les noeuds terminaux et utilisant les dictionnaires. Le résultat est un logiciel indéterminisme «a priori» sur un discours précis.

1.2.1. L'arborescence des indicateurs

Un indicateur principal définit le sujet traité, il possède comme descendance des indicateurs *domaines* qui constitueront les paragraphes de textes. Chaque domaine est constitué d'indicateurs quantitatifs ou qualitatifs dont certains possèdent des occurrences (par exemple la production peut être analysée par branche, par pays,...). Ces indicateurs forment les discours sur chaque série à commenter. La hiérarchie sujet, domaine, indicateur final peut être comparée à celle d'un modèle économétrique où les domaines sont des blocs et les indicateurs sont des équations.

1.2.2. L'arborescence du contenu des indicateurs

Tous les indicateurs sont structurés de la même façon: ils possèdent des propriétés, et pour chacune d'elle, existe une valeur (parfois par défaut). Un indicateur possède environ deux cents propriétés. Il est possible, pour chaque application, de rajouter des propriétés originales.

1.2.3. Les dictionnaires

Ils forment une arborescence élémentaire à partir de points d'entrées. Les dictionnaires sont des objets particuliers qui regroupent le vocabulaire commun aux indicateurs, mais aussi la majeure partie des règles grammaticales associées et des conjugaisons. Le contenu des dictionnaires est défini par un index qui est un type abstrait particulier. On trouvera les traitements suivants:

- Les conjugaisons: l'entrée est un verbe, ou un groupe verbal, à l'infinitif. Sont conjugués: le présent, le passé, et le conditionnel.

- Les accords: à chaque mot, ou groupe de mots, correspond le même ensemble au masculin singulier, féminin singulier, masculin pluriel, féminin pluriel, neutre singulier, neutre pluriel (ces deux cas sont pour d'autres langues), ainsi que deux cas spécifiques (par exemple en français l'élision avec h).
- Les synonymes: six synonymes sont donnés pour chaque mot ou expression.
- Le vocabulaire complémentaire: il est constitué de mots isolés, principalement invariables: prépositions, conjonctions, adverbes, etc.

1.3. Les solutions informatiques

1.3.1. Les indicateurs

L'objet que manipule le statisticien est un indicateur, c'est-à-dire une structure dans laquelle se trouve rangée toute l'information indispensable pour interpréter une (ou plusieurs) série(s). En pratique, une grande partie de l'information étant partageable, un indicateur ne porte que l'information qui lui est spécifique. Le partage de l'information se fait ensuite par en grande partie par héritage - ce qui implique l'organisation hiérarchique des indicateurs - mais aussi par accès à des dictionnaires. Il existe ainsi une information commune à tous les indicateurs.

Le contenu d'un indicateur, et donc l'information qu'il peut porter, est susceptible de modifications d'une application à une autre, ou encore en fonction de développement du système; les indicateurs sont donc construits à partir de prototypes, c'est-à-dire de modèles, aisément modifiables. Ces modèles définissent un objet organisé d'une certaine façon, avec des attributs et des valeurs. Ce sont donc des types abstraits ou des prototypes.

Le modèle «indicateur» décrit, sous une forme arborescente, ce que doit contenir un indicateur. Chaque noeud joue un rôle d'attribut. Ces attributs sont en nombre variables. A un attribut peut être associé un sous-arbre ou une valeur. Les valeurs associées dans les indicateurs instances sont aussi modifiables. Voici quelques-uns des principaux attributs existants:

- des types, permettant de classer chaque indicateur. A chaque groupe d'indicateurs sera associé des traitements, et parfois un discours possible particulier.
- des noms pour l'indicateur,
- des informations sur la périodicité de la, ou des séries, propre(s) à l'indicateur,
- des formats d'impression de valeurs,

- du vocabulaire particulier: verbes, adjectifs, adverbes, prépositions, et des règles grammaticales,
- des listes d'indicateurs occurrences,
- des formules de calcul expliquant comment est obtenue la série de base. C'est parfois une différence (un solde par exemple), parfois une somme de plusieurs séries, voire une somme pondérée. Il y a aussi les formules décrivant les variations, lesquelles peuvent être calculées de nombreuses façons et avec différents décalages; enfin, on trouvera quelquefois des formules d'accélération.
- des index de jugement, et des outils permettant de les construire,
- la description des périodes à analyser en terme de variation et éventuellement d'accélération (ou de décélération)
- les valeurs et les noms de la, ou des séries, propre(s) à l'indicateur,
- des programmes et des traitements.

Cette liste n'est pas exhaustive. Chaque utilisateur peut rajouter de l'information. En pratique, plus de deux cents attributs définissent ainsi la sémantique utile d'un indicateur.

1.3.2. Les traitements

Pour obtenir la rédaction d'un texte de diagnostic par un logiciel, il est indispensable de donner à celui-ci la sémantique de l'analyse à faire; il est cependant inconcevable de fournir la totalité de celle-ci; c'est le système lui-même qui doit la construire par un processus d'apprentissage. Le statisticien ne fournira que des informations minimales telles le nom et les valeurs des séries à analyser. Dans **Alien** trois étapes de traitements se succèdent; La première est la construction de l'arborescence des indicateurs et le remplissage de ceux-ci, ensuite se déroule une étape de calculs sur les valeurs, enfin a lieu une sélection de l'information pertinente.

1.3.2.1. *La mécanique d'apprentissage*

La sémantique nécessaire à la rédaction est construite automatiquement à partir d'un modèle standard d'indicateur et de particularités propres à chaque indicateur. Cette mécanique d'apprentissage est paramétrable et peut être complétée en substituant aux données construites automatiquement d'autres informations décrites manuellement (dans un fichier).

Informatiquement «parlant» l'apprentissage implique l'utilisation de trois fichiers de «connaissances» spécifiques au domaine traité: la description minimale des indicateurs (nom, types, liens avec les séries), les informations communes à tous, et enfin des compléments d'informations particulières (facultatives).

De nombreux choix a priori seront ici définis par le statisticien: Entre autres des choix sur les types d'indicateurs, les périodes analysables, et les paramètres de calculs d'index de jugements. Par défaut, c'est le système qui effectuera ces choix.

1.3.2.2. *L'analyse numérique extensive*

Une fois la base de connaissances construite, Alien réalise une analyse extensive des valeurs numériques. Ceci consiste à appliquer sur toutes les périodes de temps analysables —celles-ci ont été fixées lors de la phase d'apprentissage— des formules mathématiques prédéfinies. Il s'agit d'obtenir le maximum de données quantitatives pour chaque indicateur: les niveaux des valeurs, mais aussi les variations et parfois les accélérations. Pour chaque information, un jugement qualitatif sera porté et codifié sur une échelle de sept valeurs: de «très faible» à «très fort». Ceci constituera des index de jugement.

Dans une analyse manuelle, un statisticien portera généralement un jugement sur les valeurs qu'il observe. Ainsi une augmentation du chômage, des prix, ou encore de la production, sera qualifiée par un expert de faible, de moyenne ou de forte! Comment de tels qualificatifs se forment-ils?

En fait, il n'existe pas de règles absolues. Ainsi entre 1976 et 1978 l'Insee qualifiait ainsi diverses hausses de prix:

- en 1978: faible, avec 7,9%
- en 1976:
  - vive, avec 9,6%
  - modérée, avec 7%
  - quasi-stable, avec 2,7%

  Chacun aura constaté que la qualification d'un chiffre se fait souvent dans un contexte historique, mais doit-on toujours faire référence au passé? ou doit-on au contraire avoir une référence absolue? Dans **Alien** différentes options sont possibles:

- les tranches peuvent être définies manuellement,

- les tranches peuvent être calculées automatiquement à partir de l'historique de la série (par défaut), Dans ce dernier cas, la «longueur» de la série est importante et influence les résultats.

- Si les tranches sont calculées automatiquement, l'usager peut encore fixer certaines options: Alien définit une valeur centrale, correspondant à une situation moyenne. Pour une série en niveau ou en stock, c'est la moyenne; pour les séries en variation c'est 0 si on raisonne en différence, 1 en rapport, 100 en pourcentage. Mais l'usager peut aussi fixer librement ces valeurs centrales.

  Dans le calcul des jugements, les valeurs proches de cette moyenne feront partie de la tranche moyenne et plus on s'éloignera de celle-ci, plus on déterminera les tranches permettant de qualifier: assez faible, assez fort —faible, fort— très faible, très fort (ces adjectifs sont bien évidement dépendants de la série traitée). L'éloignement de la valeur moyenne est aussi paramétrable.

1.3.3. La recherche de l'information pertinente

Une fois qu'un développement extensif sur les valeurs et les jugements a été réalisé, il n'est pas possible de faire une rédaction qui utiliserait tout ce matériaux. Une sélection est indispensable car on dispose de trop d'information. En effet, un texte bien écrit va directement à l'essentiel. La troisième étape des traitements dans **Alien** réalise cette sélection. Seule l'information la plus «forte» est conservée et mise en avant à partir de règles et de choix prédéfinis, mais eux aussi modifiables. Le résultat de cette opération est un «vecteur narratif», c'est-à-dire une séquence d'informations qui défini le «quoi dire».

1.3.4. La génération et les structures narratives stéréotypées

Lorsque les indicateurs sont complètement renseignés et que le vecteur narratif est codé, il ne reste plus qu'à rédiger. Jusqu'ici, toutes les étapes étaient indépendantes de la langue. A partir de maintenant, ce ne sera plus le cas. On va traduire le vecteur narratif à l'aide de deux outils. Le premier est un ensemble de vecteurs d'ordres de rédaction. Il existe plusieurs vecteurs d'ordres, ou séquences d'ordres, ce sont les «schémas narratifs» ou «structures narratives». Le second outil est bien sûr constitué des dictionnaires.

Le changement de langue implique de changer de dictionnaires, mais aussi de modifier les vecteurs des schémas narratifs. Ainsi l'Ifo a réécrit entièrement ces schémas pour la langue allemande. Actuellement, pour les tests effectués en Anglais, en Italien et en Espagnol, on utilise les structures narratives de la langue française. Le résultat est donc imparfait.

Les structures narratives gèrent la plupart des problèmes courant de lexicalité:

- l'élision qui consiste dans la suppression d'une voyelle finale devant une autre voyelle ou un h muet. L'élision est gérée pour les articles, mais aussi devant des mots invariables, tels «alors que». L'une des difficultés consiste à gérer les h (muets ou non). Les mots débutants par des «h» sont donc codés spécifiquement.
- les accords et les conjugaisons sont traités par les dictionnaires et des fonctions associées. Une variable globale du système, **syntax**, identifie à tout moment le genre et le nombre actif; elle peut être «sauvegardée» pour introduire dans un discours des éléments ayant un genre et un nombre différent du discours central; lorsque l'on retourne à ce discours, le genre et le nombre qui étaient actifs retrouvent leur état antérieur.
- les indicateurs quantitatifs ont un discours qui intègre, de diverses façons, les valeurs qui interviennent dans l'analyse. Mais ces valeurs ne doivent généralement pas apparaître lorsqu'elles sont égales à 0.
- l'enchaînement des prépositions et des pronoms; on dit:

  «*la production augmente sauf dans la mécanique où elle diminue*»
  «*Les carnets de commande sont stationnaires, sauf pour les carnets étrangers qui se sont dégarnis*».

- l'enchaînement stylistique des propositions: il est souvent plus élégant de faire suivre une proposition contenant un verbe transitif par une autre contenant une forme intransitive, et réciproquement. Il existe aussi des enchaînements particuliers avec l'auxiliaire être. Alien assure une telle gestion en mémorisant, à tout moment, le sujet traité.
- l'enchaînement des discours entre indicateurs (par des nuances: cependant, mais ...).

1.3.5. Le concept d'application

Une application **Alien** est un ensemble de connaissances sur un sujet donné. C'est sur ce sujet que sera réalisée la génération de texte.

Une application est constituée tout simplement de fichiers de connaissances et d'une base de données. Ces connaissances sont les connaissances particulières au sujet traité, et non les connaissances générales comme les dictionnaires. Elles sont formées des éléments suivants:

- le squelette de l'application: c'est le descriptif minimal pour CHAQUE futur indicateur: on donnera ici les types de l'indicateur, son nom et la, ou les, série(s) à utiliser,
- le modèle d'indicateur renseigné: c'est un prototype d'indicateur, mais complètement garni avec des valeurs communes pour tous les indicateurs,
- le fichier de complément stylistique: il est facultatif et permet d'améliorer l'apprentissage automatique. Il contiendra les tournures et le vocabulaire propre à l'application.

Le travail le plus délicat est de renseigner le squelette, en particulier ce dernier doit contenir une description typologique des indicateurs, ce qui est parfois délicat à définir. Ainsi plusieurs types sont utilisés. A titre d'illustration, examinons le type principal (il existe aussi des types secondaires).

Ce type permet de rattacher la série de base de l'indicateur à une famille parmi plusieurs possibles:

- type 0 et 1: les valeurs de la série à analyser sont significatives d'un niveau, d'un stock, d'une quantité absolue, et ce niveau a une signification économique par lui-même. Par exemple, un effectif salarié, un stock de produits finis, un chiffre d'affaires. Dans le type 1 l'étude de la variation de ce niveau ne pourra être faite, car elle n'aurait pas de sens.
- type 2: les valeurs de la série à analyser sont significatives d'un niveau, mais celui-ci ne veut rien dire; c'est le cas d'un indice, comme l'indice des prix, par exemple. On ne peut donc le commenter. Par contre, la variation de la série, entre deux périodes, est significative.
- type 3: les valeurs de la série à analyser sont ici significatives d'une variation. Le niveau est inconnu. La variation, entre deux périodes, de la série implique une accélération (ou un ralentissement, ou encore d'une stabilité de la variation).

Le choix du type implique des commentaires, des analyses, parfois fort différentes. Il est possible d'obtenir des commentaires fort différents sur des séries en ne modifiant que le type principal dans le descriptif de l'application!

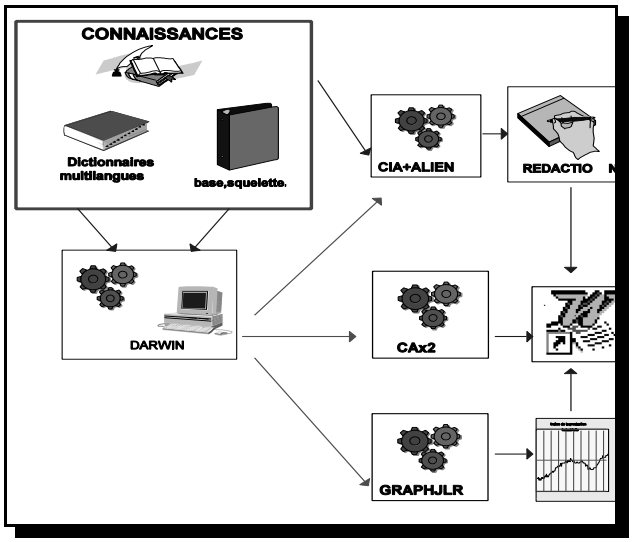
Si le type est mal choisi, le commentaire sera peut-être tout simplement faux! Mais ceci ne peut être reproché à Alien: un tel commentaire faux sera aussi bien fait par un être humain qui connaîtrait mal son sujet!

2. L'INTERFACE UTILISATEUR: DARWIN

Alien reste complexe à utiliser, aussi un environnement agréable d'utilisation, **Darwin**, à été réalisé. **Darwin** est un environnement de travail sous Windows. Il apporte une utilisation facile des textes générés à travers des éditeurs, un outil graphique de visualisation des séries (**Graphjlr**) et une mise en forme finale du texte, des graphiques et des tableaux par appel d'un traitement de texte, tel WINWORD. Cette mise en forme est automatique par l'intermédiaire du logiciel **Cax2**.

Darwin est ses logiciels associés sont écrits en langage C avec l'Api Windows 32 bits. Ils sont compatibles Win 95 et Windows 3.1.

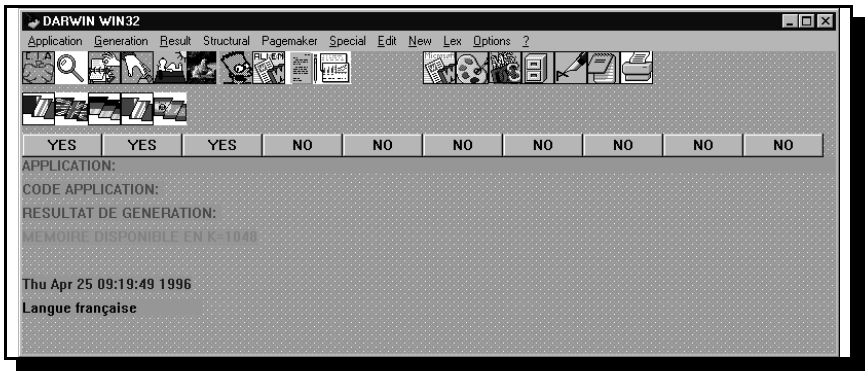
Outre un environnement de génération de texte, **Darwin** permet aussi de manipuler, de générer, de modifier les éléments d'une application; il autorise encore de choisir la langue de rédaction (cette partie est en cours de développement).



Dans l'avenir, **Darwin** automatisera une partie des tâches de construction des nouvelles applications, des dictionnaires, et peut-être même des structures narratives. Examinons les premières possibilités de cet outil.

2.1. Les capacites du logiciel

Avant tout Darwin permet de réaliser des générations de texte à partir d'Alien, mais il informe aussi l'usager sur le contenu d'une application, il réalise la mise en page (sous Word) avec des graphiques; enfin, il permet facilement le changement de langue.

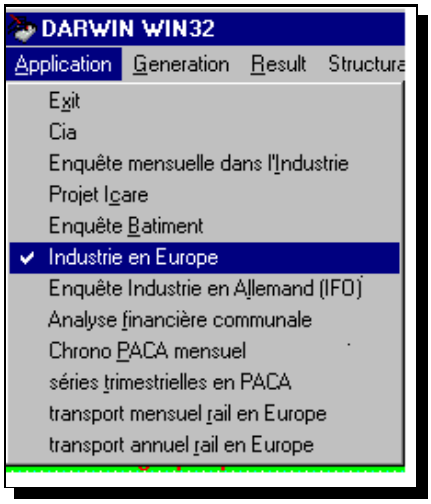


2.1.1. La génération

Réaliser une génération se fait en trois étapes: choisir une application, lancer la génération, examiner les résultats.

2.1.1.1. Le menu application

Toute utilisation implique d'avoir sélectionné une **Application** dans un menu dynamique:



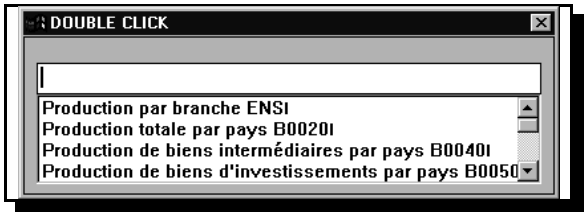
La sélection se fait sur le menu principal. Elle fait référence au contenu de fichiers d'initialisation qui définissent où se trouvent les applications.

2.1.1.2. Le menu génération

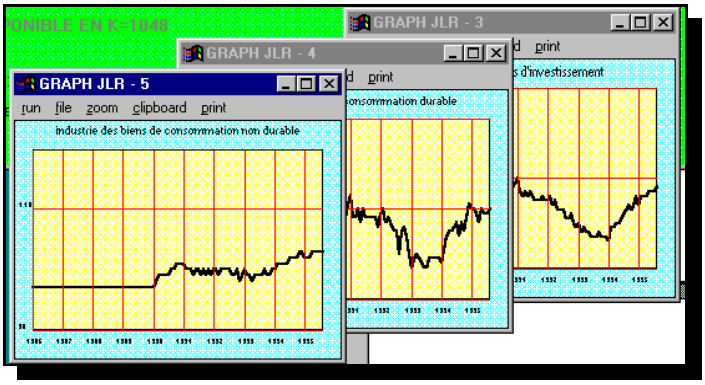
Il fait appel au logiciel d'intelligence artificielle Cia. La génération peut être, ou non, visible a l'écran.

2.1.1.3. Le menu résultat

La génération étant bien terminée, le menu **Result** permet de visionner les résultats. Par exemple à travers des éditeurs de texte, et donc de la modifier. Il est aussi possible de visionner les graphiques et les tableaux associés. Si la génération a été très importante, parfois une centaine de fichiers sont construit à raison d'un par paragraphe, Darwin permet à l'utilisateur de choisir le paragraphe qu'il désire visionner, ou dont il désire examiner les graphiques.



Les graphiques apparaissent superposés, avec un menu qui permet de nombreuses manipulations (zoom total ou partiel, sauvegardes...).



2.1.2. L'information

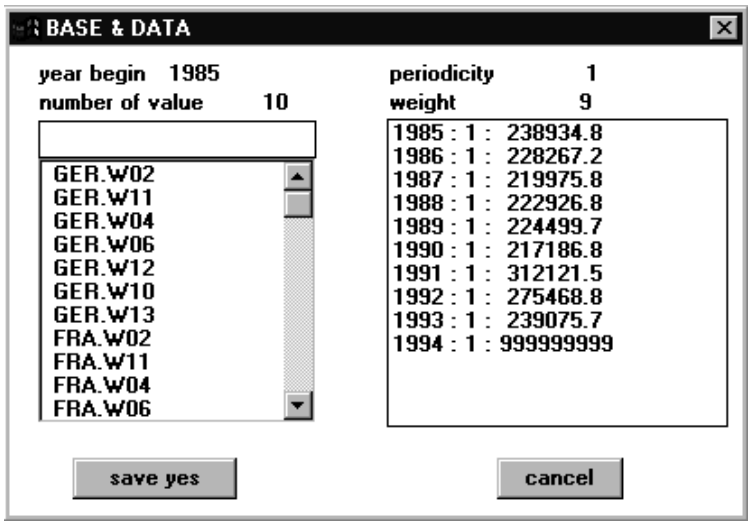
2.1.2.1. Les éditeurs simples

Un menu **Edit** fournit des éditeurs permettant de visionner rapidement et de modifier tous les fichiers indispensables pour une application.

Un menu lexical permet de visionner les dictionnaires dans le contexte d'une langue:

2.1.2.2. Les éditeurs structurels

Ces éditeurs particuliers autorisent une vision structurée des connaissances d'une application: Ainsi, par exemple, la base de données est visible par le nom des séries qu'**Alien** utilise; Et à partir d'un clic et d'un ascenseur, chaque série peut être explorée.



2.1.3. La mise en page

Un menu **Word** permet une mise en page automatique sous un autre traitement de texte de la génération issue d'**Alien**, et éventuellement modifiée par l'utilisateur. Un fond de page doit être décrit ainsi qu'une liste de commandes dans un fichier particulier. Les graphiques et tableaux peuvent être insérés. L'ensemble des opérations est géré par le logiciel **Cax2**. Les versions 2,6 et 7 de Word sont reconnues.

2.1.4. L'utilisation multilangue

Des drapeaux permettent de visionner les dictionnaires et de choisir une génération dans une langue particulière. En fait seul l'anglais est en cours de développement. L'italien et l'espagnol sont envisagés. Une version allemande partielle a été développée par l'**Ifo**.

3. LES UTILISATIONS

L'insee et Eurostat utilisent expérimentalement ce logiciel.

3.1. A l'Insee: la conjoncture

Alien est utilisé tous les mois pour l'enquête mensuelle dans l'industrie. Un commentaire de quelques pages est fabriqué, avec tableaux et graphiques (cf. annexe).

Simultanément, un projet de retour d'information aux entreprises du diagnostic conjoncturel sur leur secteur est maintenant opérationnel. A titre d'information Alien compose actuellement deux pages de texte pour 41 produits industriels en quelques secondes.

3.2. A Eurostat: les transports et l'industrie

Dans le cadre du projet de recherche SUPCOM, le logiciel est adapté et testé pour les résultats des statistiques sur les transports (données annuelles et mensuelles) et des données sectorielles par branche.

3.3. D'autres utilisations

L'Institut de recherche économique **Ifo** à Munich a une version en langue allemande de **Darwin**. Un projet en cours envisage avec d'autres partenaires une version plus complète intégrant aussi la langue hollandaise.

4. CONCLUSION

Darwin et Alien fonctionnent à ce jour, mais suscitent de nombreuses et parfois de violentes critiques. On peut classer les opposants à son utilisation en trois groupes:

- ceux qui sont certains que l'analyse statistique faite par l'être humain ne peut, en aucune manière, être reproduite par une machine,
- ceux qui pensent que la reproductibilité sera possible dans le futur, mais qu'elle est encore trop complexe et trop mal connue. Donc, celle qui peut être faite actuellement est insuffisante.
- Bien sûr, tous ces opposants ont partiellement raison. Aucune machine ne pourra égaler une analyse humaine faite par un spécialiste de haut niveau. Par contre, il est vrai qu'il est possible de formaliser des commentaires simples, mais qu'une telle méthode reste complexe et difficile. Enfin le problème des gains de productivité que cela pourrait entraîner reste entier.

5. EXEMPLES

5.1. Exemple 1: Eurostat

Production de l'industrie manufacturière par pays B0200

Sur la période juin-aout, l'indice de la production de l'industrie manufacturière dans la communauté européenne a continué d'augmenter faiblement de 0.8%, en moyenne mobile sur les trois derniers mois par rapport à la période mars-mai 1995. Alors que l'indice correspondant au Japon a subi encore un sensible recul de -2.4%. Le même indice aux Etats-Unis est resté stationnaire à 0.1%. En ce qui concerne les états membres de l'Union Européenne, et pour le même indice et la même périodicité, l'augmentation a été vérifiée particulièrement pour l'Irlande (5.1%) et pour l'Autriche (3.8%); en revanche pour le Danemark (-0.9%) l'indice de la production a diminué faiblement.

5.2. Exemple 2: Eurostat

Trafic ferroviaire total ENS

Dans l'Union européenne en 1994, le trafic ferroviaire total (national et international) (sauf Allemagne de l'Ouest, Danemark, Espagne, Italie, Luxembourg) est resté très faible (303147.0 tonnes) mais a connu une sensible augmentation (1.7%). Cette augmentation a été notée pour la Belgique (10.2%), pour la France (5.9%) et pour les Pays-bas (6.5%); tandis que la situation est opposée principalement pour la Grèce ($-58.\%$). Le trafic total par chemin de fer en tonnes kilomètres (excepté Allemagne de l'Ouest, Danemark, Espagne, Italie, Luxembourg) est demeuré faible (69133.00 tonnes) mais a connu une très nette augmentation (4.2%). Cette augmentation a été observée pour la Belgique (7.0%), pour la France (7.9%) et pour les Pays-bas (5.1%); à l'opposé en particulier pour la Grèce ($-35.\%$) et pour la Grande-Bretagne (-5.7%) le trafic total par chemin de fer en tonnes kilomètres a diminué.

5.3. Exemple 3: INSEE

Ensemble de l'industrie ENS

Selon les chefs d'entreprise interrogés à l'enquête mensuelle d'avril, l'activité industrielle a continué de diminuer légèrement, à un rythme plus rapide qu'au trimestre précédent. Les stocks de produits finis sont toujours jugés faiblement supérieurs à la normale mais ont semblé subir une baisse pour le deuxième mois consécutif. Les carnets de commandes sont restés faiblement dégarnis mais ont connu encore une légère amélioration. Les perspectives personnelles de production pour les prochains mois étaient plutôt optimistes; à l'opposé les perspectives générales étaient plutôt pessimistes. Les industriels interrogés prévoient des augmentations de prix à la production pour l'ensemble de l'industrie; à l'inverse pour leurs propres produits ils s'attendent à des baisses de prix à la production.

6. BIBLIOGRAPHIE

[1] **Roos, J.L.** (1995). *ALIEN: Un outil pour modéliser la rédaction de diagnostics économiques*. Journées de Méthodologie Statistique - 15 et 16 décembre 1993, Publication Insee (1995).

[2] **Roos, J.L.** (1995). *Comment assurer la crédibilité des discours économiques*. IA 95, Génie Linguistique, Montpellier, Juin 1995.

[3] **Roos, J.L.** (1992). *Intelligence Artificielle en Langage C*. Editions Eyrolles, (1992).

Darwin et ALIEN ont bénéficié de l'aide et du financement de la Commission Européenne à travers les projets DOSES (contrat 2661006) et SUPCOM (contrat 56610005).

Ils ont toujours été par ailleurs soutenus, et ce depuis plusieurs années, par la Division Enquêtes de Conjoncture de l'INSEE.

La version en langue allemande a été faite par l'IFO (INSTITUT FUR WIRTSCHAFTSFORSCHUNG, Munich).

ENGLISH SUMMARY

TEXT GENERATION FROM STATISTICAL DATA

Dr. JEAN LOUIS ROOS*

Institut National de la Statistique et des Etudes Economiques

ALIEN is a knowledge based system that modelizes the economist's behaviour It is used at the French Statistical Institute and at Eurostat. It can be seen as an assistant that help economists to write economic diagnosis in natural language as well as a human economist could do.

Alien is very useful for analysis in very large data base and/or for regular diagnosis. It study the shape of all series in an application, the instantaneous relationship and determine what is important and what is not.

ALIEN is used under a user friendly software (DARWIN) developed for Windows 95. DARWIN build text generation with ALIEN, but also construct graphs and table, help the user to analyse the data base and the semantic of an application. Last, DARWIN build a nice layout of the mixed comment, graph and table with WINWORD or an equivalent software.

Keywords: Weighted Principal Components Analysis, Multiple Correspondence Analysis, Mixed Tables, Indicator Variables, Weighting

* Dr. Jean Louis Roos. Institut National de la Statistique et des Etudes Economiques. Avenue Albert Einstein BP 26000, 13791 Aix En Provence Cedex 3. Tel 42 16 29 59. Fax 42 16 29 00. roos@cil3a.insee.atlas.fr

– Received october 1996.

– Accepted february 1997.

More and more frequently statistical databases are very large. If statisticians have to produce regular commentaries on the data contained in such databases, their work may be particularly protracted and tedious as they will have to examine substantial volumes of figures. Moreover, the data have a semantic aspect which never (or only very rarely) appears in the databases, so that the commentary produced by statisticians will to a great extent depend on their personal knowledge. For this reason, any change in the person responsible for producing regular statistical studies will sometimes dramatically affect the quality of the analysis texts. Strangely enough, the software manufacturers have put a lot of work into developing powerful consultation tools, but have done nothing about interpreting the contents of such databases.

As we see it, statisticians urgently need a tool to meet three requirements:

- **Statisticians need an assistant**

  Statisticians should regularly be able to ask an assistant for a correctly drafted text analysing the information in a database. It should be possible to accompany texts of this kind with tables and graphs and it should be easy to modify them using a word-processing system. The assistant should be efficient enough not to overlook anything important and, obviously, the analysis must be «stable» over time and totally independent of external subjective factors.

- **Statisticians need to be able to deal with «Bulk Commentaires»**

  Often it is not possible to use all the information available for producing a commentary, simply because the bulk of information to be processed is too great —as in the case of sectoral analyses or analysis by local-authority area or country etc. Sometimes for reasons of time alone, a human being is unable to produce dozens or hundreds of diagnoses. The difficulties will be even greater if the documents need to be accompanied by graphs and formatted.

- **Statisticians need guidelines**

  The term «guidelines» means the definition of recommended good practice in various fields - in other words, what one should do, say or write in a given situation. Guidelines of this kind have been developed particularly in the field of medicine, by formalising diagnoses or treatments on which consensus had been reached. By a process of simplification, a guideline enables a situation and a treatment to be described in a formalised fashion which is accepted by a majority of practitioners. Admittedly very relative, this consensus nevertheless provides an efficient and reasoned response to a given disorder.

It is equally important for statisticians to have guidelines of this kind for each of the databases with which they work, as otherwise there is a great risk of interpretation errors.

As part of its specific work on formalising short-term analysis, INSEE has started to develop a programme to meet these requirements. EUROSTAT very quickly joined in with project and has provided a large part of the finance. The software is now in use at both INSEE and EUROSTAT. The text generation programme is «**Alien**» (Assistant Logique pour l'Interprétation Experte de données Numerique), and its operating environment under Windows is «**Darwin**» (Diagnostic, Analyse et Rapport sous Windows). The whole package is, as it were, a computerised assistant for statisticians.

1. TEXT GENERATION: ALIEN

Alien is a software which formalises —in fact models— the mechanisms leading a statistician to produce a given diagnosis from a set of quantitative data. This formalisation is based on the following observation: the majority of simple analysis texts have a stereotyped aspect -regardless of the language in which they are written. This means that very repetitive analysis schemes exist, which can be modelled.

1.1. Theoretical principles

The principles underlying **Alien** are fairly simple: statisticians no longer manipulate series but rather indicators, i.e. entities with a semantic aspect. Secondly, statisticians use stereotyped forms of discourse.

1.1.1. The indicators

Alien is based on the idea that economists do not manipulate statistical series, but rather a set of information about the series, including a knowledge of the vocabulary associated with each series.

1.1.2. Stereotyped discourse

Languages are so rich that it would be impossible, without using enormous resources, to define a method for writing all the possible discourses for a given diagnosis. However, experience shows that the types of discourse used for statistical commentaries are very stereotyped, because of laziness or at least out of habit.

1.1.3. Dictionaries

Dictionaries perform several functions in the drafting of a text. Obviously, their prime purpose is to provide the vocabulary, but they also show most of the most important grammatical rules, such as plural forms or conjugations. The same is true

of the **Alien** dictionaries. Just like a human being, **Alien** will also use dictionaries of synonyms or dictionaries of the dialects we have just mentioned.

1.2. The formal system: a double tree-structure

Writing is a particular human activity which involves frequent backtracking, and this is also true of statistical analysis. It is therefore difficult to produce a totally deterministic software. In practice, **Alien** is based on a double information tree-structure, text generation being handled by a series of small processing operations organised on the terminal nodes and using the dictionaries.

1.2.1. Indicator tree-structure

A main indicator defines the subject. The main indicators each have their own sub-sets of domain indicators which will constitute paragraphs of text. Each domain is made up of quantitative or qualitative indicators, some of which have tokens (for example, production may be analysed by branch or country, etc.). These indicators form the discourses on each series for which a commentary is required.

1.2.2. Content indicator tree-structure

All the indicators are structured in the same way. They have properties and each one is assigned a value (sometimes by default). An indicator has around 200 properties and it is possible to add original properties for each application.

1.2.3. The dictionaries

The dictionaries form a basic-tree structure based on «entries». They are specific entities which group the vocabulary common to the indicators and most of the associated grammatical rules and conjugations.

2. THE USER INTERFACE: DARWIN

Alien is still complicated to use, and therefore a user-friendly interface has been developed. This interface, known as **Darwin**, is a working environment under Windows. It provides easy use of the text generated via editors, a graphic tool for displaying series and final formatting of the text, graphs and tables using a word processing system, such as WINWORD.

Darwin and its associated programs are written in C with the 32-bit Windows Api. They are compatible with Windows 95 and Windows 3.1.

Apart from providing a text-generation environment, **Darwin** makes it possible to manipulate, generate and modify the elements in an application and to choose the language in which the text is to be produced (this section is still under development).

2.1. Capacities of the software

2.1.1. The generation menu

This uses the **Cia** artificial intelligence software. Generation may be displayed on the screen, or not as required.

2.1.2. The result menu

When generation has been completed, the Result menu enables the results to be displayed through, for example, text editors, and hence modifications are possible. It is also possible to display the associated graphs and tables. If the item generated is very large (sometimes a hundred or so files are constructed at the rate of one per paragraph), **Darwin** enables users to choose the paragraph they wish to view or in which they want to examine the graphs.

2.1.3. The structural editors

These specific editors permit structured viewing of the knowledge of an application. For example, a database is viewed in terms of the name of the series used by **Alien**; each series can be explored by clicking on the vertical scroll bar.

2.1.4. Formatting

A layout menu permits automatic formatting of **Alien** output using a different processing system, possibly with user modifications. A basic page structure and a list of commands must be defined in a given file. The graphs and tables can be inserted.

2.1.5. Multilingual use

Flags make it possible to display the dictionaries and choose generation in a given language. Only the English version is in fact being developed. Italian and Spanish are planned and a partial version in German has been developed by **Ifo**.

3. USES

INSEE and EUROSTAT are using this program on an experimental basis.